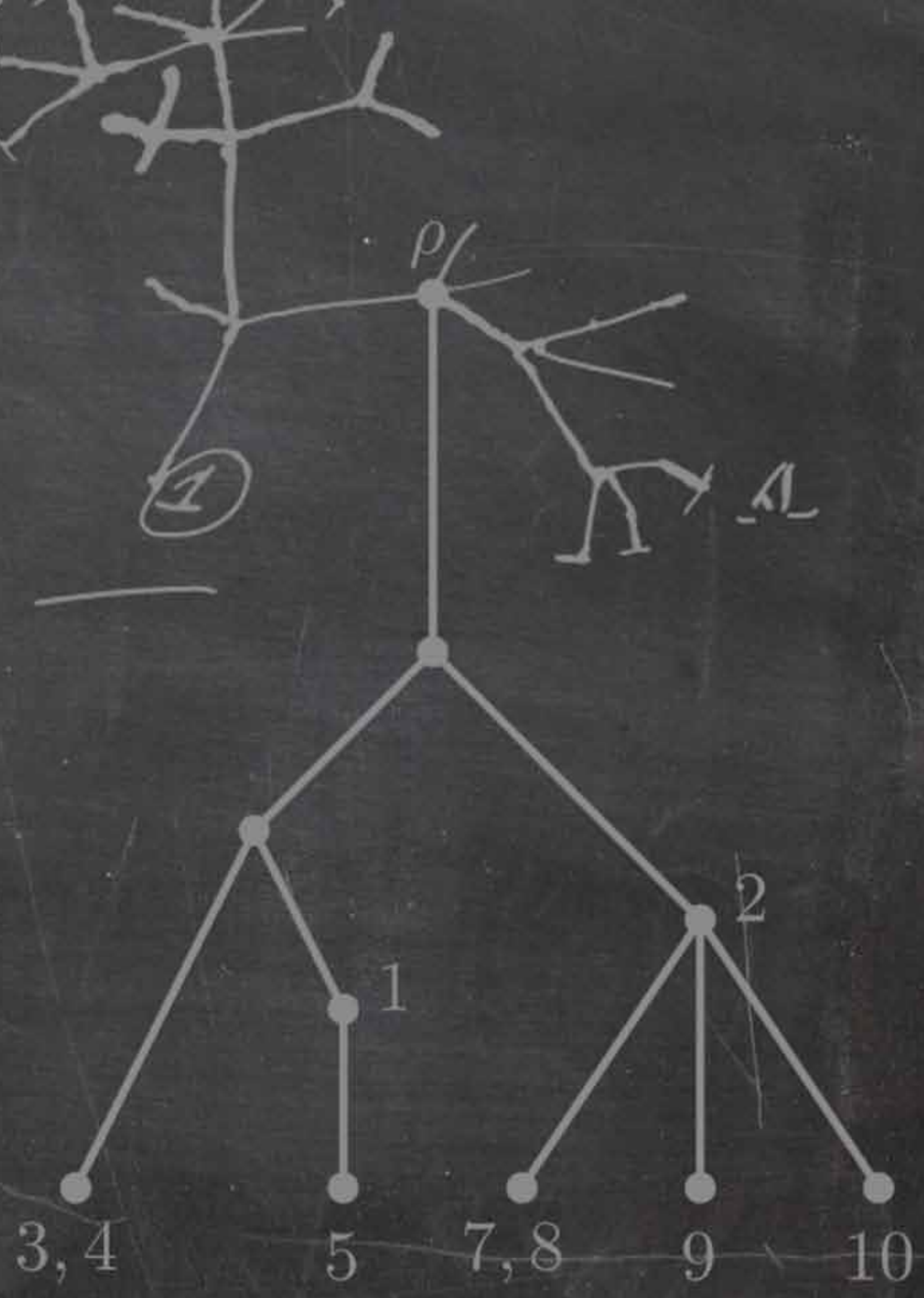
Phylogenetic Trees Peter Regner



Phylogenetic Trees

Selected Combinatorial Problems

Peter Regner

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

DIPLOMARBEIT

Phylogenetic Trees

Selected Combinatorial Problems

ausgeführt am

Institut für Diskrete Mathematik und Geometrie
der Technischen Universität Wien

unter der Anleitung von

Ao.Univ.Prof. Dr. Bernhard Gittenberger

durch

Peter Regner
1090 Wien

25. April 2012

Phylogenetic Trees

Selected Combinatorial Problems

Peter Regner

Cover design by Anna Vasof in corporation with Peter Regner.
Parts of Figure 0.1 were used (source: [15, p. 36], reproduced from [80]).

Please send any questions, comments, corrections or other requests to:

peter.regner-phylgen@suuf.cc

This document (or an updated unofficial version), all Mathematica notebooks from Appendix A and additional material are available electronically at:

http://suuf.cc/phyl-trees

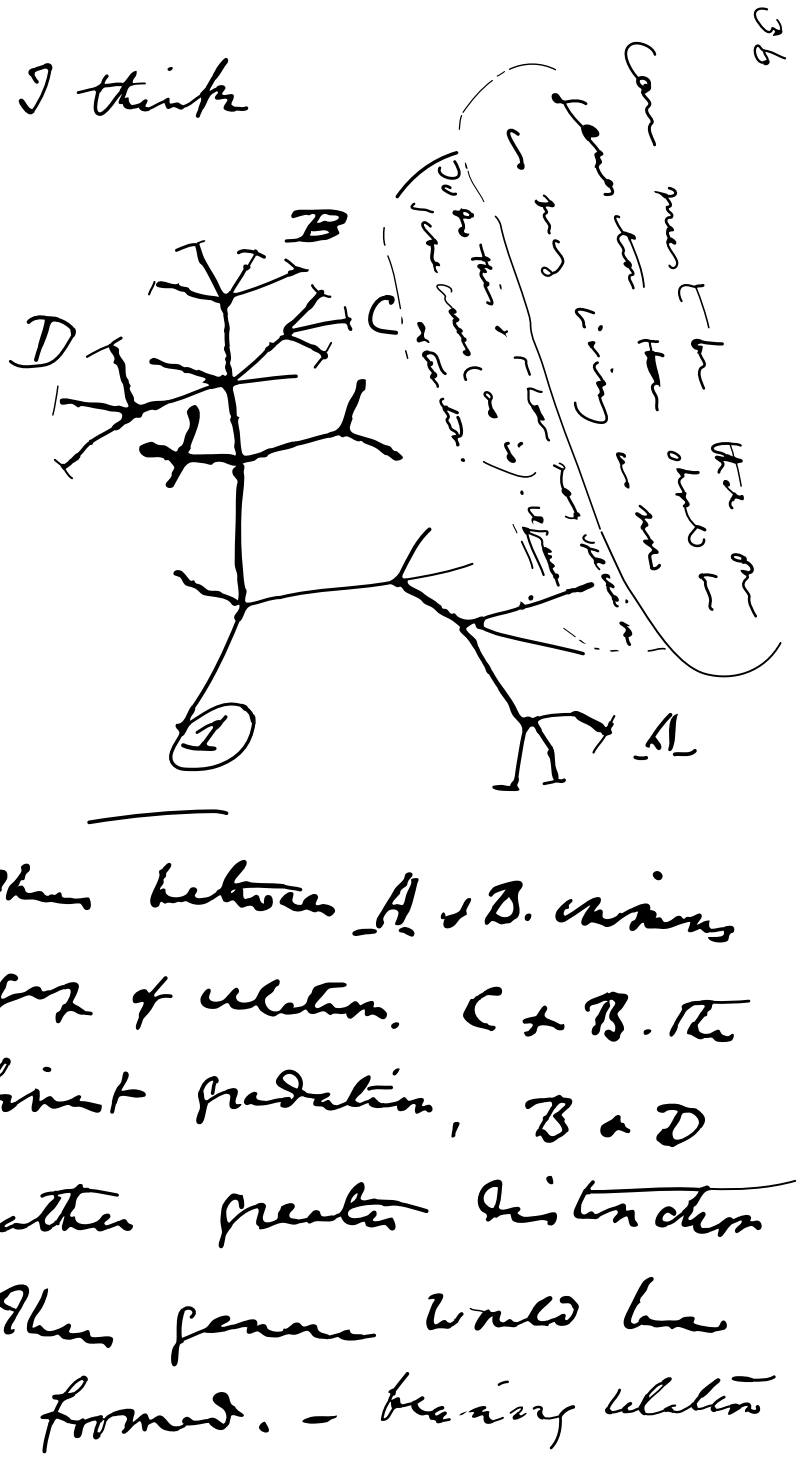


Figure 0.1: Diagram of an evolutionary tree, one of Charles Darwin's early sketches (source: [15, p. 36], reproduced from [80]).

Abstract

In evolutionary biology phylogenetic trees are used to represent evolutionary relationships within a group of species. Typically treelike branching diagrams are used, Figure 0.1, Figure 1.1 and Figure 1.2 show early diagrams by Darwin and Haeckel. In graph theoretic terminology a phylogenetic tree corresponds to a rooted leaf-labeled tree, i.e. a (finite) simple, connected, acyclic graph, where one vertex is distinguished as root and distinct labels are assigned to the leaves of the tree. The labels refer to the various species under consideration and the internal nodes of the tree represent hypothetic ancestral species (see [63, p. 19f.]). The availability of genetic data in the 1960s made it possible to develop formal models of the evolution of species and to apply mathematical methods to infer the evolutionary history of a group of species (see [52]). Still, phylogenetics is an ongoing field of research, the existing models and algorithms are improved and new questions arise.

Chapter 1 gives a rough overview of the field of study and mentions some of the topics not covered in this thesis. In Chapter 2 all fundamental objects will be defined and basic concepts used later will be introduced. The well-known *maximum parsimony* approach is described as well as the *symmetric N_r-model*.

In Chapter 3 and Chapter 4 more specific problems are discussed. Chapter 3 contains a collection of several enumeration problems concerning phylogenetic trees, which were solved by different authors in the last decades. The size of several classes of phylogenetic trees will be determined and two problems concerning random phylogenetic trees will be discussed.

In Chapter 4 maximum parsimony and the symmetric N_r-model will be compared by studying the reconstruction of states of ancestral species. Results by Li et al. [49] and Fischer and Thatte [24] concerning the accuracy of this reconstruction with the Fitch-Hartigan algorithm are presented. Finally, initial results (proved within the scope of this thesis) towards generalizing a theorem by Fischer and Thatte [24] are outlined. In particular, the Mathematica package `Phylgen` was developed to examine special cases of this theorem.

Zusammenfassung

Phylogenetische Bäume werden im Bereich der Evolutionsbiologie verwendet, um evolutionäre Beziehungen innerhalb einer Gruppe von Arten darzustellen. In Abbildung 0.1, Abbildung 1.1 und Abbildung 1.2 sind frühe Versionen solcher baumähnlichen Verzweigungsdiagramme von Darwin und Haeckel zu sehen. Ein phylogenetischer Baum entspricht – im Sinne der Graphentheorie – einem gewurzelten Baum, dessen Blätter markiert sind. Die Blätter korrespondieren mit den betrachteten Arten und die internen Knoten des Baums können als ihre hypothetische Vorfahren gesehen werden (siehe [63, p. 19]). Die Verfügbarkeit von genetischen Daten seit den 1960er-Jahren hat es ermöglicht, formale Modelle für die Evolution der Arten zu entwickeln und mathematische Methoden anzuwenden, um die evolutionäre Geschichte von Arten zu rekonstruieren (siehe [52]).

In Kapitel 1 wird das Forschungsgebiet grob umrissen, einschließlich einiger Themen, die in dieser Arbeit nicht näher ausgeführt werden können. In Kapitel 2 werden alle später verwendeten Begriffe und Objekte formal definiert und grundlegende Konzepte werden vorgestellt. Sowohl der weit verbreitete Maximum-Parsimony-Ansatz („maximale Sparsamkeit“) als auch das symmetrische N_r-Modell werden erläutert.

In Kapitel 3 und Kapitel 4 werden speziellere Problemstellungen behandelt. Kapitel 3 umfasst eine Sammlung verschiedener auf phylogenetische Bäume bezogener Abzählprobleme, die in den vergangenen Jahrzehnten gelöst wurden. Die Anzahl der Bäume in unterschiedlichen Klassen phylogenetischer Bäume wird bestimmt. Ferner werden zwei Fragestellungen mit wahrscheinlichkeitstheoretischem Ansatz diskutiert.

In Kapitel 4 wird Maximum-Parsimony mit dem symmetrischen N_r-Modell verglichen, indem die Rekonstruktion von Merkmalen der Vorfahren untersucht wird. Es werden Resultate von Li et al. [49] und Fischer und Thatte [24] vorgestellt, die die Zuverlässigkeit dieser Rekonstruktion behandeln. Schließlich werden erste Teilresultate zur Verallgemeinerung eines Theorems von Fischer und Thatte [24] vorgestellt, die im Rahmen dieser Arbeit erzielt werden konnten. Insbesondere wurde das Mathematika-Paket `Phylgen` entwickelt, um Spezialfälle dieses Theorems zu untersuchen.

Preface

A personal note

There is a plenty of mathematics out there[1]. Some parts of it can be considered as pure mathematics or basic research, while others are targeting applications, either inside mathematics, in other scientific fields or elsewhere in the real world. And lastly, some problems can be considered essentially as playful approach to mathematics, e.g. to prove the NP-completeness of the computer game Tetris (see [18]), or to find a solution of the 100 prisoners puzzle (see [84]). Such a classification is not meant to justify a priori some areas of mathematics and question the right to exist for others. But especially because there are so many considerable mathematical problems, it makes sense to ask why we are interested in some of them particularly. In general, my fascination for mathematics arises from the mysterious power of formal methods (such as the *symbolic method*, which is outlined briefly in Section 2.4 and will be extensively used in Chapter 3). In my opinion, this is exciting especially if such methods are applicable in the real world, or if there are certain *pure* problems—problems where the question *and* the answer, once found, can be grasped immediately, but without formal methods a solution cannot be derived easily. A neat example of the latter is, how one can compute the number of rotationally distinct ways of drawing one of the two possible diagonals in each of the faces of a cube[2]. In addition, a playful touch is another source of fascination, but it is not at all easy to explain why.

Some fields of mathematics actually fall in two or even more of these categories. A famous example is number theory, which for long has been considered as pure mathematics, but is enjoying a renaissance as application in cryptology. A similar combination of application and basic research can be found in the topic of this thesis: How to apply methods of discrete mathematics to the field of evolutionary biology?

While not being addressed in the following, I can also think of playful aspects of this sub-

[1] The question "How much mathematics can there be?" is discussed in [17, p. 24f.].

[2] It is immediately clear that the answer is a number between 1 and $2^6 = 64$ and it would even possible to draw all possible cubes, but without the help of *Burnside's lemma* one loses easily track of rotationally identical cubes.

ject. What about examining the notoriously branching out of GNU/Linux distributions or the tree of doctoral advisors (see [19]) in so-called academic genealogy with methods known from evolutionary biology? But, let us do the serious work first.

Acknowledgements

First and foremost, I would like to thank my advisor Bernhard Gittenberger, who introduced me to the fascinating world of combinatorics. His support made it possible to combine topics of basic and applied research in this thesis. Furthermore, I am deeply grateful to Mareike Fischer, Arndt von Haeseler and their team at the Center for Integrative Bioinformatics Vienna (CIBIV). Without their help the part presented in Chapter 4 would not have been possible, and many of my questions concerning biology would have remained unanswered. Special thanks go to my parents, who supported me with many helpful comments throughout the development of this thesis and with discussions of problems of all kinds. Additionally, I would like to thank Martin Lackner for his comments and Charlie Allen for helping me out with translation issues. Last, not least, I want to thank Anna Vasof who is responsible for the great cover design and supported me through all the time.

Peter Regner
Vienna, April 2012

Contents

Chapter 1

Introduction

The basic ideas for the understanding of the development of the different forms of life dates back to ancient Greek philosophy, but also in Chinese philosophy similar thoughts can be found. For example, the Encyclopædia Britannica [77] writes about the Greek pre-Socratic philosopher Empedocles[1]:

> “ [...] the most interesting and most matured part of his views dealt with the first origin of plants and animals, and with the physiology of man. As the elements (his deities) entered into combinations, there appeared quaint results—heads without necks, arms without shoulders. Then as these fragmentary structures met, there were seen horned heads on human bodies, bodies of oxen with men’s heads, and figures of double sex. But most of these products of natural forces disappeared as suddenly as they arose; only in those rare cases where the several parts were found adapted to each other, and casual member fitted into casual member, did the complex structures thus formed last. Thus from spontaneous aggregations of casual aggregates, which suited each other as if this had been intended, did the organic universe originally spring. Soon various influences reduced the creatures of double sex to a male and a female, and the world was replenished with organic life. It is impossible not to see in this theory a crude anticipation of the ‘survival of the fittest’ theory of modern evolutionists. ”

Nevertheless, many centuries were still to pass before Charles Darwin[2] published his book *On the Origin of Species* Darwin [14] in 1859. It can be considered to be one of the cornerstones of modern evolutionary biology. After few decades his theory was widely accepted, in spite of

[1] Empedocles, ca. 490–430 BC
[2] Charles Darwin, 12 February 1809–19 April 1882

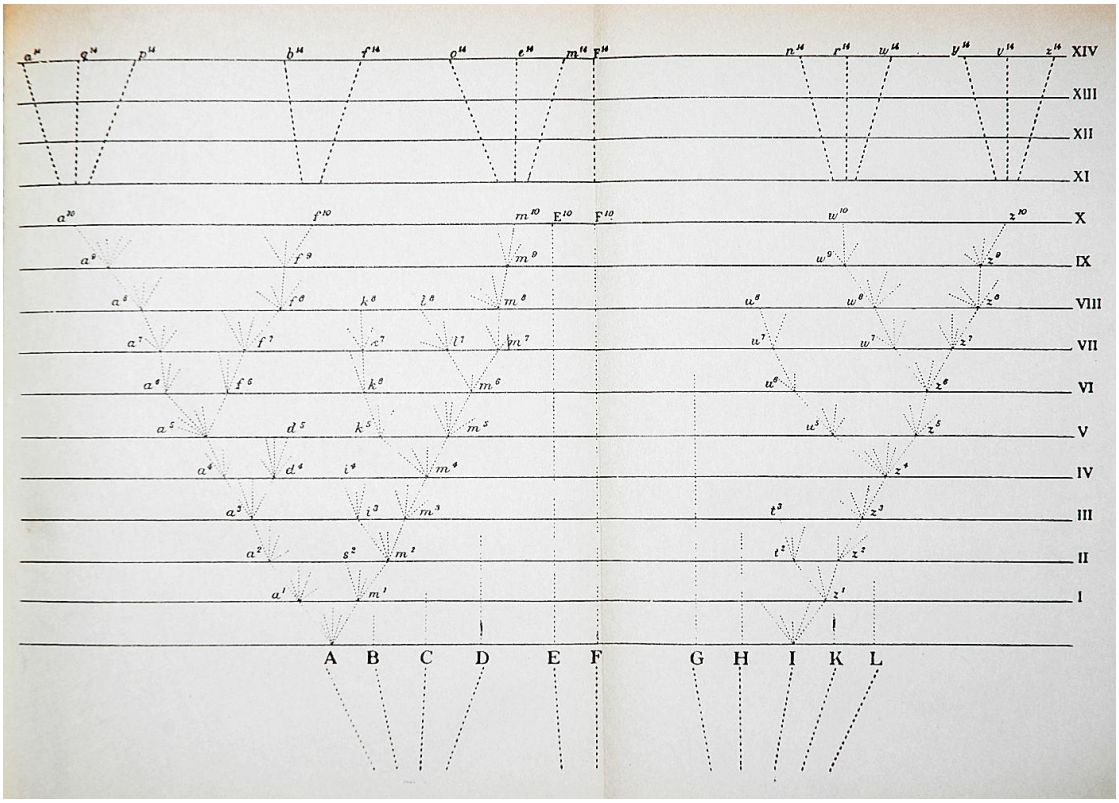


Figure 1.1: Treelike diagram to illustrate the branching of species (source: [14, p. 176], reproduced from [81]).

political and religious contradictions with the prevailing view in Western society back then. Also, there were many supporters for Lamarck's hypothesis. Lamarck assumed that an organism's characteristics, acquired during its lifetime, can be inherited. But due to a lack of evidence for this theory Darwin's conception of *natural selection* has prevailed. Darwin proposed that all forms of life evolved from ancestral species. He actually claimed that there is one common ancestor for all organisms from which the extant species diverged through random variation and natural selection:

> "Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed."[3]

Therefore it makes sense to illustrate the historical evolution of species in a treelike branching diagram as he did in one of his earlier sketches, shown in Figure 0.1, and in a diagram in [14]

[3] Darwin [14, p. 484]

shown in Figure 1.1.

Such branching diagrams are still used today when studying evolution. They correspond to trees as known from graph theory or computer science. The question of, how to analyze and construct these evolutionary trees under certain assumptions with mathematical methods, is the core issues of a wide field of ongoing research, combining knowledge and methods from biology, mathematics and computer science. In this thesis we are going to shed light on certain aspects of these questions. This will be a quite specific selection of topics—it is not possible to give detailed attention to all of them. Felsenstein [23, p. xix] estimated in 2004 that "there are about 3,000 papers on methods for inferring phylogenies." These were even too many to be included in the comprehensive textbook [23] by Felsenstein. In addition to this recommendable book we refer the interested reader to *Phylogenetics* by Semple and Steel [63], which provides a profound overview of the mathematical foundations of phylogenetics, and to the survey by Allman and Rhodes [3]. Another book concentrating on "the fundamental mathematical concepts" is [36]. In this chapter we only give a rough overview of the field of activity by providing some more historical background and introductory information.

Phylogenetics—inferring phylogenies by algorithmic methods. A phylogeny describes the evolution of species or specific groups of organisms and therefore also the relatedness between them. The term is derived from the two ancient Greek words φῦλον (tribe, race [50, p. 1698]) and γένεσις (origin, birth [50, p. 305]), and *phylogenetics* is the field of science which deals with these topics. The aim is to understand evolution and the diversity of the different forms of organisms and also to build a system for their classification and naming (the science of doing that is called taxonomy[4]). Carl Linnaeus[5], the founder of Linnaean taxonomy, classified life only by studying morphological characteristics of organisms. He compared form, size, shape, and structure of organisms or parts of them, for example, the number of stamens of plants. In contrast the understanding of evolution allows alternative ways of classification. In cladistics, for example, one uses certain subtrees of a phylogenetic tree to classify species. (There are also schools of biological systematists which make use of phylogenies for this purpose.) Roughly speaking instead of characterizing organisms by their external appearance, one can characterize them by their origin. With respect to this characterization those species which share a common history are closely related to each other.

[4] τάξις = arrangement, order, rank [50, p. 1526]
[5] Carl Linnaeus, 1707–1778

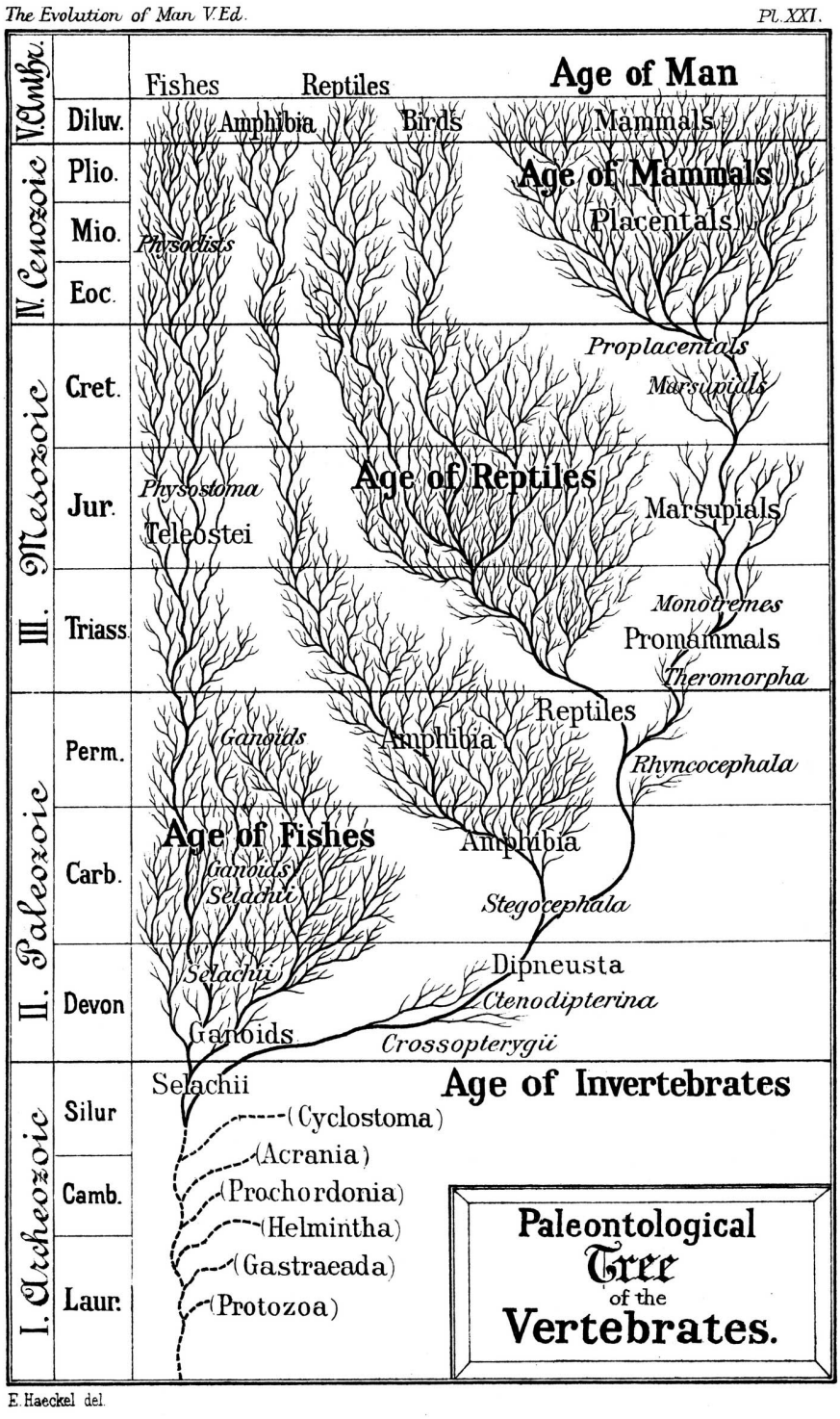


Figure 1.2: *Paleontological Tree of the Vertebrates* by Haeckel [40, Pl. XXI] (reproduced from [82]).

The field of phylogenetics has its origin in times of Darwin. His findings and the work of his contemporaries lay the basis for the hypothesis of a common ancestor and the treelike branching of species. The drawings of phylogenetic trees by Darwin and others at that time (see Figure 0.1, Figure 1.1, Figure 1.2) are very similar to phylogenetic trees used in modern evolutionary biology. In fact, Darwin's sketches include already all fundamental characteristics.

Nevertheless, it took another 100 years before a more formal approach was taken. The availability of the first computers at universities opened up new possibilities for numerical and algorithmic methods. This fact probably was the main reason why a formal way has prevailed. On the other hand, the progress in the field of genetics in the 1950s provided a lot of valuable information. For instance, Allman and Rhodes [3, p. 21] write

> "The availability of sequence data produced a revolution in several ways. First, the volume of available data for any given collection of species grew tremendously. Obtaining data became less of a problem than how to sort through it. Second, since sequences are so amenable to mathematical description, it became possible to formalize the inference process, bringing to bear mathematical tools."

According to Felsenstein [23, p. 123] the paper by Michener and Sokal [51] (published in 1957) might be considered as the first publication about numerical inference of phylogenies. Several articles about numerical clustering methods were published, but only Michener interpreted the taxonomic classification also as valid phylogeny. Therefore from a historical point of view one could say that (evolutionary) taxonomy led to algorithmic inference of phylogenies and not the other way around. In the 1960s several different methods were developed in order to reconstruct ancestral states or the underlying tree—the basic questions and ideas are outlined in the next paragraphs. This introduction is not meant to explain details, but rather to point to topics which are not covered here and to show what awaits the reader in this thesis. Chapter 2 contains a detailed introduction to the fundamental terminology and properties of the objects mentioned in the following paragraphs.

An introductory example. A simple example will help us to illustrate the problem of reconstructing the evolutionary history of a group of species and to explain some basic terms. The example is taken from [23, chapt. 1], but we will try to stick closer to mathematical conventions of notations and terminology[6]. The target of this section is to provide a simple overview of the

[6]These differences are irrelevant to the findings, but as fun fact the following might be worth mentioning. One can observe that biologists tend to draw trees more similar to the woody plant, while in mathematics and computer

Species\Characters	χ_1	χ_2	χ_3	χ_4	χ_5	χ_6
1	α	β	β	α	α	β
2	α	α	β	α	α	α
3	α	α	β	β	β	β
4	β	β	α	β	β	β
5	β	β	α	α	α	β

Table 1.1: Character states for the species in X for the characters $\chi_1, \chi_2, \ldots, \chi_6$.

topic without explaining definitions and concepts in detail.

Consider a set of five (extant) species $X = \{1, 2, 3, 4, 5\}$ and six traits $\chi_1, \ldots, \chi_6$, in phylogenetics usually called *characters* (see e.g. [63, chapt. 4]). The elements in X are also often called taxa or *OTU*s (*Operational Taxonomic Units*) [12, p. 721] referring to a group of organisms considered as a unit in taxonomy. A character can be any trait which is present or absent or in any other specifiable state in all individuals of the species under consideration. So for $i = 1, \ldots, 6$ we have functions $\chi_i : X \to C_i$, where C_i is a set of possible states for character χ_i (see Definition 2.3 on page 18). Nowadays genetic data might play the biggest role in the field of algorithmic phylogenetics, but in general a character can be "morphological (e.g. wings versus no-wings), biochemical, physiological, behavioural, embryological, or genetic (e.g. the nucleotide at a particular DNA sequence position or the order of certain genes on a chromosome)" [63, p. 65]. In the following the source of the data for the characters will not matter for us, but only the number of different states for each character. With such an abstract approach the developed methods can be applied not only to biology but also to other scientific fields as we will explain later. In Table 1.1 the states for the characters $\chi_1, \chi_2, \ldots, \chi_6$ are shown. All characters χ_i are *binary characters* that is $|C_i| = 2$. We denote the states with α and β.

Suppose the phylogenetic tree $\mathcal{T}$ in Figure 1.3 is suggested by biologists for the species in X. This is a rooted binary tree in the sense of graph theory and therefore a direction is given implicitly—we view the edges as directed away from the root. The leaves are labeled with elements of X. The internal nodes of the tree are sometimes called *HTU*s (*Hypothetical Taxonomic Units*) in contrast to OTUs [12, p. 721]. They can be viewed either as hypothetical ancestral

science more abstract figures of trees are common. To be precise, trees in graph theory or computer science are usually illustrated upside down with their leaves at the bottom and the root at the top, while phylogenetic trees in evolutionary biology are drawn in the natural way with their leaves at the top and the root at the bottom. But of course, this is not due to a different level of abstraction of the different fields of science, but more likely because trees in mathematics and computer science often start growing at the root while phylogenetic trees actually are constructed by use of the leaves. Phylogenetic trees are actually also often drawn with the root at the left side and the leaves at the right side, because this is very useful for labeling the leaves with longer names.

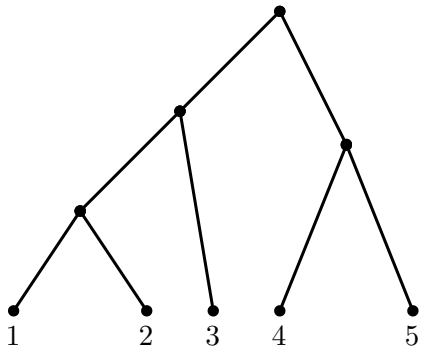


Figure 1.3: Proposed phylogeny $\mathcal{T}$ with the underlying tree T for the species in $X = \{1, 2, 3, 4, 5\}$.

species or as speciation events [63, p. 19f.]. In both cases the tree represents the evolution of the species in X and their historical relatedness as already mentioned and illustrated with the historical examples in Figure 0.1, Figure 1.1 and Figure 1.2. So every species is represented by a vertex of the tree and it evolved from the species represented by its parent vertices. This brings up several questions: How can we reconstruct the character states for the hypothetical ancestral species, that is, are there functions $\overline{\chi_i} : V \to C_i$ which extend the characters χ_i in some natural way to all nodes of the tree? How can we justify the proposed phylogenetic tree by use of the information provided by the characters χ_i? Or is it even possible to reconstruct the correct phylogenetic tree from a set of given characters?

There is not one simple answer to these questions, but the ideas and concepts of the different answers will be outlined in the following. Those concepts which are used later will be formalized and presented in detail in Chapter 2.

Figure 1.4 illustrates two possible extensions of character χ_1 to all internal nodes. In both cases there is one edge (u, v) with $\overline{\chi_1}(u) \neq \overline{\chi_1}(v)$. This represents the situation that the species u and v differ in the trait represented by character χ_1. When v evolved from u, the state of the character changed. This event is called *substitution* (see e.g. [63, chapt. 8]). A common assumption (see [23, p. 3], [63, p. 67f.]) is that every character state arises only once in the tree, or equivalently, that there are exactly $|\chi(C)| - 1$ substitutions for a character χ with state set C (see [63, p. 85, prop. 5.1.3]; details are discussed in Section 2.2, particularly in Proposition 2.6 and in Proposition 2.8). Otherwise, if there are more than $|C| - 1$ substitutions for a character extension, we speak of *homoplasy* (see [23, p. 25]). This implies that a character state arises and then changes back to a previous state (*reverse transition*) or that a state evolves independently several times. In the latter case the state occurs in different subtrees of a vertex v, but v is in a

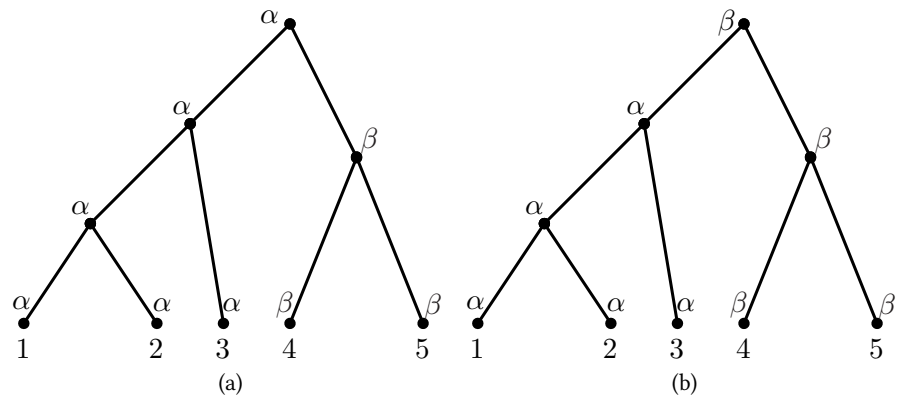


Figure 1.4: Possible extensions of character χ_1 with one substitution. α and β illustrate the character states of the according species, while 1–5 denote the species initially given.

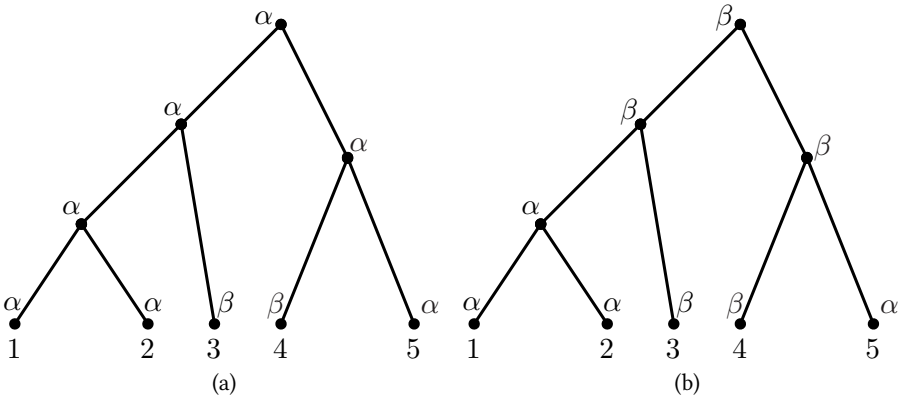


Figure 1.5: Possible extensions of character χ_4 with two substitutions.

different state (*convergent transition*), see [63, p. 67f.] for details. Both reverse transitions and convergent transitions occur in nature, but they can be considered as "relatively rare for certain types of genetic data" [63, p. 68].

In Figure 1.5 the two possible extensions for character χ_4 with two substitutions are illustrated. For this simple example it is immediately clear that there is no extension $\overline{\chi_4}$ with less substitutions. Hence, there is no homoplasy-free character extension for character χ_4. But it is still possible to search for the *most parsimonious* character extension $\overline{\chi_4}$. That is a character extension with the least possible number of substitutions—in the case of χ_4 this is one of the character extensions displayed in Figure 1.5. This approach is called *maximum parsimony*, sometimes abbreviated *MP* (see also [63, chapt. 5]). One could count the number of substitutions

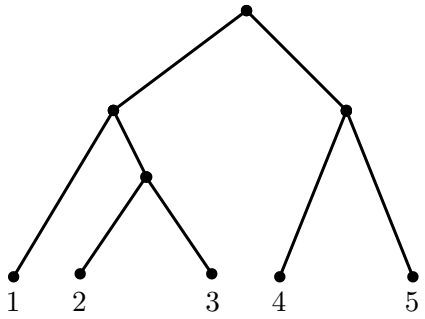


Figure 1.6: An alternative phylogenetic tree $\mathcal{T}'$ with underlying tree T' for the species in $X = \{1, 2, 3, 4, 5\}$.

for all $|C|^{|V|-|X|}$ possible extensions of a character in order to find the most parsimonious one. But for large trees or state sets one should rather use a more efficient method, such as the *Fitch-Hartigan algorithm* with a runtime in $O(|C| \cdot |X|)$ (see [26, 43] for the original papers published in 1971 and 1973 respectively and [63, pp. 89–91], [23, pp. 11-13] for other resources). Basically, the algorithm traverses the tree starting at the leaves with the given states for the character and then infers a set of states for their parent vertices by choosing for each parent vertex the states, which occur most frequently in its children. In our case this procedure for character χ_1 results in $\{\alpha, \beta\}$ for the root ρ while all other states correspond to the two extensions presented in Figure 1.4. Thus, by assuming maximum parsimony the states of all internal nodes except ρ can be reconstructed unambiguously while for ρ both states $\{\alpha, \beta\}$ lead to a character extension of χ_1 with a minimal number of substitutions. By traversing the tree from the root to the leaves in a second pass one can construct explicitly a character extension with a minimal number of substitutions (see Section 2.2, Theorem 2.9 for details).

Considering all possible extensions of the characters $\chi_1, \ldots, \chi_6$ for the phylogenetic tree in Figure 1.3, the total minimal number of substitutions is $9 = 1 + 2 + 1 + 2 + 2 + 1$, which is called *parsimony score*. The phylogenetic tree $\mathcal{T}'$ in Figure 1.6 shows that the previously suggested phylogenetic tree $\mathcal{T}$ is not the most parsimonious one. In Figure 1.7 there are character extensions $\overline{\chi_i}$ for $i = 1, \ldots, 6$ on $\mathcal{T}'$ with a total number of $8 = 1+1+1+2+2+1$ substitutions. This is also the minimal number of substitutions which can be achieved by extending the given characters on any binary phylogenetic tree. One could also expect to find a phylogenetic tree with a parsimony score of 6 or 7 because we have a set of 6 binary characters, but in this case there is no such tree.

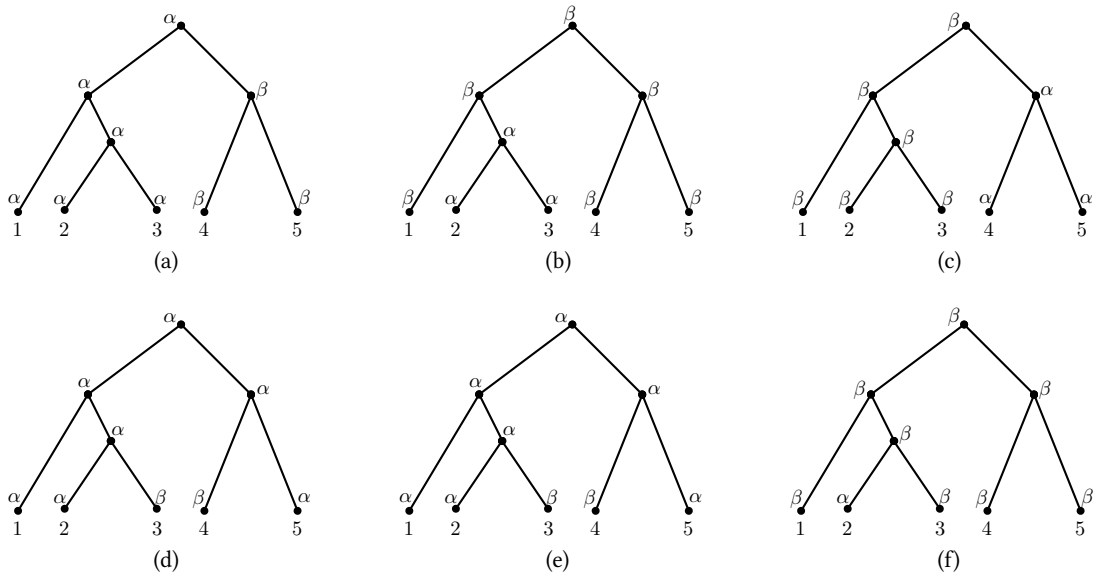


Figure 1.7: Character extensions $\overline{\chi_i}$ for $i = 1, \ldots, 6$ on $\mathcal{T}'$ with a total number of 8 substitutions.

Branch lengths. If one takes into account that phylogenetic trees represent the chronological development of a group of species, it makes sense to impose *branch lengths*—a length for each edge of the tree and therefore a distance between each vertex and its parent. This leads to a weighted tree with positive weights. A distance between two species can then be defined by the sum of the weights of the unique path between these two species. This distance can be viewed as time passed between the speciation events or as a measure of the dissimilarity between the species and its ancestral species (e.g. the Hamming distance of a set of characters, see [58]). Sometimes it is assumed that changes of characters occur at a constant rate through time—then these two views correspond with each other and we speak of a *molecular clock* or *ultrametric* trees (see [24] and [63, sect. 7.2]; some details will be given in Section 2.3 on page 33).

The quantity of dissimilarity between two species can be used as a measure of distance between them. This defines a so-called *dissimilarity map* on the set X, usually represented by a matrix (*distance matrix*) and generalizing the concept of metrics. One can use a dissimilarity map on X to reconstruct a phylogenetic tree for the species in X. Usually we then speak of *distance matrix methods* as opposed to character based methods. A very popular approach using distance matrices is the *Neighbor Joining algorithm* (often abbreviated by NJ) by Saitou and Nei [58] (see also [63, p. 157f.]). The algorithm identifies two vertices from X to form a cherry[7] of the phylo-

[7] In graph theory two leaves of a tree are said to form a *cherry* if they are connected to the same vertex in the tree

genetic tree and replaces them with one new leaf. Then recursively again a cherry is replaced by a leaf until the tree is fully defined. The algorithm aims to fulfill the *minimum evolution* criteria, that is, to minimize the sum of the weights of all edges of the tree. This assumption dates back to Edwards in the early 1960ties (see [23, p. 125f.]). It is similar to the maximum parsimony criteria for phylogenetic trees without distances. Note that not every dissimilarity map on X can be represented by the sum of weights of the corresponding path in a phylogenetic tree. If there is a representation by a specific phylogenetic tree with positive edge weights for a dissimilarity map or its corresponding distance matrix, we call it a *tree metric,* or we refer to the distance matrix as *additive* (see [63, p. 145f.] and [38, p. 456]). In this case it can be proven that the phylogenetic tree and its weights are unique up to isomorphism (see [63, thm. 7.1.8, p. 148]). An even stronger property of dissimilarity maps is to be *ultrametric* (see [63, p. 149]). In this thesis we will not discuss more details of tree reconstruction by use of distance based methods (the interested reader may take a look at [63, chapt. 7] and [23, chapt. 11] and the references therein), but branch lengths and ultrametricity will be used in a slightly different way.

To view the edge weights as passed time between the speciation events is, as already mentioned, not the only useful interpretation. Viewing the edge weights as *substitution probability* results in a probability model for the evolution of the characters. For a binary character extension $\overline{\chi}$ on $\mathcal{T}$, let $\overline{\chi}(\rho) = \alpha$ and let $0 < p_e < \frac{1}{2}$ be the weight for each edge $e = (u, v)$. Then p_e is the probability that $\overline{\chi}(u) \neq \overline{\chi}(v)$ and every vertex' state $\overline{\chi}(v)$ depends only on the state $\overline{\chi}(u)$ of its parent vertex u. Hence, the character states for each character evolve from ρ to every leaf of the tree according to a Markov process. The probability that a specific $\overline{\chi}$ evolves is given by multiplication of the according probabilities p_e or $1 - p_e$ for all edges e in the tree. Also in a more general setting one can define a tree metric by use of the joint probability that $\overline{\chi}(x) = c_1$ and $\overline{\chi}(y) = c_2$ holds for some character states $c_1, c_2 \in C$ and species $x, y \in X$. Therefore a unique phylogenetic tree is given by these probabilities (see comment above and [63, p. 192f.]).

With help of such a probability model it is also possible to analyze the correctness of tree reconstruction by maximum parsimony or distance matrix methods. When we reconstruct the parent character states of a phylogenetic tree from given character states of the leaves by use of the previously explained Fitch-Hartigan algorithm, it is interesting how accurate this reconstruction works. We define the *reconstruction accuracy* for a certain phylogenetic tree as the probability that the Fitch-Hartigan algorithm reconstructs the character state of the vertex ρ correctly for a random character under the described probability model. This means that a character extension

(see Semple and Steel [63, p. 8]).

$\overline{\chi}$ evolves under the given probability model and the Fitch-Hartigan algorithm outputs the state $\overline{\chi}(\rho)$ or a set containing $\overline{\chi}(\rho)$ and the state $\overline{\chi}(\rho)$ is chosen by a uniform distribution. Surprisingly, it turns out that for a fixed phylogenetic tree the reconstruction accuracy may become greater if the Fitch-Hartigan algorithm operates only on a subtree. So for certain phylogenetic trees the chance of correct reconstruction of the root state increases if information about the species is left out.

Probability models will be developed in detail in Section 2.3, the Fitch-Hartigan algorithm and its reconstruction accuracy are covered in Chapter 4.

Methods of rooting a phylogenetic tree. Often an unrooted phylogenetic tree is reconstructed first from the given data. But it is also of interest to find the correct position of the root vertex. This procedure is referred to as *rooting the tree* or *directing* (see [28]). One way to achieve this is the *outgroup comparison method* (also *outgroup criterion*, see e.g. [23, p. 6f.] and [75, p. 1f.]). If it is known that one species $x \in X$ (the *outgroup*) differs greatly from the others $X \setminus \{x\}$ although the species in $X \setminus \{x\}$ are closely related to each other, it can be assumed that the differences are caused by the speciation event at the branching point of x and $X \setminus \{x\}$. Then the neighbor of x is the root vertex of the tree and the species in $X \setminus \{x\}$ are located in the other subtree of the root.

Another method, called *midpoint rooting*, makes use of the assumption of a molecular clock as mentioned above (see [23, p. 6f.] and [46]). If it is assumed that the distance between the root and a leaf is equal for every leaf, the root is located at the middle between two leaves.

Other applications. There are many parallels to other scientific fields, and therefore results can be applied not only in evolutionary biology but also elsewhere. For example Hartigan [43, p. 53] writes

> “Evolutionary models are used in the classification of plant and animal life, languages, motor cars, cultures, religions.”

Buneman [7] mentioned in his important paper from 1971 that “a similar situation occurs when one has a set of manuscripts all directly or indirectly copied from a common original manuscript.” Errors occur during the process of copying and therefore one can define a distance between two manuscripts by counting the number of differences. Spencer et al. [67] used the copying history of artificially created manuscripts to compare and study the accuracy and appropriateness of different tree reconstruction algorithms. Methods applied in phylogenetics are used also in

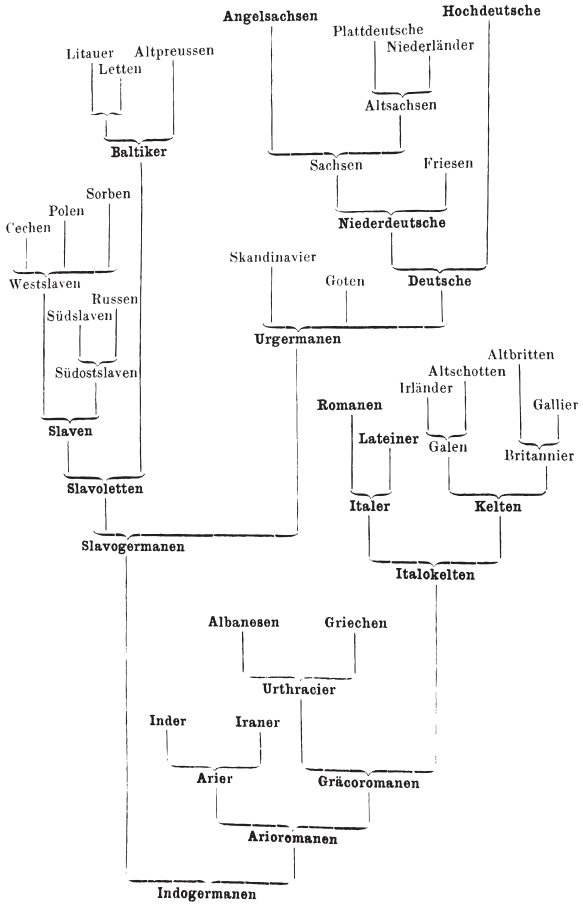


Figure 1.8: *Genealogical tree of the Indo-Germanic languages* by Haeckel [39, p. 360].

"archaeological artifacts, written works such as chain letters and medieval manuscripts" (see [67]). Already Darwin and Ghiselin [16] mentioned the parallels between the evolution of species and the development of languages:

> "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. [...] Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters."

Recent studies of parallels between phylogenetics and historical linguistics can be found in [78, 4].

Perspective. One of the main problems in phylogenetics is how to infer the evolutionary history of a group of species. The most popular algorithms for solving this problem belong to one of the three categories *maximum parsimony* (MP), *distance matrix,* or probability based methods such as *maximum likelihood* or *Bayesian methods*. Besides the reconstruction of phylogenetic trees there are also other problems of interest. For example, *phylogenetic diversity* is a measure for the heterogeneousness of a group of species. To find a subset of species of a certain size which maximizes the expected phylogenetic diversity, is called the *Noah's Ark Problem* (see [37, chapt. 6]).

Results in phylogenetics are not only interesting for biological applications, but also from a mathematical point of view. Székely et al. [75, p. 6] write about a "unified technique to solve a number of tree enumeration problems":

> "Had not we seen counting of trees with unlabelled branching vertices in biomathematics, we would hardly have ever come to this point."

In this thesis we concentrate on combinatorial problems in phylogenetics. Different types of phylogenetic trees, certain aspects of maximum parsimony, and probability models will be discussed. Other topics concerning the reconstruction of phylogenetic trees, such as distance matrix methods, will be treated marginally only. In Chapter 3 several enumeration problems with connections to phylogenetics will be studied, such as counting trees, and in Chapter 4 a specific problem concerning the reconstruction of parent character states will be presented.

Chapter 2

Preliminaries

In Chapter 1 we gave a short historical overview and mentioned different ways to view phylogenetic trees. We outlined briefly what phylogenetic trees are and how they are used as a tool to understand evolution. In this chapter we introduce phylogenetic trees as mathematical objects in order to analyze them with formal methods. This chapter lays the foundation for the following chapters. We will define all fundamental objects and introduce the basic concepts used later. We assume that the reader is familiar with the basics of graph theory. (See [63, chapt. 1] for a brief introduction to graphs, specifically focusing on phylogenetic trees. A comprehensive treatment of this matter can be found e.g. in [41].) For phylogenetic trees and phylogenetics we mainly will use the notation of [63]. For concepts of combinatorics we primarily rely on [27]. Additionally, we shortly present the main topics in Section 2.4.

2.1 Phylogenetic trees, X-trees and characters

We start with a set X (with $|X| = n$, $n \geq 2$) of distinct species, and typically we are interested in their evolutionary history, which is represented by a phylogenetic tree. In the following the species in X will be denoted by $1, 2, 3, \ldots, n$.

When we speak of a *tree*, we refer to a tree in the sense of graph theory, that is a (finite) simple, connected, acyclic graph $T = (V, E)$ with vertex set V and edges $E \subseteq \{\{u, v\} | u, v \in V\}$. Sometimes we will denote the vertex set of a graph G by $V(G)$ and the set of edges by $E(G)$. Trees under consideration will be always *unordered*, sometimes also referred to as *non-plane trees* (see e.g. [6]), i.e. there is no order of the subtrees of a node—in contrary to plane trees (see e.g. [68, pp. 293–295]). Formally the trees are undirected graphs, but in the case of rooted trees we usually view the edges as directed away from the root and then alternatively denote edges by $(u, v) \in E$ instead of $\{u, v\} \in E$ to indicate that the vertex u is closer to the root than

v. This direction is given naturally. It is the chronological direction of the evolution of species in the tree. Often also *unrooted trees* are of interest, because most algorithms reconstructing phylogenetic trees do not answer the question which internal node should be marked as root of the tree [63, p. 20]. As already mentioned (Chapter 1 on page 12), the problem of rooting the tree can be analyzed separately. With "tree" we mean in the following an unrooted tree, but occasionally we will use the term "unrooted tree" to emphasize that no vertex is distinguished as root[1].

The following definitions are very common, they differ in publications at most in details—we follow here [63, chapt. 2].

Definition 2.1. Let X be a finite set and $T = (V, E)$ a tree without vertices of degree 2 and the leaves of which are labeled with elements of X so that there is a 1:1-correspondence between the set of leaves and X. The latter can be formalized with a bijective map $\phi : X \to L$ from the set X to the set L of leaves of T. We then call the pair $\mathcal{T} = (T, \phi)$ a *phylogenetic tree.* A *rooted phylogenetic tree* is a pair $\mathcal{T} = (T, \phi)$, where T is a rooted tree with root vertex ρ of degree[2] at least 2, all other vertices are not of degree 2, and $\phi : X \to L$ is a bijective map from the set X to the set of leaves of T.

Phylogenetic trees are often called also *evolutionary trees* (e.g. in [75]). Also the terms *cladogram* (see e.g. [63, p. 20]) and *dendrogram* (see e.g. [52]) are used as synonyms.

If $\mathcal{T} = (T, \phi)$ is a phylogenetic tree, we refer to T as the *underlying tree* of $\mathcal{T}$ or the *structure* or the *tree shape* of $\mathcal{T}$. But these terms are probably self-explanatory, when they are used. If the underlying tree T is a binary tree[3], we call $\mathcal{T}$ also a *binary phylogenetic tree* (in biology sometimes called *bifurcating tree*, e.g. in [23]). In contrast, we sometimes speak of a *multifurcating phylogenetic tree* to emphasize the fact that T is not necessarily a binary tree. A phylogenetic tree which is not binary usually represents the situation where the chronological order of some speciation events is unclear, or they are too close to each other to be distinguishable. Therefore these events are represented by one single vertex in the tree. However, in biology it is usually assumed that a binary phylogenetic tree is a good representation of the evolutionary history of species (see [67]).

[1] While this is common terminology in graph theory (see e.g. [41]), computer scientists mostly consider rooted trees and hence in computer science the term "binary tree" usually refers to a rooted tree (see e.g. [5]).

[2] Also in the case of rooted trees we define the *degree* of a vertex as the number of its adjacent vertices and not as the number of its child vertices. The latter is often called *outdegree* in the case of rooted trees.

[3] A tree is called *binary* if every internal vertex has degree 3, or in case of a rooted tree if the root has degree 2 or degree 0 and every other internal vertex has degree 3.

The following definition generalizes phylogenetic trees. At the first glance the definition might appear somehow arbitrary, but it turns out that such trees can be characterized by a set of bipartitions of X, called X-*splits*, which is a nice tool to prove certain properties. More details can be found in [63, chapt. 3].

Definition 2.2. Let X be a finite set, $T = (V, E)$ a tree, and $\phi : X \to V$ a map with $v \in \phi(X)$ for every vertex $v \in V$ of degree 1 or 2. The pair $\mathcal{T} = (T, \phi)$ is called X*-tree*. If T is a rooted tree with root vertex ρ and $v \in \phi(X)$ for every vertex $v \in V \backslash \{\rho\}$ of degree 1 or 2, then $\mathcal{T} = (T, \phi)$ is called *rooted* X*-tree.*

To simplify notation we may write $V(\mathcal{T})$ to denote the set of vertices V of T and $E(\mathcal{T})$ to denote the set of edges. Note that also multiple labels can be assigned to one vertex of the tree, and some vertices may not be labeled at all. This can be interpreted in the following way. If two or more species in X cannot be distinguished by the given data, they are assigned to the same vertex of the tree. Vertices without label represent hypothetical ancestral species not known a priori, which are not in X. A vertex of degree 1 or 2 without label would not give additional information about the evolutionary history of the species in X. So we can restrict our attention to trees where all leaves and all vertices of degree 2 are labeled (see [30, p. 187]). However, in addition to these interpretations also technical reasons justify this definition of X-trees where vertices of degree 2 are allowed if they are labeled. As mentioned above, in this way X-trees correspond to so-called X-splits. In the following we will not need X-splits and therefore only refer to [63, capt. 3].

A phylogenetic tree $\mathcal{T} = (T, \phi)$ is an X-tree with the additional properties that the tree T has no vertices of degree 2 and ϕ is a bijective map to the set of leaves. So clearly X-trees generalize phylogenetic trees and rooted X-trees generalize rooted phylogenetic trees allowing also vertices of degree 2 if they are labeled, multiple labels and labels for any vertex in V.

Often a problem can be solved by dividing it into smaller problems, each of them being similar to the original problem. In the field of computer science this is often referred to as the *divide and conquer* principle. But in the same way it can be helpful in induction proofs or for enumeration problems. When dealing with rooted phylogenetic trees, this principle is usually applied by decomposing the tree into its subtrees rooted at the child nodes of the root, as illustrated in Figure 2.2. As in [63, p. 21] we want to call this procedure the *standard decomposition* of $\mathcal{T}$. We will use the standard decomposition e.g. for the Fitch-Hartigan algorithm (see Theorem 2.9), Felsenstein's recurrence relation in Section 3.1.2 (see Theorem 3.7), and in the inductive proof of Theorem 4.5.

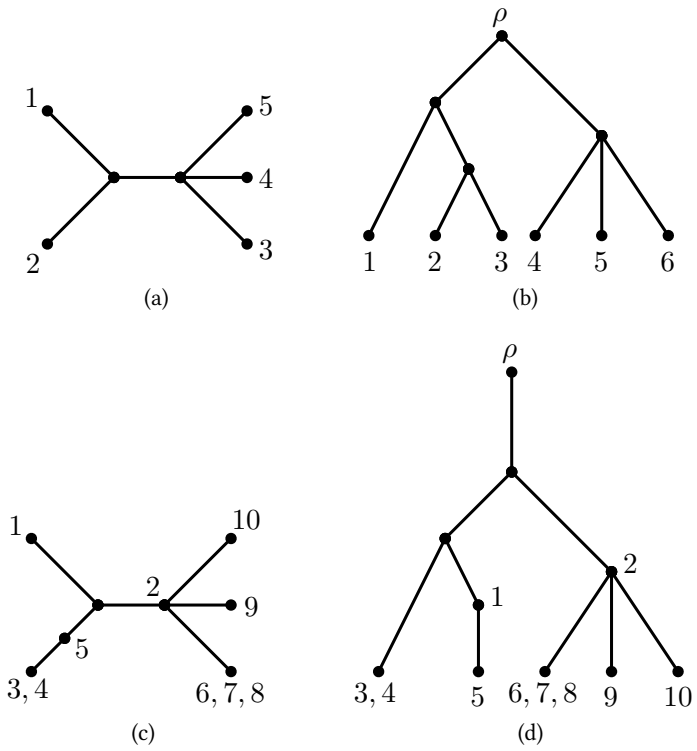


Figure 2.1: Example for an unrooted and a rooted phylogenetic tree and an unrooted and a rooted X-tree, where $X = \{1, 2, \ldots, 10\}$ is the label set of the X-trees.

Note that the term *subtree* can be interpreted ambiguously. If $T = (V, E)$ is a tree, a subtree $T' = (V', E')$ of T clearly should be a tree itself and fulfill $V' \subseteq V$, $E' \subseteq E$. But if T is a rooted tree and we mean the subtree including all vertices being descendant of a vertex $v \in V$, we speak of *the subtree rooted at* v. By *the subtrees of* T we mean the subtrees of T rooted at $v_1, v_2, \ldots, v_k$ as in Figure 2.2, where $v_1, v_2, \ldots, v_k$ are the child nodes of the root.

As already illustrated in Chapter 1, the given data about the species in X is typically represented by characters (see also [63, p. 65, p. 84]).

Definition 2.3. A *character* is a function $\chi : X \to C$, where C is a finite set of character states. For $|C| = r$ we also speak of a *r-state character*. If $\mathcal{T} = (T, \phi)$ is an X-tree with $T = (V, E)$, a *character extension* of the character χ is a function $\overline{\chi} : V \to C$ with $\overline{\chi} \circ \phi = \chi$.

Sometimes instead of character the term *leaf-coloration* is used (e.g. in [75, p. 3]). Usually we denote the character states with small letters of the Greek alphabet $C = \{\alpha, \beta, \gamma, \ldots\}$, and we will always denote character extensions with an overline to make clear that the function is

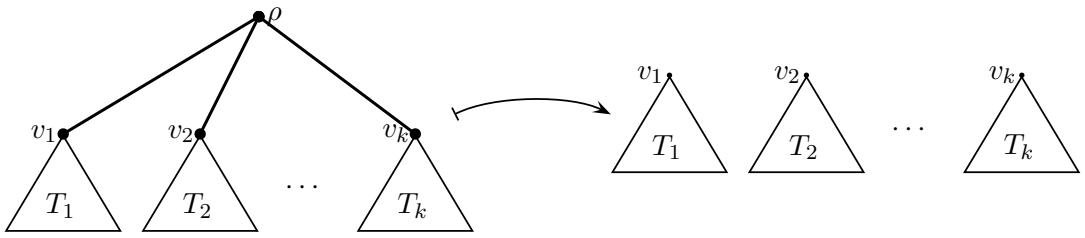


Figure 2.2: Illustration of the standard decomposition.

defined on the nodes of the tree and not on the set X. In the case of phylogenetic trees we can identify the leaves of the tree with the labels in X. Then a character extension $\overline{\chi}$ of a character χ is an extension in the sense that $\overline{\chi}|_L = \chi$ where L is the set of leaves of the tree.

If a character χ is not surjective, we have $r' := |\chi(X)| < |C| = r$ and then sometimes it makes more sense to speak of a r'-state character than of a r-state character (see [63, p. 66]). Nevertheless, in the following chapters, we will mostly refer to the number of *possible* states $|C|$ rather than to the number of *occurring* states $|\chi(X)|$ and speak of r-state characters when $r = |C|$. If statements can be misinterpreted, we will clarify what is meant. As mentioned before, a vertex with multiple labels is interpreted as aggregation of speciation events which are indistinguishable with respect to the given data. Therefore it makes sense to consider only characters with $\chi(x) = \chi(y)$ if $\phi(x) = \phi(y)$ for any $x, y \in X$. Hence, every vertex has not more than one character state, even if multiple labels are assigned (see [63, p. 84]). However, this is always the case with phylogenetic trees, and in the following characters will be considered only on phylogenetic trees (except for some basic properties concerning maximum parsimony in the next section).

2.2 Maximum parsimony

The *maximum parsimony method*, sometimes also called *minimum mutation problem* (e.g. in [38, pp. 472–474]), was already briefly illustrated in Chapter 1. It can be considered as the basic model used to infer phylogenetic trees from a set of characters. In this section the *parsimony score* of a set of characters on a phylogenetic tree will be defined as a measure for the evolutionary change of the phylogenetic tree and the characters under consideration. The most parsimonious tree in this sense is then the one involving the least evolutionary change, and it is hypothesized that it represents the correct reconstruction of the evolution for the given set of species. We will start with the formal definitions, following [63, chapt. 4–5], and then derive some properties and

present the Fitch-Hartigan algorithm, already mentioned in Chapter 1. At the end of this section we will shortly discuss the validity of the maximum parsimony approach as a model of evolution.

Definition 2.4 (Parsimony score). The *parsimony score* of a character χ on an X-tree $\mathcal{T} = (T, \phi)$ with $T = (V, E)$ is defined as the minimum number of edges connecting nodes with different character states for all possible character extensions $\overline{\chi}$ of χ

$$l(\chi, \mathcal{T}) := \min_{\overline{\chi}} |\{\{u, v\} \in E | \overline{\chi}(u) \neq \overline{\chi}(v)\}|.$$

A character extension $\overline{\chi}$ of χ with $l(\chi, \mathcal{T}) = |\{\{u, v\} \in E | \overline{\chi}(u) \neq \overline{\chi}(v)\}|$ is called a *minimum character extension*. The *parsimony score* of a set $\mathcal{C} = \{\chi_1, \chi_2, \ldots, \chi_k\}$ of characters on an X-tree is defined as the sum of the parsimony scores of all characters:

$$l(\mathcal{C}, \mathcal{T}) := \sum_{i=0}^{k} l(\chi_i, \mathcal{T}).$$

Sometimes it will simplify notation to define $\mathrm{ch}(\overline{\chi}) := |\{\{u, v\} \in E | \overline{\chi}(u) \neq \overline{\chi}(v)\}|$. We call this quantity the *changing number of* $\overline{\chi}$ (as in [63, p. 84] and [75, p. 3]). If there is more than one tree under consideration, it will be clear from the domain of $\overline{\chi}$ to which tree the changing number refers.

Trees $\mathcal{T}$ minimizing the quantity $l(\mathcal{C}, \mathcal{T})$ are called *maximum parsimony tree*s. These are the trees we are looking for if we want to reconstruct a phylogenetic tree from a set of characters by following the maximum parsimony approach. However, searching the space of phylogenetic trees for trees fulfilling the presented optimality criterion is not an easy task. There are simply too many different trees to examine all of them—also for considerably small sets of species as Section 3.1 will demonstrate thoroughly. Actually, Foulds and Graham [28] showed that finding a maximum parsimony tree for a set of characters $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$ is NP-complete also if all characters have only two states, i.e. $\chi_i : X \to C$ and $|C| = 2$ for $i = 1, \ldots, k$. They used a modified version of the *general Steiner tree problem*, which is known to be NP-complete (see [35]), to prove the NP-hardness of maximum parsimony. Maximum parsimony trees relate to *minimum Steiner trees*[4] as we will outline briefly in the following. If there are $x_1, x_2 \in X$ with $\chi_i(x_1) = \chi_i(x_2)$ for all $i = 1, \ldots, k$, there is a maximum parsimony tree, where $\phi(x_1)$ and $\phi(x_2)$ form a cherry and the problem of finding a maximum parsimony tree on X for the

[4] Let $G = (V, E)$, d a metric on V and $X \subseteq V$ a subset of V. A tree $T = (V', E')$ with $X \subseteq V' \subseteq V$ and $E' \subseteq E$ is a *minimum Steiner tree* for X in G, if the sum of all edge weights $\sum_{\{u,v\} \in E'} d(u, v)$ is minimal for all subtrees of G connecting all vertices in X (see [28] and [63, p. 97.f]).

characters in $\mathcal{C}$ is equivalent to the problem for finding a maximum parsimony tree on $X \setminus \{x_1\}$ for the characters $\chi|_{X\setminus\{x_1\}}$ in $\mathcal{C}$ restricted to $X \setminus \{x_1\}$. That means w.l.o.g we can assume that each $x \in X$ corresponds to a tuple $(\alpha_1, \ldots, \alpha_k) \in C^k$ of character states and this mapping is injective. Let $d\left(\vec{\alpha}, \vec{\beta}\right) := |\{i|1 \leq i \leq k, \alpha_i \neq \beta_i\}|$ be the Hamming distance on C^k. Then a minimum Steiner tree for the set $\{(\chi_1(x), \ldots, \chi_k(x))|x \in X\}$ of the complete graph[5] with vertex set C^k is a maximum parsimony tree for $\mathcal{C}$, if vertices of degree 2 are suppressed (see also [63, p. 97.f]).

Minimizing the number of evolutionary changes between the species is only one way to view maximum parsimony. As already mentioned in the introduction, biologists often assume characters to be *homoplasy-free*, which refers to the case that every character state does not arise more than once in the tree when the character evolves from the root of the tree (see [23, p. 25]). This can be formally defined also for unrooted trees (we follow here [63, p. 65]).

Definition 2.5. A character $\chi : X \to C$ is called *convex* on an X-tree $\mathcal{T} = (T, \phi)$ with $T = (V, E)$ if there is a character extension $\overline{\chi}$ with the property that the subgraph of T induced by $\{v \in V|\overline{\chi}(v) = \alpha\}$ is connected for each $\alpha \in C$.

The following proposition shows that every character state arises only once in case of a rooted X-tree and a convex character. The statement was inspired by the informal explanations in [23, chapt. 1] and proved within the scope of this thesis.

The statement in (ii) corresponds to the case where the state $\overline{\chi}(v)$ arises in vertex v for the first time and only there. Clearly if the statement in (i) is true, the character χ is convex on $\mathcal{T}$. Conversely if the character χ is convex, there exists a character extension such that (i) is true.

Proposition 2.6. *Let $\mathcal{T} = (T, \phi)$ be a rooted X-tree with $T = (V, E)$ and $\chi : X \to C$ a character on $\mathcal{T}$ and $\overline{\chi}$ a character extension. Then the following two statements are equivalent:*

(i) The subgraph of T induced by $\{v \in V|\overline{\chi}(v) = \alpha\}$ is connected for each $\alpha \in C$.

(ii) If $(u, v) \in E$ is an edge with u being closer to the root ρ than v and $\overline{\chi}(u) \neq \overline{\chi}(v)$, then $\overline{\chi}(\rho) \neq \overline{\chi}(v)$ and there is no other edge $(u', v') \in E$ with $\overline{\chi}(u') \neq \overline{\chi}(v')$, $\overline{\chi}(v) = \overline{\chi}(v')$ and $v \neq v'$.

Proof. First assume ¬(ii). So we have edges $(u, v), (u', v') \in E$ with $\overline{\chi}(u) \neq \overline{\chi}(v)$, $\overline{\chi}(u') \neq \overline{\chi}(v')$, $\overline{\chi}(v) = \overline{\chi}(v') =: \alpha$ and $v \neq v'$ or $(u, v) \in E$ with $\overline{\chi}(u) \neq \overline{\chi}(v)$ and $\overline{\chi}(\rho) = \overline{\chi}(v)$. First, consider the former case. At least one of the two nodes u, u' is on the unique path between v

[5] $G = (V, E)$ is a complete graph, if every two vertices $v_1, v_2 \in V, v_1 \neq v_2$ are adjacent $\{v_1, v_2\} \in E$.

and v' (otherwise u and u' could not be closer to ρ than v and v'). Since $\overline{\chi}(u) \neq \alpha, \overline{\chi}(u') \neq \alpha$ both nodes u and u' are not members of the subgraph of T induced by $\{v \in V | \overline{\chi}(v) = \alpha\}$ and therefore it is not connected (there are at least two unconnected components, one containing v and the other v' and the unique path in T connecting v and v' is interrupted at u or u'). Now, if $(u, v) \in E$ with $\overline{\chi}(u) \neq \overline{\chi}(v) =: \alpha$ and $\overline{\chi}(\rho) = \overline{\chi}(v)$, again the subgraph of T induced by $\{v \in V | \overline{\chi}(v) = \alpha\}$ is not connected because the unique path in T from ρ to v is interrupted at u in the subgraph. Thus, we have ¬(ii)⇒ ¬(i).

For the other direction assume ¬(i). So we have a character state $\alpha \in C$ where the subgraph of T induced by $\{v \in V | \overline{\chi}(v) = \alpha\}$ consists of at least two unconnected components. Choose v as the vertex closest to ρ in one of these components and v' closest to ρ in another component. So clearly $v \neq v'$. If $v \neq \rho, v' \neq \rho$, there are edges (u, v) and (u', v') with $\overline{\chi}(u) \neq \alpha$ and $\overline{\chi}(u') \neq \alpha$ (otherwise u or u' would be closer to ρ and member of the component). Now consider w.l.o.g. the case $v = \rho$. Then $v' \neq \rho$, and there is an edge (u', v') with $\overline{\chi}(u') \neq \alpha$ and $\overline{\chi}(\rho) = \overline{\chi}(v)$. Hence, we have ¬(i)⇒ ¬(ii). □

Convexity of characters can be characterized also directly without relating to a character extension as shown in the following proposition (both the statement and its proof are taken from [63, Proposition 4.1.3, p. 66]). Figure 2.3 shows an example.

Proposition 2.7. *Let $\chi : X \to C$ be a character on an X-tree $\mathcal{T} = (T, \phi)$ with $T = (V, E)$. Then the following two statements are equivalent:*

(i) χ is convex on $\mathcal{T}$.

(ii) The members of $\{T(\alpha) | \alpha \in C\}$ are pairwise vertex disjoint where $T(\alpha)$ denotes the subtree of T induced by the vertices in $\phi\big(\chi^{-1}(\{\alpha\})\big)$.

Proof. First assume (i). So there is a character extension $\overline{\chi}$ with the property that the subgraph of T induced by $\{v \in V | \overline{\chi}(v) = \alpha\}$ is connected for each $\alpha \in C$ and furthermore also acyclic since T is a tree. Clearly we have $\phi\big(\chi^{-1}(\alpha)\big) \subseteq \{v \in V | \overline{\chi}(v) = \alpha\}$ and therefore $T(\alpha)$ is a subtree of the subgraph of T induced by $\{v \in V | \overline{\chi}(v) = \alpha\}$. For $\alpha_1, \alpha_2 \in C$ with $\alpha_1 \neq \alpha_2$ the sets $\{v \in V | \overline{\chi}(v) = \alpha_1\}$ and $\{v \in V | \overline{\chi}(v) = \alpha_2\}$ are disjoint and therefore $T(\alpha_1)$ and $T(\alpha_2)$ are vertex disjoint.

Conversely, suppose (ii) is true. We have to construct an appropriate character extension $\overline{\chi}$. For any state $\alpha \in C$ and all vertices $v \in T(\alpha)$ we have to set $\overline{\chi}(v) := \alpha$. For the remaining vertices, there are several possible choices of values for $\overline{\chi}$. Let F be the subgraph of T induced

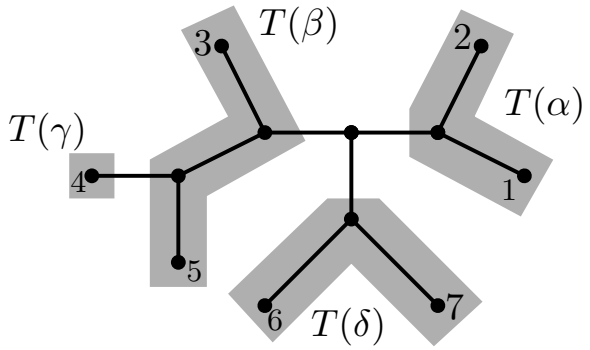


Figure 2.3: χ is a character on $\mathcal{T}$ with $\chi(1) = \alpha$, $\chi(2) = \alpha$, $\chi(3) = \beta$, $\chi(4) = \gamma$, $\chi(5) = \beta$, $\chi(6) = \delta$, $\chi(7) = \delta$. The subtrees $T(\alpha)$, $T(\beta)$, $T(\gamma)$, $T(\delta)$ are illustrated with gray background. The example and the figure are from [63, p. 67], it illustrates the setting in Proposition 2.7.

by $\{v \in V | \forall \alpha \in C : v \notin V(T(\alpha))\}$. Now for each component of F choose any arbitrary vertex $u \notin V(F)$ which is adjacent to a vertex of the component. Then define for all nodes v of the component $\overline{\chi}(v) := \overline{\chi}(u)$ (note that $u \in T(\alpha)$ for some $\alpha \in C$). The function $\overline{\chi}$ is now defined on all vertices $v \in V$ and clearly a character extension of χ. Furthermore for $\alpha \in C$ the graph $T(\alpha)$ is connected by definition and other vertices with $\overline{\chi}(v) = \alpha$ are in F and connected to $T(\alpha)$ by definition. Hence, χ is convex. □

Now we are going to show the relation between convex characters and the parsimony of characters. Convex characters are exactly the ones with the least possible amount of evolutionary change for the given number of occurring character states. In general it is not always possible to find a phylogenetic tree so that any given set of characters is homoplasy-free, but we can view the maximum parsimony approach also as strategy to reduce homoplasy as much as possible. The following proposition and its proof are from [63, p. 85].

Proposition 2.8. *Let χ be a character on X with $r = |\chi(X)|$ and $\mathcal{T} = (T, \phi)$ an X-tree. Then,*

$$l(\chi, \mathcal{T}) \geq r - 1,$$

where $l(\chi, \mathcal{T}) = r - 1$ if and only if χ is convex on $\mathcal{T}$.

Proof. Let $\overline{\chi}$ be a minimum character extension for χ on $\mathcal{T}$ and T' the tree obtained from T by contracting all edges in $E(T) \setminus \{\{u, v\} \in E(T) | \overline{\chi}(u) \neq \overline{\chi}(v)\}$. Since T' is a tree, we have $|V(T')| - 1 = |E(T')|$ and by definition of T' also $|E(T')| = |\{\{u, v\} \in E(T) | \overline{\chi}(u) \neq \overline{\chi}(v)\}|$.

We chose $\overline{\chi}$ as minimum character extension, thus $l(\chi, \mathcal{T}) = |\{\{u, v\} \in E(T) | \overline{\chi}(u) \neq \overline{\chi}(v)\}|$. In total this yields

$$l(\chi, \mathcal{T}) = |\{\{u, v\} \in E(T) | \overline{\chi}(u) \neq \overline{\chi}(v)\}| = \left|E\left(T'\right)\right| = \left|V\left(T'\right)\right| - 1.$$

Now consider the map $\overline{\chi}|_{V(T')}$ to derive the claimed equality and inequality. The function $\overline{\chi}|_{V(T')}$ is well defined, because each node $v \in V(T')$ corresponds to nodes $v_1, v_2, \ldots, v_k \in V(T)$ for some $k \geq 1$ and $\overline{\chi}(v_i) = \overline{\chi}(v_j)$ for all $1 \leq i, j \leq k$. In addition the cardinality of the image of $\overline{\chi}|_{V(T')}$ equals r, because $\overline{\chi}$ is a minimum character extension and we did not remove or add any character states in the construction of T'. Therefore we have $|V(T')| \geq r$.

Furthermore the map $\overline{\chi}|_{V(T')}$ is injective if and only if the subgraph of T induced by the set $\{v \in V | \overline{\chi}(v) = \alpha\}$ is connected for each $\alpha \in \chi(X)$. Therefore the equality $|V(T')| = r$ holds if and only if χ is convex on $\mathcal{T}$. □

Fitch-Hartigan algorithm. As earlier mentioned, it is not easy to find a phylogenetic tree which minimizes the parsimony score for a given character. And, if the phylogenetic tree and a character are given, still, the number of possible character extensions is exponential in the number of internal nodes of the tree. Nevertheless, this problem of reconstructing parent character states can be solved efficiently also for large sets of species and without examining every possible character extension. In the following we are going to present the *Fitch-Hartigan algorithm*, which reconstructs a set of possible character states for all parent nodes in a rooted phylogenetic tree. In a second pass traversing the tree from the root towards the leaves, a minimum character extension can be obtained. So, the algorithm provides a reconstruction of the ancestral character states for all tree nodes following the maximum parsimony approach. The algorithm has a runtime in $O(|C| \cdot |X|)$ (see [63, p. 90]) while $|C|^{|V(\mathcal{T})|-|X|}$ is the number of all possible character extensions for a character $\chi : X \to C$ on a phylogenetic tree $\mathcal{T}$ with a set of species X.

A version for binary phylogenetic trees was first published by Fitch [26] in 1971 while Hartigan [43] developed independently a generalized version providing also a proof for the correctness of the algorithm (other resources include [23, pp. 11–13], [38, p. 478f.] and [49, p. 648]). We follow here the version by Semple and Steel [63, pp. 89–91] to describe the algorithm formally and use the main ideas by Hartigan [43] for a proof strongly adopted to our setting[6].

[6] Strangely none of the authors except Hartigan [43] explain the correctness of the algorithm in detail. While Semple and Steel [63, p. 90] and Gusfield [38, p. 478f.] leave the proof as an exercise, Felsenstein [23, pp. 11–13] encourages to glance at a figure, but also admits that "on larger trees the moment's glance will not work, but the [...] algorithm will continue to work". Also at first glance one gets a rough feeling that the algorithm does what it claims to do,

Theorem 2.9 (Fitch-Hartigan algorithm). *Let $\mathcal{T} = (T, \phi)$ be a rooted phylogenetic tree with $T = (V, E)$, root vertex ρ and label set X and χ a character on $\mathcal{T}$. Define the maps $\psi : V \to 2^C \setminus \{\emptyset\}$ and $l : V \to \mathbb{N}$ for all vertices $v \in V$ recursively as follows:*

(i) If v is a leaf, set $\psi(v) := \{\chi(\phi^{-1}(v))\}$ and $l(v) := 0$.

(ii) Denote the child vertices of v by $v_1, v_2, \ldots, v_k$ and suppose that $\psi(v_i)$ and $l(v_i)$ have been defined for $i = 1, 2, \ldots, k$. Furthermore, we use the function $f : V \to \mathbb{N}^{>0}$ to simplify notation and define

$$f(v) := \max_{\alpha \in C} |\{v_i | 1 \leq i \leq k, \alpha \in \psi(v_i)\}|.$$

Now, let $\psi(v)$ be the set of character states occurring $f(v)$ times in the sets $\psi(v_1), \ldots, \psi(v_k)$

$$\psi(v) := \{\alpha \in C | f(v) = |\{v_i | 1 \leq i \leq k, \alpha \in \psi(v_i)\}|\}$$

and

$$l(v) := k - f(v) + \sum_{i=1}^{k} l(v_i).$$

Then the parsimony score of χ is determined by $l(\chi, \mathcal{T}) = l(\rho)$ and $\psi(\rho)$ is the set of reconstructed character states for ρ, i.e.

$$\psi(\rho) = \{\alpha \in C | \text{there is a minimum extension } \overline{\chi} \text{ of } \chi \text{ with } \overline{\chi}(\rho) = \alpha\}. \tag{2.1}$$

Furthermore an explicit minimum character extension can be constructed by a subsequent backward pass (see [38, p. 478f.]) defining the character states $\overline{\chi}(v)$ for all $v \in V$ now starting at ρ traversing towards the leaves as follows.

(iii) If v is the root vertex choose $\overline{\chi}(v)$ arbitrary from $\psi(v)$.

(iv) Denote the child vertices of v by $v_1, v_2, \ldots, v_k$ and suppose that $\overline{\chi}(v)$ has been defined. For $i = 1, 2, \ldots, k$ define

$$\overline{\chi}(v_i) := \begin{cases} \overline{\chi}(v), & \text{if } \overline{\chi}(v) \in \psi(v_i), \\ \text{any arbitrary } \alpha \in \psi(v_i), & \text{otherwise.} \end{cases}$$

but it is not that obvious why the obtained character extension is minimal.

Proof. We are going to use the standard decomposition (see Figure 2.2) to prove the statements by induction. Clearly all statements are true if $|V| = 1$. Now let $\mathcal{T}$ be a rooted phylogenetic tree with root ρ and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ its k subtrees with roots $y_1, y_2, \ldots, y_k$ as in Figure 2.2. As induction hypothesis we assume for $i = 1, 2, \ldots, k$ that $\psi(y_i)$ is the set of character states $\alpha \in C$, such that there is a minimum character extension $\overline{\chi}_i$ on $\mathcal{T}_i$ with $\overline{\chi}_i(y_i) = \alpha$ as in (2.1). Let $\overline{\chi}$ be a minimum character extension on $\mathcal{T}$. Now consider two cases for each $i \in \{1, 2, \ldots, k\}$ in order to determine the changing number of $\overline{\chi}$ restricted to the subtree of $\mathcal{T}$ induced by $V(\mathcal{T}_i) \cup \{\rho\}$ (that is the subtree $\mathcal{T}_i$ together with the edge $\{\rho, y_i\}$).

Case 1. If $\overline{\chi}(\rho) \in \psi(y_i)$, there exists a minimum character extension $\overline{\chi}_i$ of $\chi|_{V(\mathcal{T}_i)}$ on $\mathcal{T}_i$ with $\overline{\chi}_i(y_i) = \overline{\chi}(\rho)$. This means that $\overline{\chi}|_{V(\mathcal{T}_i)}$ is also a minimum character extension on $\mathcal{T}_i$. Otherwise we would have $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)}\big) > \mathrm{ch}(\overline{\chi}_i)$ which is a contradiction to $\overline{\chi}$ being a minimum character extension on $\mathcal{T}$. Hence, we have $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)}\big) = l\big(\chi|_{V(\mathcal{T}_i)}, \mathcal{T}_i\big)$ and by use of the induction hypothesis $l\big(\chi|_{V(\mathcal{T}_i)}, \mathcal{T}_i\big) = l(y_i)$ and because of $\overline{\chi}(y_i) = \overline{\chi}(\rho)$ we get $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) = l(y_i)$.

Case 2. On the other hand, if $\overline{\chi}(\rho) \notin \psi(y_i)$, then either $\overline{\chi}|_{V(\mathcal{T}_i)}$ is a minimum character extension on $\mathcal{T}_i$ and $\overline{\chi}(y_i) \neq \overline{\chi}(\rho)$ (let this be *Case 2a*) or $\overline{\chi}|_{V(\mathcal{T}_i)}$ is not a minimum character extension on $\mathcal{T}_i$ (let this be *Case 2b*). In the former case (Case 2a) $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)}\big) = l\big(\chi|_{V(\mathcal{T}_i)}, \mathcal{T}_i\big) = l(y_i)$, as previously, is the changing number of $\overline{\chi}$ restricted to $\mathcal{T}_i$ but also $\overline{\chi}(y_i) \neq \overline{\chi}(\rho)$. Thus in Case 2a we have in total $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) = l(y_i) + 1$. In Case 2b $\overline{\chi}|_{V(\mathcal{T}_i)}$ is not a minimum character extension on $\mathcal{T}_i$ and therefore $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)}\big) > l(y_i)$. But at the same time $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) \leq l(y_i) + 1$, because otherwise this would be a contradiction to $\overline{\chi}$ being a minimum character extension on $\mathcal{T}$ (replacing the values of $\overline{\chi}|_{V(\mathcal{T}_i)}$ by the values of some minimum character extension $\overline{\chi}_i$ on $\mathcal{T}_i$ would lead to a character extension on $\mathcal{T}$ with a smaller changing number). Therefore the only possibility is that $\overline{\chi}(\rho) = \overline{\chi}(y_i)$ and $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)}\big) = l(y_i) + 1$. Thus, in Case 2b we get also in total $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) = l(y_i) + 1$ (the example in Figure 2.4b illustrates Case 2b).

We chose $\overline{\chi}$ as minimum character extension and therefore the parsimony score is given by $l(\chi, \mathcal{T}) = \mathrm{ch}(\overline{\chi})$. But we can count the edges $\{u, v\}$ with $\overline{\chi}(u) \neq \overline{\chi}(v)$ also separately for each subtree and get $l(\chi, \mathcal{T}) = \sum_{i=1}^{k} \mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big)$. Clearly $\overline{\chi}$ is a minimum character extension if and only if the number of $i \in \{1, \ldots, k\}$, such that Case 1 occurs, is maximal for all possible choices of $\overline{\chi}(\rho) \in C$. This number equals $f(\rho)$ as defined in Step (ii) of the algorithm and the set $\psi(\rho)$ contains exactly the character states $\overline{\chi}(\rho)$ for any arbitrary minimum character extension $\overline{\chi}$ on $\mathcal{T}$. Furthermore Case 1 occurs exactly $f(\rho)$ times and each time we have $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) = l(y_i)$. Hence, Case 2 occurs $k - f(\rho)$ times with $\mathrm{ch}\big(\overline{\chi}|_{V(\mathcal{T}_i)\cup\{\rho\}}\big) = l(y_i) + 1$ each time. In total

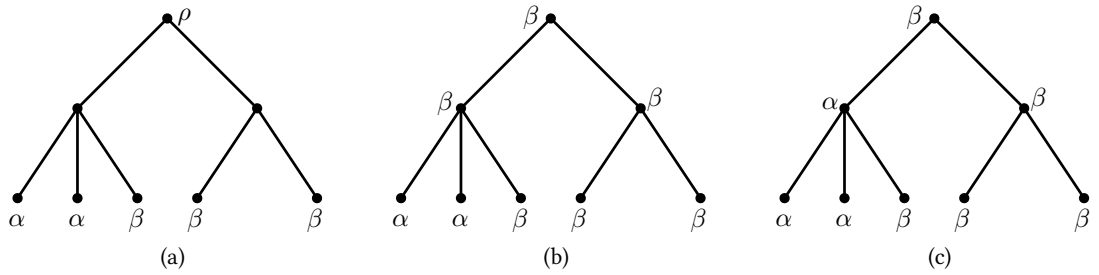


Figure 2.4: A rooted phylogenetic tree and a character with the character states α and β (in Figure 2.4a) and two of its minimum character extensions (in Figure 2.4b and Figure 2.4c). The character extension in Figure 2.4b is an example for Case 2b in the proof of Theorem 2.9.

this yields $l(\rho) = k - f(\rho) + \sum_{i=1}^{k} l(y_i)$. The backward pass chooses any arbitrary $\alpha \in \psi(\rho)$ and then decides for Case 1 or Case 2a. As we have shown, there is always a minimum character extension $\overline{\chi}$ on $\mathcal{T}$ which can be constructed in this way. □

Note that the algorithm cannot construct every possible minimum character extension. Hartigan [43] describes a slightly enhanced version of the algorithm, where the algorithm remembers also for each node v the set of states α, such that a character extension $\overline{\chi}_v$ on the subtree rooted at v exists where $\overline{\chi}_v(v) = \alpha$ and $\mathrm{ch}(\overline{\chi}_v) = l(v) + 1$. In this way also the minimum character extensions can be constructed, where Case 2b occurs.

If $\psi(\rho) = \{\alpha\}$, we say that the state α was *reconstructed unambiguously*, and otherwise if $|\psi(\rho)| > 1$, we say that the state was *reconstructed ambiguously* (see [24, 49]).

Remark 2.10. Note that Step (ii) in Theorem 2.9 can be described in a simpler way if $\mathcal{T}$ is a rooted binary phylogenetic tree (see [49, p. 648] and [70]). Let v be an internal node and v_1, v_2 its child vertices. Then we have either $f(v) = 1$ or $f(v) = 2$. The former is the case if every state occurs at maximum once in the sets $\psi(v_1)$ and $\psi(v_2)$. Hence, the two sets are disjoint and all states occurring once are given by the union of the two sets. On the other hand, if $f(v) = 2$, there is at least one state α with $\alpha \in \psi(v_1)$ and $\alpha \in \psi(v_2)$. Therefore the two sets are not disjoint and all states occurring in both sets are given by the intersection. Hence, in total the set of reconstructed character states for v can be defined by

$$\psi(v) = \begin{cases} \psi(v_1) \cup \psi(v_2), & \text{if } \psi(v_1) \cap \psi(v_2) = \emptyset \\ \psi(v_1) \cap \psi(v_2), & \text{otherwise.} \end{cases}$$

Validity of maximum parsimony as model of evolution. In 1965 Camin and Sokal [8] published a study about phylogenetic tree reconstruction based on an analysis of the evolution of artificial organisms (see [23, p. 129f.]). Joseph Camin let evolve imaginary species (by drawing them on paper) "according to rules known so far only to him, but which are believed to be consistent with what is generally known of transspecific evolution" [8, p. 311]. A set of resulting characters was then handed to students and systematists. The most accurate reconstruction of the known evolutionary history of the imaginary species was achieved by the students who minimized the number of changes of the character states (see [23, p. 129f.]). A similar study using the copying history of artificially created manuscripts was done by Spencer et al. [67]. The reconstruction of the (known) copying tree with a maximum parsimony approach was "reasonably accurate". Also a reconstruction of "a set of manuscript copies (text tradition) of an Icelandic saga", where some sources of the manuscripts are known, matched closely these known relationships (see [67, p. 503]).

Often the maximum parsimony approach is also justified by Ockham's Razor, i.e. that a more parsimonious phylogenetic tree is a simpler explanation and therefore should be preferred compared to a more complex one (see [63, p. 84]). However, many authors disagree with this reasoning and suggest a statistical approach (see [23, p. 138ff.] and [65] for a discussion of the different points of view in this debate).

Moreover, it is not even guaranteed that the evolution of species is treelike. For example Nakhleh et al. [53] write:

> "...yet it is widely understood and accepted that trees oversimplify the evolutionary histories of many groups of organisms, most prominently bacteria (because of horizontal gene transfer) and plants (because of hybrid speciation)."

Phenomenons such as *hybrid speciation* and *horizontal gene transfer* cannot be represented in phylogenetic trees, but they occur in nature (see [74, 57]). Hybrid speciation refers to the situation, where a species has two ancestral species, and horizontal gene transfer describes the situation, where genetic material moves from one species to another one without mating or cell division. In these cases one can use sets of phylogenetic trees or certain digraphs, such as *phylogenetic networks*, to model the evolutionary history (see also [63, sect. 2.7], [53] and [36, chapt. 7]).

2.3 Markov models on phylogenetic trees

Besides distance based models, mentioned briefly in Chapter 1, and the maximum parsimony approach, presented in the previous section, a third approach are stochastic models, which will be introduced in this section. The maximum parsimony method makes use of certain properties, which are assumed for the evolution of the traits under consideration. Now we are going to model how a trait might evolve if we know the state at the root vertex and assume probabilities for the events that one character state changes to another one. It might be e.g. more likely that a specific trait arises, if its ancestor exhibits a certain other trait. If we compare the character states of a species and one of its descendants under the so-called N_r-model, the most likely result will be that no change happened. As we will explain later in more detail, this is referred to as *conservation* and the probability that it happens as *conservation probability* (see e.g. [24]). A probability model with certain probabilities for a specific tree induces a probability distribution on the set of all possible characters $\chi : X \to C$ as we will explain in detail in this section. But typically a character is given and we are interested in the correct reconstruction of the phylogenetic tree. For a given set of characters $\mathcal{C}$, the method of *maximum likelihood* assumes $\mathcal{T}$ to be the correctly reconstructed tree if the probability that the characters in $\mathcal{C}$ evolve on $\mathcal{T}$ is maximal for all phylogenetic trees. However, no general algorithm is known to find these optimal trees and therefore different heuristics are used (see [63, sect. 8.9] for details).

A probability model can be used to justify and analyze other methods such as maximum parsimony—but often it shows also their limits. If a character is assumed to evolve according to a certain probability model, we can determine the probability that an algorithm correctly reconstructs the original tree from a random set of characters. For example, Felsenstein [22] showed that even for a set of only four species $X = \{1, 2, 3, 4\}$ the maximum parsimony might be "positively misleading". Considering a set of random binary characters $\mathcal{C} = \{\chi_1, \chi_2, \ldots, \chi_k\}$ with certain parameters for the stochastic model, the probability of reconstructing the correct tree converges to 0 as $k \to \infty$ (see also [63, sect. 8.7, pp. 202–204]). Of course, this is an undesired property of a reconstruction method, often referred to as the *statistical inconsistency* of maximum parsimony. Similarly, we will discuss certain properties of character states reconstructed by the Fitch-Hartigan algorithm in Chapter 4. Other aspects of links between maximum parsimony and maximum likelihood are discussed in [25, 71, 76].

This section shall explain only how such a model usually is defined ([63, chapt. 8] provides some more details and a good overview).

In the following we will denote by $\mathbb{P}(\xi = a)$ the probability that a random variable ξ takes

the value a. The event that ξ takes a is denoted by $\{\xi = a\}$. One can see $\mathbb{P}(\xi = a)$ also as abbreviation for $\mathbb{P}(\{\xi = a\})$. The conditional probability of an event A, given an event B, is denoted by $\mathbb{P}(A|B)$ and defined as usual by

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

where necessarily $\mathbb{P}(B) > 0$.

Throughout the whole section we will consider only rooted phylogenetic trees and C will be a set of character states. The character state $\overline{\chi}(\rho) \in C$ is selected randomly under some probability distribution on C, or it can be considered as initially given. All other character states $\overline{\chi}(v) \in C$ for $v \in V(\mathcal{T}) \setminus \{\rho\}$ are random too, but they depend statistically on the character states of their parent nodes, since the species v evolved from the species represented by the parent node of v. This is modeled by a Markov process as follows (we follow here the definitions and notation of [63, pp. 185–188], the whole section sums up [63, chapt. 8] unless otherwise stated, however this approach is widely accepted and can be found similarly in other resources too).

Definition 2.11. Let $\mathcal{T}$ be a rooted phylogenetic tree with vertex set V and edge set E. Let $<$ be any strict total order on V, such that $v_1 < v_2$ if $(v_1, v_2) \in E$[7]. A set of random variables $\{\xi_v | v \in V\}$ with values in C is said to be a *Markov process* on $\mathcal{T}$ if

$$\mathbb{P}\left(\xi_v = \alpha | \bigcap_{w<v} \{\xi_w = \alpha_w\}\right) = \mathbb{P}(\xi_v = \alpha | \xi_u = \alpha_u), \tag{2.2}$$

for any vertices $v, u \in V$ with $(u, v) \in E$ and any character states $\alpha \in C$ and $\alpha_w \in C$ for all $w \in V$. Equation (2.2) is called the *Markov property*.

This means that—if the character states evolve from the root towards the leaves according to a Markov process—the probability of an event $\{\xi_v = \alpha\}$ for a node $v \in V$ and $\alpha \in C$ typically depends on the character state of its parent node u, but this event is statistically independent from its previous development until u.

For an edge $e = (u, v) \in E$ the probability $\mathbb{P}(\xi_v \neq \xi_u)$ is called *substitution probability* of e and is denoted by $p(e)$. The probability $\mathbb{P}(\xi_v = \xi_u)$ is called the *conservation probability* and equals $1 - p(e)$. As earlier mentioned we speak of a *substitution*, if $\overline{\chi}(u) \neq \overline{\chi}(v)$ for some character extension $\overline{\chi}$. Otherwise, if $\overline{\chi}(u) = \overline{\chi}(v)$ the state of the character was conserved when

[7] Recall that we might regard rooted trees as directed graphs and therefore write $(u, v) \in E$ instead of $\{u, v\} \in E$, if the vertex u is closer to the root than v.

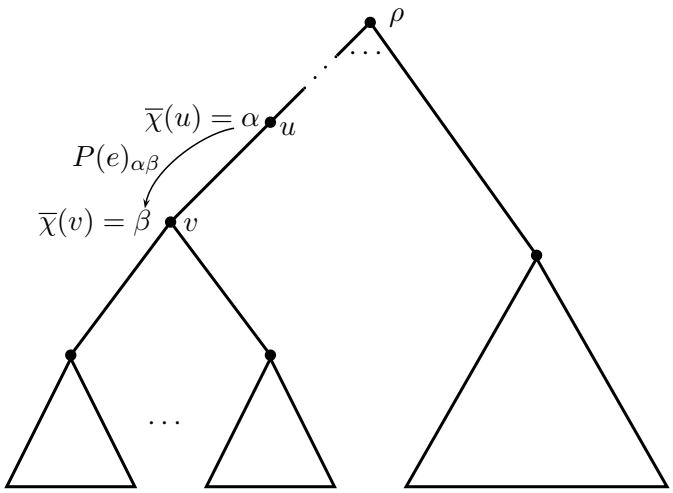


Figure 2.5: The character extension $\overline{\chi}$ evolves from u to v and given that $\overline{\chi}(u) = \alpha$, the probability that $\overline{\chi}(v) = \beta$ is $P(e)_{\alpha\beta}$.

species u evolved to species v. As indicated already in Chapter 1, under certain conditions we can view the substitution probabilities as edge weights (such edge weights are often referred to as branch lengths in biology). But first we are going a to present a more general setting where a *transition matrix* is assigned to each edge.

For every edge $e = (u, v) \in E$ we define the *transition matrix* $P(e)$ as a $|C| \times |C|$ matrix with rows and columns indexed over C and its entry at row α and column β

$$P(e)_{\alpha\beta} := \mathbb{P}(\xi_v = \beta | \xi_u = \alpha).$$

The probability of the evolution of a specific character. It is now possible to describe the probability that a specific character evolves by means of the transition matrices. Let $\overline{\chi}$ be a character extension on $\mathcal{T}$. The probability that a random character extension equals $\overline{\chi}$ is given by

$$p(\overline{\chi}) = \mathbb{P}\left(\bigcap_{v \in V} \{\xi_v = \overline{\chi}(v)\} \right), \tag{2.3}$$

which is the joint probability of the events $\{\xi_v = \overline{\chi}(v)\}$ for each vertex $v \in V$. Now the multiplication theorem $\mathbb{P}\Big(\cap_{j=1}^{k} A_j\Big) = \mathbb{P}(A_1) \cdot \prod_{i=2}^{k} \mathbb{P}\Big(A_i | \cap_{j=1}^{i-1} A_j\Big)$ can be applied. In order to do so, denote the vertices by $\{\rho = v_1, v_2, v_3, \ldots, v_k\} = V$, such that $i < j$ if $(v_i, v_j) \in E$ is an

edge of the tree. This yields

$$p(\overline{\chi}) = \mathbb{P}(\xi_\rho = \overline{\chi}(\rho)) \cdot \prod_{i=2}^{k} \mathbb{P}\left(\xi_{v_i} = \overline{\chi}(v_i) | \bigcap_{j<i} \{\xi_j = \overline{\chi}(v_j)\} \right).$$

By use of the Markov property in (2.2) and by expressing the transition probabilities by means of the entries of the transition matrix, this simplifies to

$$p(\overline{\chi}) = \mathbb{P}(\xi_\rho = \overline{\chi}(\rho)) \cdot \prod_{e=(u,v)\in E} P(e)_{\overline{\chi}(u)\overline{\chi}(v)}. \tag{2.4}$$

Furthermore, it is possible to express the probability $p(\chi)$ of the event, that a specific character χ evolves—regardless of the character states of the internal nodes of the tree. This event is given by $\bigcap_{x\in X} \{\xi_{\phi(x)} = \chi(x)\}$ for an phylogenetic tree $\mathcal{T} = (T, \phi)$ with label set X. Clearly it can be expressed also as the disjoint union

$$\dot{\bigcup_{\overline{\chi}\circ\phi=\chi}} \bigcap_{v\in V} \{\xi_v = \overline{\chi}(v)\}$$

over all character extensions $\overline{\chi}$ of χ on $\mathcal{T}$. Hence, the probability is a sum of (2.3)

$$p(\chi) = \sum_{\overline{\chi}\circ\phi=\chi} p(\overline{\chi}),$$

where the summation is again over all character extensions $\overline{\chi}$ of χ on $\mathcal{T}$. Using the expression for $p(\overline{\chi})$ in (2.4) this finally yields

$$p(\chi) = \sum_{\overline{\chi}\circ\phi=\chi} \mathbb{P}(\xi_\rho = \overline{\chi}(\rho)) \cdot \prod_{e=(u,v)\in E} P(e)_{\overline{\chi}(u)\overline{\chi}(v)}.$$

Note that an exponential number of summands is necessary to compute this sum. To be precise, there are $|C|^{|V|-|X|}$ possible (not necessarily minimum) character extensions $\overline{\chi}$ of χ where $|C|$ is the number of possible character states, $|V|$ the number of vertices in the tree and $|X|$ the number of labels and therefore also the number of leaves in the tree.

The N_r-model. A very common special case of a Markov process on a rooted phylogenetic tree is the *N_r-model*, introduced by Neyman [54] and also known as the *r-state Neyman model*, (see e.g. [24]). In biology the N_4-model often is referred to as the *Jukes-Cantor model* introduced

by Jukes and Cantor [47] in 1969. Especially the cases $r = 4$ and $r = 20$ are of biological interest because there are four nucleobases in the DNA (adenine, cytosine, guanine and thymine) and twenty amino acids occurring in nature (see e.g. [28, p. 43f.] and [76, p. 584]).

Definition 2.12. Let $\mathcal{T}$ be a rooted phylogenetic tree, C a set of character states with $|C| = r$, $\{\xi_v | v \in V\}$ a Markov process on $\mathcal{T}$ and $P(e)$ the transition matrix associated with edge e for every $e \in E(\mathcal{T})$. If ξ_ρ is uniformly distributed on C and if the substitutions on every edge $e = (u, v) \in E(\mathcal{T})$ from any $\alpha \in C$ to any $\beta \in C$ with $\alpha \neq \beta$ are equally likely, we speak of the N_r-model. To be precise, this means that the transition matrices are of the form

$$P(e)_{\alpha\beta} = \begin{cases} 1 - p(e), & \text{if } \alpha = \beta, \\ \frac{1}{r-1} p(e), & \text{otherwise,} \end{cases} \tag{2.5}$$

where $p(e)$ is the substitution probability for the edge $e \in E(\mathcal{T})$. Additionally, the following constraint may be satisfied for all $e \in E(\mathcal{T})$

$$0 < p(e) < \frac{r-1}{r}. \tag{2.6}$$

We refer to the triple $\mathcal{T} = (T, \phi, p)$ as *phylogenetic tree under the N_r-model*, where $p : E(T) \to \mathbb{R}$ is a map as described above.

The constraint in (2.6) may not seem unnatural in any case, but in addition it is important for technical reasons, explained detailed in [63, (8.18), p. 197f.]. With this simplifications it makes sense to view $p(e)$ as edge weight for the edge e. In Chapter 4 we are going to analyze the accuracy of the Fitch-Hartigan algorithm under the N_r-model.

Ultrametric trees. As already mentioned in Chapter 1, it is often assumed that the rate of mutations is constant through time. This means that the evolutionary change between a species and one of its ancestral species is proportional to the time passed. That is why, in this case, biologists speak of a *molecular clock* or a *clock-like* tree (see e.g. [24]). The distance from the root to the species at the leaves is equal for all leaves of the tree regardless if you view the length of the edges as passed time or as measure of evolutionary change between the species (the considered species are present species and then, of course, the time passed during their development from their common ancestor is equal for all of them). We will not go into details of distance based methods of inferring phylogenetic trees from distance data between the species. But we want to mention some interesting aspects at this point (Semple and Steel [63, chapt. 7] give a detailed

overview). Given positive edge weights $w_e \in \mathbb{R}^{>0}$ on a phylogenetic tree, one can define a metric on $V(\mathcal{T})$ by using the unique path between two nodes and define their distance as the sum of the weights along this path. Such a metric is called *tree metric*. A map $\delta : X \times X \to \mathbb{R}$ is called *dissimilarity map* if $\delta(x,x) = 0$ and $\delta(y,x) = \delta(x,y)$ for all $x, y \in X$. A dissimilarity map is the information we can get from the species in X. Therefore a very nice result is, that if δ is a dissimilarity map and if there is a phylogenetic tree $\mathcal{T}$ and a tree metric d on $\mathcal{T}$ such that δ is identical[8] with d on $X \times X$, then $\mathcal{T}$ and d are unique up to isomorphism (see [63, thm. 7.1.8, p. 148]). Furthermore tree metrics can be characterized by the *four-point condition* (see [63, thm. 7.2.6, p. 152]), that is: If δ is a dissimilarity map then δ is extendable to a tree metric d on a suitable phylogenetic tree if and only δ satisfies for $w, x, y, z \in X$

$$\delta(w,x) + \delta(y,z) \leq \max\{\delta(w,y) + \delta(x,z), \delta(w,z) + \delta(x,y)\}.$$

Ultrametrics as defined in [63] are a special case of tree metrics and therefore are also unique up to isomorphism as explained. Ultrametrics are used also in other fields of mathematics, for instance in p-adic analysis (see e.g. [61]).

Anyhow, for our purpose we want to define ultrametricity slightly different. Instead of defining a distance between two species by their dissimilarity, we consider a Markov process on a rooted phylogenetic tree and use the substitution probability as distance.

Definition 2.13. Let $\mathcal{T}$ be a rooted phylogenetic tree with label set X and $\{\xi_v | v \in V(\mathcal{T})\}$ a Markov process on $\mathcal{T}$. $\mathcal{T}$ is said to be *ultrametric* if the substitution probability from the root to all leaves is equal, that is

$$\mathbb{P}(\xi_\rho \neq \xi_u) = \mathbb{P}(\xi_\rho \neq \xi_v) =: p$$

for any two leaves u, v of $\mathcal{T}$. In this case we call p the *height of the tree*.

The following remark illuminates the way we define distances on trees. This is especially interesting for our definition of ultrametric trees, but it is valid for any rooted phylogenetic tree under the N_r-model.

Remark 2.14. Note that in contrast to tree metrics on phylogenetic trees in our setting the distances induced by the substitution probabilities are not additive in the following sense. Consider a rooted phylogenetic tree $\mathcal{T}$ under the N_r-model. If we have two edges $(v_0, v_1) \in E(\mathcal{T})$ and

[8] Note that this is an imprecise simplification and shall only briefly illustrate the results. A tree metric is defined on $V \times V$ and a dissimilarity map on $X \times X$, but in case of phylogenetic trees the set of leaves as subset of V can be identified with X and in this sense one needs to understand "identical" on $X \times X$.

$(v_1, v_2) \in E(\mathcal{T})$ with substitution probabilities $p_{0,1} := \mathbb{P}(\xi_{v_0} \neq \xi_{v_1})$ and $p_{1,2} := \mathbb{P}(\xi_{v_1} \neq \xi_{v_2})$ and $p_{0,2} := \mathbb{P}(\xi_{v_0} \neq \xi_{v_2})$ the substitution probability from v_0 to v_2 (see Figure 2.6), of course, it cannot be $p_{0,2} = p_{0,1} + p_{1,2}$ as one might suggest in analogy to tree metrics. Instead the following lemma holds (for $r = 2$ this fact is also used in [24] but not explicitly mentioned or proved).

Lemma 2.15. *Let $\mathcal{T} = (T, \phi)$ be a rooted phylogenetic tree under the N_r-model with a set of random variables $\{\xi_v | v \in V\}$ and a state set C, where $r = |C|$. Furthermore, let $v_0 = \rho, v_1, \ldots, v_n, v_{n+1}$ be a path in T, i.e. $(v_i, v_{i+1}) \in E(\mathcal{T})$ for $i = 0, 1, \ldots, n$, and denote the substitution probabilities from v_i to v_j by $p_{i,j} := \mathbb{P}\big(\xi_{v_i} \neq \xi_{v_j}\big)$ for any $0 \leq i < j \leq n+1$ (see Figure 2.6). Then the substitution probability from v_0 to v_{n+1} can be calculated from the edge weights by the following recursive formula*

$$p_{0,n+1} = p_{n,n+1} + p_{0,n} - \frac{r}{r-1} \cdot p_{0,n} \cdot p_{n,n+1}.$$

If one uses the transformation $P_{i,j} := 1 - \frac{r}{r-1} p_{i,j}$ this yields the following explicit expression for the substitution probability from v_0 to v_n

$$p_{0,n} = \frac{r-1}{r}(1 - P_{0,1} \cdot P_{1,2} \cdot \ldots \cdot P_{n-1,n}). \tag{2.7}$$

Neither is $\rho = v_0$ a necessary condition nor that v_{n+1} is a leaf, but this is the way we are going to use this lemma later on.

Proof. The probability event $\big\{\xi_{v_0} \neq \xi_{v_{n+1}}\big\}$, that v_0 is in a different state than v_{n+1}, is the disjoint union of the three events $\big\{\xi_{v_n} \neq \xi_{v_{n+1}}\big\} \cap \{\xi_{v_0} = \xi_{v_n}\}$, $\big\{\xi_{v_n} = \xi_{v_{n+1}}\big\} \cap \{\xi_{v_0} \neq \xi_{v_n}\}$ and $\big\{\xi_{v_n} \neq \xi_{v_{n+1}}\big\} \cap \{\xi_{v_0} \neq \xi_{v_n}\} \cap \big\{\xi_{v_0} \neq \xi_{v_{n+1}}\big\}$. These are the events that a change of the state happens between v_n and v_{n+1} but not between v_0 and v_n, and the other way around, and that both between v_n and v_{n+1} and v_0 and v_n the state changes, but it is different also between v_0 and v_{n+1}. The latter may occur only if $r > 2$. In the first two cases the events are intersections of two independent events and therefore their probability is given by the multiplication of the individual events

$$\mathbb{P}\big(\big\{\xi_{v_n} \neq \xi_{v_{n+1}}\big\} \cap \{\xi_{v_0} = \xi_{v_n}\}\big) = p_{n,n+1} \cdot (1 - p_{0,n})$$

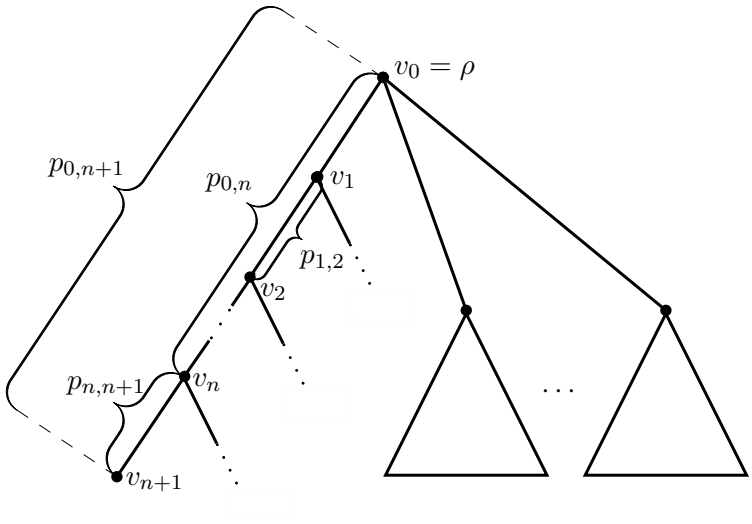


Figure 2.6: Illustration of the substitution probability using the N_r-model.

and

$$\mathbb{P}\big(\{\xi_{v_n} = \xi_{v_{n+1}}\} \cap \{\xi_{v_0} \neq \xi_{v_n}\}\big) = (1 - p_{n,n+1}) \cdot p_{0,n}.$$

Given that $\{\xi_{v_0} \neq \xi_{v_n}\}$ occurs, denote $\alpha := \xi_{v_0}$ and $\beta := \xi_{v_n}$, and $p_{0,n}$ is the probability of this event. Now by (2.5) we have $\mathbb{P}\big(\xi_{v_{n+1}} = \gamma | \xi_{v_n} = \beta\big) = \frac{1}{r-1} \cdot p_{n,n+1}$ and for $\gamma \in C \setminus \{\alpha \cup \beta\}$ these are $r - 2$ disjoint events, being independent from $\{\xi_{v_0} \neq \xi_{v_n}\}$. In total this yields

$$\begin{aligned} p_{0,n+1} &= p_{n,n+1} \cdot (1 - p_{0,n}) + (1 - p_{n,n+1}) \cdot p_{0,n} + \frac{r-2}{r-1} \cdot p_{0,n} \cdot p_{n,n+1} = \\ &= p_{n,n+1} + p_{0,n} - \frac{r}{r-1} \cdot p_{0,n} \cdot p_{n,n+1}. \end{aligned}$$

Now by induction one can prove (2.7). This is straightforward, but quite technical. Therefore details are omitted here.

□

2.4 Combinatorial basics

Chapter 3 is entirely devoted to enumeration problems in connection with phylogenetics. In this section we want to outline some of the history and the basics of combinatorics, especially the tools we are going to use. However, readers less familiar with the basics of combinatorics are

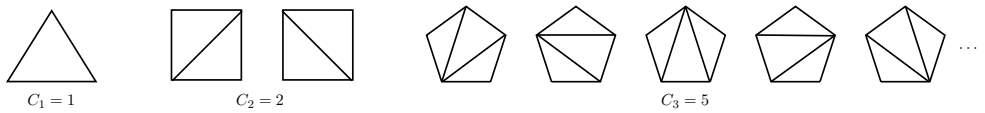


Figure 2.7: The number of polygons with $n + 2$ sides decomposed into triangles.

encouraged to consult the comprehensive introduction to *Analytic Combinatorics* by Flajolet and Sedgewick [27] or other material.

Given a specific class of objects, in combinatorics typically we want to enumerate all objects of size n in this class. This number, might be used, for example to determine the probability for choosing such objects with uniform probability. Of course, a necessary constraint is that the set of objects of size n is finite. If this is the case for all $n \in \mathbb{N}$, we speak of a *combinatorial class* and the related sequence with values in $\mathbb{N}$ is called *counting sequence*. We follow here the definitions of [27, chapt. 1].

Definition 2.16. A *combinatorial class* $\mathcal{A}$ is a finite or countably infinite set of objects, associated with a size function $|.|_{\mathcal{A}} : \mathcal{A} \to \mathbb{N}$ such that all subsets $\mathcal{A}_n := \{a \in \mathcal{A} || a|_{\mathcal{A}} = n\}$ of $\mathcal{A}$ are finite. The *counting sequence* of $\mathcal{A}$ is the sequence $(a_n)_{n\in\mathbb{N}}$ where $a_n \in \mathbb{N}$ is the number of objects in $\mathcal{A}$ of size n. This means, a_n is the size of the preimage of $\{n\}$ with respect to the size function $|.|_{\mathcal{A}}$

$$a_n = |\{a \in \mathcal{A} || a|_{\mathcal{A}} = n\}|.$$

Often considered combinatorial classes include classes of permutations[9], classes of graphs and classes of trees. The latter is the category of problems we will consider in Chapter 3. The origin of many of these problems dates back to the 18th and 19th century. For example, Euler wrote in 1751 to his friend Goldbach,

> “Ich bin neulich auf eine Betrachtung gefallen, welche mir nicht wenig merkwürdig vorkam. Dieselbe betrifft, auf wie vielerley Arten ein gegebenes polygonum durch Diagonallinien in triangula zerschnitten werden könne.[10]”

This combinatorial class consists of polygons decomposed into triangles by diagonal lines (see Figure 2.7). The size function is defined by the number of sides of the polygon minus 2. The according counting sequence $(C_n)_{n\geq 1}$ is therefore given by the number of decomposed polygons

[9] A permutation is a bijective function from $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}^{>0}$.

[10] “I have recently encountered a question, which appears to me rather noteworthy. It concerns the number of ways in which a given [convex] polygon can be decomposed into triangles by diagonal lines.” (The quotation is from [34, pp. 549–552], the translation to English is from [27, Figure I.2, p. 20].)

with $n+2$ sides, which can be shown to equal the famous *Catalan numbers*

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

At the same time C_n is the number of rooted plane binary trees with exactly n internal nodes (see [27, p. 738]). Euler further writes,

> “Ueber die Progression der Zahlen $1, 2, 5, 14, 42, 132$, etc. habe ich auch diese Eigenschaft angemerket, dass $1 + 2a + 5a^2 + 14a^3 + 42a^4 + 132a^5 + \text{etc.} = \frac{1-2a-\sqrt{1-4a}}{2aa}$. Also wenn $a = \frac{1}{4}$, so ist $1 + \frac{2}{4} + \frac{5}{4^2} + \frac{14}{4^3} + \frac{42}{4^4} + \text{etc.} = 4$.[11] ”

Although it is unclear if Euler knew a proof for these statements, he provided an explicit expression for the counting sequence C_n and for the series $\sum_n C_n z^n$ in his letter to Goldbach. It seems as if he had used properties of the sequence C_n to conclude that the series $\sum_n C_n \frac{1}{4^n}$ approaches 4. We will do the opposite—formal power series of the form $\sum_n a_n z^n$ will provide a useful tool to determine recurrence relations, explicit formulas and the asymptotic behavior of the sequence a_n.

Definition 2.17. The formal power series $A(z) = \sum_{n\geq 0} a_n z^n$ is called *ordinary generating function* (OGF) of the sequence $(a_n)_{n\geq 0}$. The formal power series $\overline{A}(z) = \sum_{n\geq 0} a_n \frac{z^n}{n!}$ is called *exponential generating function* (EGF) of the sequence $(a_n)_{n\geq 0}$.

We will stick to the convention that the combinatorial class and its counting sequence is denoted by the same group of letters (e.g. $(a_n)_{n\geq 0}$ denotes the counting sequence, $\mathcal{A}$ the combinatorial class and $A(z)$ the generating function).

On the one hand the formal power series $\sum_n a_n z^n$ can be considered simply as different notation of the sequence $(a_n)_{n\geq 0}$, on the other hand formal power series reveal powerful methods for the manipulation of sequences. The sum, product (sometimes also called *Cauchy product*) and powers of formal power series are defined in analogy to polynomials, e.g.

$$A(z) + B(z) := \sum_{n\geq 0}(a_n + b_n)z^n$$

and

$$A(z) \cdot B(z) := \sum_{n\geq 0}\left(\sum_{k=0}^{n} a_k b_{n-k}\right)z^n,$$

[11] “Regarding the progression of the numbers $1, 2, 5, 14, 42, 132$, and so on, I have also observed the following property: $1 + 2a + 5a^2 + 14a^3 + 42a^4 + 132a^5 + \text{etc.} = \frac{1-2a-\sqrt{1-4a}}{2aa}$.” (See [34, pp. 549–552], the translation is from [27, Figure I.2, p. 20].) Euler then concludes: “So if $a = \frac{1}{4}$, then $1 + \frac{2}{4} + \frac{5}{4^2} + \frac{14}{4^3} + \frac{42}{4^4} + \text{etc.} = 4$.”

where $A(z) = \sum_{n\geq 0} a_n z^n$ and $B(z) = \sum_{n\geq 0} b_n z^n$ (see [27, sect. A.5] and [44, chapt. 1] for details).

The symbolic method. Many combinatorial classes can be constructed by applying certain operations to some elementary combinatorial classes. Such a construction can be translated directly into an equation for the generating function.

Definition 2.18. The combinatorial class $\mathcal{E}$ is called *neutral class* and contains only one object of size 0. The combinatorial class $\mathcal{Z}$ is called *atomic class* and contains only one labeled object of size 1. For disjoint[12] combinatorial classes $\mathcal{A}$ and $\mathcal{B}$ the combinatorial class $\mathcal{A} + \mathcal{B}$ is defined by

$$\mathcal{A} + \mathcal{B} := \mathcal{A} \dot{\cup} \mathcal{B}$$

and its associated size function by

$$|x|_{\mathcal{A}+\mathcal{B}} := \begin{cases} |x|_{\mathcal{A}}, & \text{if } x \in \mathcal{A}, \\ |x|_{\mathcal{B}}, & \text{if } x \in \mathcal{B}, \end{cases}$$

where $x \in \mathcal{A} \dot{\cup} \mathcal{B}$.

Other operations for combinatorial classes can be defined in a similar way. In Chapter 3 the combinatorial classes $\mathcal{A} \times \mathcal{B}$, $\mathcal{A} \star \mathcal{B}$, $\textsc{Set}_2(\mathcal{A})$ (see proof of Theorem 3.2), $\textsc{Set}_{\geq 2}(\mathcal{A})$ (see Section 3.1.2) and $\textsc{MSet}_2(\mathcal{A})$ (see Section 3.3) will be used, formal definitions for these can be found in [27]. $\mathcal{A} \times \mathcal{B}$ and $\mathcal{A} \star \mathcal{B}$ bot consist of pairs (a, b), where $a \in \mathcal{A}$ and $b \in \mathcal{B}$. $\textsc{Set}_2(\mathcal{A})$ consists of sets $\{a_1, a_2\}$, where $a_1 \in \mathcal{A}$ and $a_2 \in \mathcal{A}$. $\textsc{Set}_{\geq 2}(\mathcal{A})$ consists of finite sets of at least 2 elements $a_i \in \mathcal{A}$. $\textsc{MSet}_2(\mathcal{A})$ consists of multisets of 2 elements $a_i \in \mathcal{A}$. Some of these operations ($\star$, $\textsc{Set}_2$ and $\textsc{Set}_{\geq 2}$) are defined only for classes of labeled structures, such as phylogenetic trees, others only for classes of unlabeled structures ($\times$ and $\textsc{MSet}_2$). Combinatorial classes of unlabeled structures correspond to OGFs, while combinatorial classes of labeled structures correspond to EGFs as we will briefly outline in the following.

Lemma 2.19. *Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be combinatorial classes of unlabeled objects and $A(z)$, $B(z)$ and*

[12] If $\mathcal{A}$ and $\mathcal{B}$ are not disjoint, disjoint isomorphic copies of $\mathcal{A}$ and $\mathcal{B}$ can be used.

$C(z)$ their corresponding OGFs. Then the following identities hold

$$\begin{aligned}
&\mathcal{A} = \mathcal{B} + \mathcal{C} \Rightarrow A(z) = B(z) + C(z)\\
&\mathcal{A} = \mathcal{B} \times \mathcal{C} \Rightarrow A(z) = B(z) \cdot C(z)\\
&\mathcal{A} = \textsc{MSet}_2(\mathcal{B}) \Rightarrow A(z) = \frac{1}{2}B(z)^2 + \frac{1}{2}B\big(z^2\big).
\end{aligned}$$

Lemma 2.20. *Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be combinatorial classes of labeled objects and $A(z)$, $B(z)$ and $C(z)$ their corresponding EGFs. Then the following identities hold*

$$\begin{aligned}
\mathcal{A} &= \mathcal{B} + \mathcal{C} \Rightarrow A(z) = B(z) + C(z)\\
\mathcal{A} &= \mathcal{B} \star \mathcal{C} \Rightarrow A(z) = B(z) \cdot C(z)\\
\mathcal{A} &= \textsc{Set}_2(\mathcal{B}) \Rightarrow A(z) = \frac{1}{2}B(z)^2\\
\mathcal{A} &= \textsc{Set}_{\geq 2}(\mathcal{B}) \Rightarrow A(z) = e^{B(z)^2} - 1 - B(z).
\end{aligned}$$

Proofs for the last two lemmata can be found in [27, sect. I.2.2] and [27, sect. II.2.1].

Bivariate generating functions. To examine properties of double sequences, it makes sense to introduce generating functions in two variables.

Definition 2.21. The formal power series $\widehat{A}(x,y) = \sum_{n,m\geq 0} a_{n,m}\frac{x^n}{\omega_n}y^m$ in two variables is called *bivariate generating function* (BGF) of the double sequence $(a_{n,m})_{n,m\geq 0}$, where $\omega_n = 1$ (ordinary BGF) or $\omega_n = n!$ (exponential BGF).

The *formal derivative* $A'(z)$ of a series $\sum_{n\geq 0} a_n z^n$ is defined by

$$A'(z) = \sum_{n\geq 0}(n+1)a_{n+1}z^n.$$

We will also use the notation $\partial_z^n A(z)$ for the n-th derivative. This can be used to determine the mean and variance (and also higher moments) of a parameter of the objects of a specific size in a combinatorial class. For a combinatorial class $\mathcal{A}$, a parameter is a function $\eta : \mathcal{A} \to \mathbb{N}$. Let $a_{n,m} := |\{a \in \mathcal{A} || a|_{\mathcal{A}} = n, \eta(a) = m\}|$ be a double sequence and $A(x,y) = \sum_{n,m\geq 0} a_{n,m}\frac{x^n}{\omega_n}y^m$ its associated BGF (where $\omega_n = 1$ if the objects are unlabeled and $\omega_n = n!$ if n labels are assigned to every object of size n). In Chapter 3 double sequences $a_{n,m}$ will be used to denote the number of certain trees with n (labeled) leaves and m internal nodes, and

the mean refers to the average number of internal nodes in a tree with n leaves. Considering a random object a of size n selected uniformly from $\mathcal{A}_n$, the mean μ_n and the variance σ_n^2 of the parameter η can be determined by means of the derivative of $A(x, y)$ with respect to y and evaluation at $y = 1$. In [27, p. 158f.] the following identities are proven in detail

$$\begin{aligned} \mu_n &= \frac{[x^n]A_y(x,1)}{[x^n]A(x,1)}, \\ \sigma_n^2 &= \frac{[x^n]A_{yy}(x,1) + A_y(x,1)}{[x^n]A(x,1)} - \mu_n^2, \end{aligned}$$

where $A_y(x, y)$ denotes the first and $A_{yy}(x, y)$ the second derivative with respect to y.

Asymptotic analysis. Often we are interested in the growth rate of a sequence a_n and not in its exact values. Sometimes it is not even possible to establish an exact formula. However, if a combinatorial class admits an iterative specification in terms of the operators $+$, $\times$, $\textsc{MSet}_k$ etc. or $+$, $\star$, $\textsc{Set}_k$ etc. for labeled structures, then the *exponential order* K of the counting sequence is a computable real number, i.e. $K := \limsup_{n\to\infty} \sqrt[n]{|a_n|}$ and K can be computed to within any given precision by a computer program terminating in finite time (see [27, sect. IV.4] for a precise formulation of this statement and a proof).

Although we defined generating functions as *formal* power series, they can be considered also as functions $A(z) : \Omega \to \mathbb{C}$ for a proper subset $\Omega \subseteq \mathbb{C}$, such that $A(z)$ converges for all $z \in \Omega$. This approach turns out to be useful, in order to determine the asymptotic behavior of the coefficients of $A(z)$.

Definition 2.22. The *exponential order* K of a sequence $(a_n)_{n\in\mathbb{N}}$ is defined by

$$K := \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

The sequence then—given that K is finite–is of the form $a_n = K^n\theta(n)$ for an appropriate function $\theta(n)$ which satisfies $\limsup_{n\to\infty} \sqrt[n]{|\theta(n)|} = 1$ and is called the *subexponential factor*.

Theorem 2.23. *Let $A(z) = \sum_n a_n z^n$ be a power series and R its radius of convergence. If $R > 0$ the exponential order K of $(a_n)_{n\in\mathbb{N}}$ is then given by*

$$K = \frac{1}{R}.$$

A proof can be found in [27, sect. IV.3.2]. Pringsheim's Theorem (see [27, sect. IV.3.1]) states

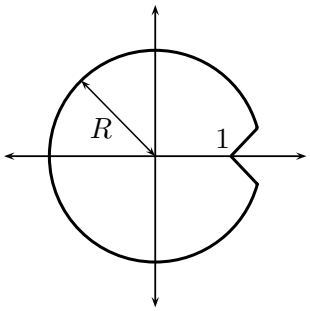


Figure 2.8: Illustration of a Δ-domain.

that $R \in \mathbb{R}^{>0}$ is a singularity of $A(z)$ if all coefficients a_n are nonnegative. Hence, if the smallest singularity of $A(z)$ along the positive real axis is known, one can immediately compute the exponential order of $(a_n)_{n\geq 0}$. Also, if there are negative coefficients a_n, it can be shown that there is a singularity on the boundary of the disc of convergence. Such singularities are called *dominant singularities*:

Definition 2.24. Let $A(z)$ be a function analytic at 0 and $R > 0$ the radius of convergence of its power series expansion at the origin. Singularities of $A(z)$ with modulus R are called *dominant singularities*.

Flajolet and Sedgewick [27, p. 227] therefore declare, "the location of a function's singularities dictates the exponential growth of its coefficients" (*First Principle of Coefficient Asymptotics*), and in addition, "the nature of a function's singularities determines the associate subexponential factor" (*Second Principle of Coefficient Asymptotics*). Singularity analysis as it is summarized in [27, sect. IV.4] corresponds to the Second Principle of Coefficient Asymptotics and reveals information about the subexponential factor of the coefficients of certain functions. The main result is summarized in Theorem 2.26 (see also [27, Corollary VI.1]).

Definition 2.25. Let $\Delta \subseteq \mathbb{C}$ a domain of the form (see Figure 2.8)

$$\Delta := \{z || z| < R, z \neq 1, |arg(z-1)| > \phi\},$$

where $R > 1$ and $0 < \phi < \frac{\pi}{2}$. The domain Δ is called Δ-domain and a function $f : \Delta \to \mathbb{C}$ analytic in Δ is called Δ-analytic.

Theorem 2.26. *If a function $f : \Delta \to \mathbb{C}$ is Δ-analytic and for some $\alpha \in \mathbb{R} \setminus \mathbb{Z}$*

$$f(z) \sim (1-z)^{\alpha},$$

as $z \to 1$, $z \in \Delta$, *then the coefficients of* f *satisfy*

$$[z^n]f(z) \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)},$$

where $\Gamma(\alpha)$ *denotes the Gamma function.*

Chapter 3

Enumeration problems concerning phylogenetic trees

This chapter is a collection of different enumeration problems dealing with phylogenetic trees as it was done before also by Székely et al. [75], Murtagh [52], and Foulds and Robinson [29, 30, 31, 32, 33]. Some of these problems have applications in phylogenetics or elsewhere, some others can be considered rather as applications of combinatorial methods. Such a method is the so-called *symbolic method* (as it is called in [27]) for constructing generating functions or for describing them with equations by use of ready recipes as briefly outlined in Section 2.4. We also will study the asymptotic behavior of sequences by using analytic properties of their generating functions. Although biological motivation is used in this chapter primarily as inspiration, we will mention some direct applications of the results in phylogenetics.

3.1 Tree counting

Having defined phylogenetic trees as graph theoretical objects, a natural question arises from a combinatorial point of view: how many different trees are there? Felsenstein claims in [23, p. 36] that "one use for the numbers [of phylogenetic trees] was 'to frighten taxonomists'." In taxonomy often one is interested in the correct reconstruction of the phylogenetic tree for a group of species or equivalently in selecting the *correct* or *best* tree defined by the species. To solve this problem one might suggest to examine every possible tree of corresponding size and then, for example by pairwise comparison under certain criteria, to find the one best fitting the given data. However, the following results show the limitations for such algorithms. Usually the space of phylogenetic trees is too huge to examine each single tree—only if a very small number of species is studied, this would be possible. Of course, to see these difficulties and to

recognize certain approaches as dead end streets, rough estimates or lower bounds would suffice. Felsenstein [21, p. 7] actually writes about suggested enumerations of certain phylogenetic trees, "There seems to me to be little point in following up these possibilities, as the enumeration of evolutionary trees has somewhat restricted interest." But from a mathematical perspective these results are definitely worth being mentioned. Even Felsenstein admits that "one may have a proposed notation system for a particular category of trees. By considering the ratio between the number of different trees and the number of different configurations of the notation system, one has a measure of the efficiency of the notation system." Likewise the enumeration results can be used to calculate miscellaneous probabilities under the uniform model as, for example, the probability that two random phylogenetic trees are isomorphic (see Section 3.3 and [6]).

But there are also other direct applications of counting trees. Determining the size of particular classes of trees—namely neighborhoods with respect to some appropriate metric in the space of all phylogenetic trees—yields theoretical background information for some greedy algorithms, used to reconstruct phylogenetic trees (see [23, chapt. 4]). As already earlier mentioned, to find the maximum parsimony trees for a given set of characters is a NP-hard problem (see [28]). Therefore several heuristics are applied in practice. One way is to use a hill-climbing method: choose a first tree, examine all trees in a predefined neighborhood of the tree and continue the search at the best tree in the neighborhood. Such algorithms are able to find local extrema, but not global ones. Usually certain tree rearrangement operations are used to define these neighborhoods (see [63, sect. 2.6]). Hill-climbing algorithms are, among other things, the reason why the size of such neighborhoods is of interest. In some cases the neighborhood of a tree does not depend on its shape but only on the number of its nodes (see [2, 66]) and can be determined explicitly.

Also theoretical aspects of tree counting played a big role in the history of combinatorics. The problems considered by Schröder [62] in 1870 and studies by Cayley can be seen as foundation of combinatorics as mathematical discipline. For instance the generating function of unordered rooted trees with respect to the number of their vertices was given by Cayley in 1857 (see [27, sect. I.5.2]). Otter [55] writes "The mathematical theory of trees was first discussed by Cayley". Schröder's results will be used in the following sections.

3.1.1 Rooted binary phylogenetic trees

A problem equivalent to counting rooted binary phylogenetic trees was first studied by Schröder in 1870 (for the original article refer to [62], other references concerning this problem include [69, p. 15], [11, p. 223f.], [27, p. 129], [30], [63, p. 17], [6] and [23, chapt. 3]). He considered the question

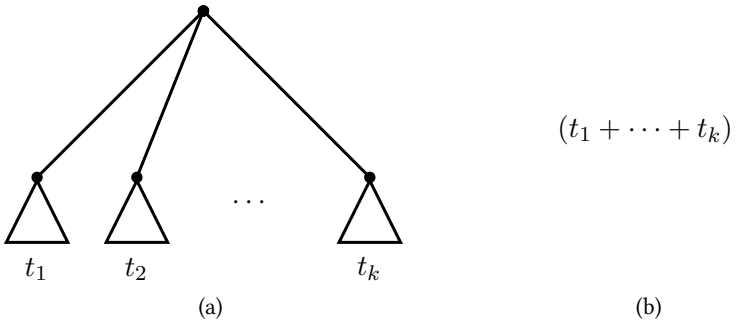


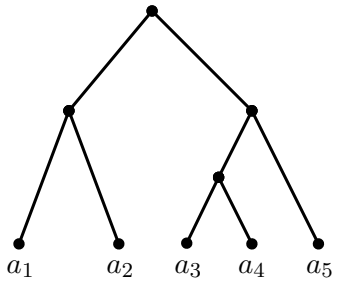
Figure 3.1: Correspondence between rooted plane trees and sums with brackets.

in how many ways a sum of n distinct summands could be denoted. Due to the commutative law we are allowed to rearrange the order of the objects and it is an immediate result that there are $1 \cdot 2 \cdot \ldots \cdot n = n!$ such arrangements. Hence, if the sum is considered as n-ary operator, for summands $a_1, a_2, \ldots, a_n$ there are $n!$ ways to denote the sum $a_{i_1} + a_{i_2} + \cdots + a_{i_n}$ where each corresponds to an ordered n-tuple $(a_{i_1}, \ldots, a_{i_n})$. But if the sum is considered as binary operator, the summands are grouped in brackets. Due to the associative law $x + (y + z) = (x + y) + z$, all different ways to parenthesize the summands result in the same sum. Since the number of arrangements is well-known, Schröder considered the order of the summands to be fixed and studied the number of ways to parenthesize a sum of n summands. In his first problem he studied fully parenthesized sums, i.e. each pair of parentheses contains exactly two summands, and in his second problem he studied arbitrarily parenthesized sums, i.e. each pair of parentheses contains $k \geq 2$ summands.

Such sums correspond bijectively to *rooted plane trees*[1]. Every summand can be perceived as leaf and a pair of brackets corresponds to an internal node. More precisely, the sum a_1 with only one summand a_1 corresponds to the tree consisting of only one leaf. Recursively for terms $t_1, \ldots, t_k$ and any $k \geq 2$ the term $(t_1 + \ldots + t_k)$ corresponds to the tree with a new internal node v as root and the k trees corresponding to the terms $t_1, \ldots, t_k$ as subtrees of v (see Figure 3.1 and the example in Figure 3.2). If the sum is considered as binary operator and fully parenthesized, the related trees are binary too. Note that we do not allow double brackets, implying that the corresponding trees have no vertices of degree 2. Furthermore, also on the outermost level the sums are always parenthesized and any pair of parentheses contains at least two summands.

Based on these initial questions (Schröder's first and second problem), he also studied a similar problem where the considered objects are not ordered. In his third and fourth problem, he imag-

[1]Recall that phylogenetic trees are by definition unordered trees.



(a)

$$((a_1 + a_2) + ((a_3 + a_4) + a_5))$$

(b)

Figure 3.2: Example of a binary rooted tree and its related sum.

ines the objects to be grouped together in nested cells—again each cell containing either exactly two or any number of objects (or other cells). These nested cells are related to phylogenetic trees in the same way as sums to rooted plane leaf-labeled trees as described above. Interestingly, also Hartigan [43] introduces phylogenetic trees by use of nested sets of the species in X instead of using purely graph theoretic terminology.

Therefore the trees considered in Theorem 3.2 relate to Schröder's third problem, but we want to prove it with more modern terminology following the notation of [27]. Surprisingly, there is a very nice formula for the sequence, namely the odd factorials also called *double factorial numbers* and denoted by $n!!$ (this is sequence A001147 in [64]).

Definition 3.1. Let $\mathcal{B}$ the combinatorial class of all rooted binary phylogenetic trees and $\mathcal{B}_n$ the subset of rooted binary phylogenetic trees with label set $|X| = n$. $b_n := |\mathcal{B}_n|$ denotes the number of rooted binary phylogenetic trees.

Theorem 3.2. *The number b_n of rooted binary phylogenetic trees with n leaves satisfies*

$$b_n = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3) = \frac{(2n-3)!}{2^{n-2} \cdot (n-2)!} = (2n-3)!!.$$

Proof. Each tree in $\mathcal{B}$ is either of size 1 (if the tree has only one node being a leaf), or it can be constructed by attaching two trees to a root. Because phylogenetic trees are unordered trees, there is no order between these two trees. Therefore a specification of $\mathcal{B}$ is given by

$$\mathcal{B} = \mathcal{Z} + \textsc{Set}_2(\mathcal{B}), \tag{3.1}$$

where $\mathcal{Z}$ is the labeled atomic class which contains only one labeled object of size 1 and $\textsc{Set}_2(\mathcal{B})$ denotes the class of sets of size 2 with labeled elements of $\mathcal{B}$ (see Section 2.4).

According to Lemma 2.20 (see also [27, fig. II.18, p. 148]) the specification in (3.1) translates to $B(z) = z + \frac{1}{2} \cdot B(z)^2$ where $B(z) = \sum_{n \geq 0} b_n \frac{z^n}{n!}$ denotes the EGF for the series b_n. Hence, the EGF is given by

$$B(z) = 1 \pm \sqrt{1 - 2z}, \tag{3.2}$$

where we cannot decide yet for one of the two roots. We could now derive the coefficients by means of the general binomial theorem, but we will choose a more direct approach and use Taylor's formula:

$$\begin{aligned}
\partial_z^n B(z) &= \partial_z^n 1 \pm \sqrt{1 - 2z} = \\
&= \partial_z^{n-1}(\pm 1) \cdot (-1) \cdot (1 - 2z)^{-\frac{1}{2}} = \\
&= \partial_z^{n-2}(\pm 1) \cdot (-1) \cdot (1 - 2z)^{-\frac{3}{2}} = \\
&= \partial_z^{n-3}(\pm 1) \cdot (-1) \cdot 3 \cdot (1 - 2z)^{-\frac{5}{2}} = \\
&\vdots \\
&= (\pm 1) \cdot (-1) \cdot 3 \cdot 5 \cdot \ldots \cdot (2n - 3) \cdot (1 - 2z)^{-\frac{2n-1}{2}},
\end{aligned} \tag{3.3}$$

where (3.3) can be proved by induction. This yields $b_n = n! \cdot [z^n]B(z) = n! \cdot \frac{\partial_z^n B(0)}{n!} = (\pm 1) \cdot (-1) \cdot 3 \cdot 5 \cdot \ldots \cdot (2n - 3)$. b_n has to be positive, so we can decide on one of the roots in (3.2) and we have

$$B(z) = 1 - \sqrt{1 - 2z}. \tag{3.4}$$

□

Remark 3.3. Theorem 3.2 can be proved also directly by induction (see [63, prop. 2.1.4, p. 17]). But to do so we need to know the result in advance, and we do not obtain the EGF in the form (3.4).

Asymptotic results. Applying Stirling's formula and using $\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = e^{-1}$ yields the following asymptotic results (see [63, p. 17f.])

$$b_n = \frac{(2n-2)!}{2^{n-1}(n-1)!} \sim 2^{1-n} \cdot \sqrt{2} \cdot 2^{2n-2} \cdot (n-1)^{n-1} \cdot e^{1-n} = \sqrt{2} \cdot \left(\frac{2n-2}{e}\right)^{n-1}.$$

This further simplifies to

$$\sqrt{2} \cdot \left(\frac{2n-2}{e}\right)^{n-1} \sim \sqrt{2} \cdot 2^{n-1} n^{n-1} e^{-n}.$$

3.1.2 Rooted phylogenetic trees (multifurcating)

In this section we will analyze the number r_n of multifurcating rooted phylogenetic trees with n leaves (such trees are also called *labeled hierarchies* [27, p. 128]). This sequence was first studied by Schröder [62] in his fourth problem. As described in the previous section, he imagined objects grouped together in nested cells which correspond to rooted phylogenetic trees. In this section also more than two child nodes for each internal node are allowed, so every rooted binary phylogenetic tree is also a rooted phylogenetic tree and we have $r_n > b_n$ for $n \geq 3$, where b_n denotes the number of rooted binary phylogenetic trees as discussed in the previous section. In [69, p. 13f.] this sequence is introduced by partitioning a set of size n into at least two blocks and then doing the same recursively with these blocks until only singletons are left. The result is then called a *total partition of the set* and the number of different total partitions is counted. There are also several connections to other enumeration problems and applications of the following results. In the following we will provide two different ways to calculate r_n recursively—one due to Schröder [62] and one due to Felsenstein [21]. Furthermore we are going to derive the generating function and asymptotic results by Flajolet and Sedgewick [27].

Throughout the whole section, we will use the following notation.

Definition 3.4. The number of multifurcating rooted phylogenetic trees with n leaves is denoted by r_n. The number of multifurcating rooted phylogenetic trees with n leaves and m internal nodes is denoted by $r_{m,n}$.

An implicit formula for the generating function. Recall that phylogenetic trees are unordered trees, labeled at their leaves (with exactly one label per leaf), and every internal node has outdegree at least 2 if we view the edges directed away from the root. Each rooted phylogenetic tree can be identified either with one labeled leaf or with a set containing at least 2 rooted phylogenetic trees. Thus, one can construct the combinatorial class corresponding to the sequence r_n by use of the operator $\textsc{Set}_{\geq 2}$ introduced in Section 2.4:

$$\mathcal{R} = \mathcal{Z} + \textsc{Set}_{\geq 2}(\mathcal{R})$$

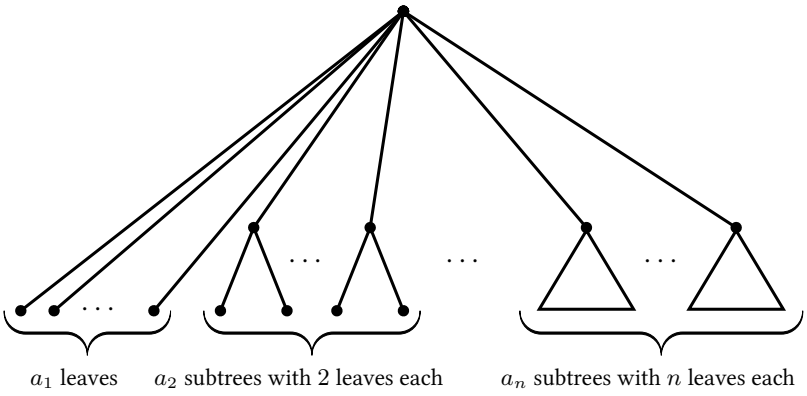


Figure 3.3: Illustration of Schröder's recursion for the number of rooted phylogenetic trees.

According to Lemma 2.20 the EGF $R(z) = \sum_{n\geq 0} \frac{1}{n!} r_n z^n$ satisfies the implicit equation

$$R(z) = z + e^{R(z)} - 1 - R(z) \tag{3.5}$$

(see also Flajolet and Sedgewick [27, fig. II.18, p. 148], [27, p. 472] and [29, p. 171]).

In [69, p. 13f.] $R(z)$ is expressed explicitly by means of the compositional inverse, denoted by $F^{\langle -1\rangle}(z)$, that is $F^{\langle -1\rangle}(F(z)) = F\big(F^{\langle -1\rangle}(z)\big) = z$ for $F(z) = \sum_{n\geq 0} f_n z^n$ with $f_0 = 0$ and $f_1 \neq 0$. From (3.5) we have $z = 1 + 2R(z) - e^{R(z)}$, which leads to

$$R(z) = (1 + 2z - e^z)^{\langle -1\rangle}.$$

Stanley [69, p. 14] states that "it does not seem possible to obtain a simpler result".

Recursions. The recurrence relation, established by Schröder [62] in 1870, is probably more of historical interest. He uses it to derive the equation for the generating function already stated in (3.5).

Theorem 3.5 (Schröder's recurrence relation). *The number r_n of rooted phylogenetic trees satisfies the following recurrence relation*

$$r_{n+1} = \sum_{\substack{a_1,a_2,\ldots,a_n\in\mathbb{N} \\ a_1+2\cdot a_2+\cdots+n\cdot a_n=n+1}} \frac{(n+1)!\cdot r_1^{a_1}\cdot r_2^{a_2}\cdot\ldots\cdot r_n^{a_n}}{a_1!\cdot\ldots\cdot a_n!\cdot(1!)^{a_1}\cdot\ldots\cdot(n!)^{a_n}},$$

where the summation is over all solutions $a_1, a_2, \ldots, a_n \in \mathbb{N}$ of the equation $a_1 + 2\cdot a_2 + \cdots + n\cdot a_n = n+1$.

$m\backslash n$	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
2		3	10	25	56	119	246
3			15	105	490	1 918	6 825
4				105	1 260	9 450	56 980
5					945	17 325	190 575
6						10 395	270 270
7							135 135
$\sum_m r_{n,m} = r_n$	1	4	26	236	2 752	39 208	660 032

Table 3.1: Illustration of Felsenstein's recurrence relation (the table is inspired by the table in [23, p. 27]). The arrows indicate which values $r_{n,7}$ are needed to compute the values $r_{n,8}$.

Proof. Given a rooted phylogenetic tree with $n+1$ leaves, by a_i for $i = 1, \dots, n$ we denote the number of subtrees directly attached to ρ which contain exactly i leaves (see Figure 3.3). Thus the total number of leaves is given by $1 \cdot a_1 + 2 \cdot a_2 + \dots + n \cdot a_n = n+1$. There are $a_1 + a_2 + \dots + a_n$ subtrees directly attached to ρ and each of them is a leaf or contains a set of leaves. If we consider a fixed set of labels for the leaves in each of the subtrees, then there are r_i possibilities for the a_i subtrees with i leaves for every $i = 1, \dots, n$. Hence, there are $r_1^{a_1} \cdot r_2^{a_2} \cdot \ldots \cdot r_n^{a_n} \cdot N$ rooted phylogenetic trees of size $n+1$ for the given $a_1, a_2, \dots, a_n$ and for some factor N which indicates the possible permutations of the labels. There are $(n+1)!$ possibilities to arrange the $n+1$ labels. But for each subtree with i leaves there are $i!$ permutations of the labels within the subtree, so for every subtree of ρ we counted every arrangement of the other labels $i!$ times instead of only once—in total for all subtrees this is $(1!)^{a_1} \cdot \ldots \cdot (n!)^{a_n}$. We also should not distinguish between permutations where only whole subtrees with equal number of leaves are permuted but no label goes to another subtree. There are $a_1! \cdot \ldots \cdot a_n!$ ways to do so. Hence, all together we have $N = \frac{(n+1)!}{a_1! \cdot \ldots \cdot a_n! \cdot (1!)^{a_1} \cdot \ldots \cdot (n!)^{a_n}}$. This leads to the claimed recursive formula. □

Felsenstein [21, p. 29] describes Schröder's methods as "somewhat complex" and therefore states his own recursive formula (see also [23, p. 25ff.]). And indeed, in order to compute the first n_0 values of the sequence r_n for a sufficiently small n_0, this might be the most promising approach (see also Remark 3.14). The following lemma was not explicitly stated by Felsenstein in this way, but it will prove useful also later.

Lemma 3.6. *Let $\mathcal{T} = (T, \phi)$ be a rooted phylogenetic tree with m internal vertices and n leaves.*

Then the following inequality holds

$$n - 1 \geq m,$$

where $n - 1 = m$ if and only if $\mathcal{T}$ is a binary phylogenetic tree.

Proof. We denote the vertex set by V and the set of edges by E, i.e. $T = (V, E)$. Because T is a tree, the number of edges are given by the number of vertices minus 1, i.e.

$$|E| = |V| - 1 = m + n - 1.$$

For all $v \in V$ let denote d_v the outdegree of vertex v. If v is a leaf, we have $d_v = 0$. For all internal nodes v is $d_v \geq 2$ and therefore $\sum_{v \in V} d_v \geq 2 \cdot m$. In addition equality holds, i.e. $d_v = 2$ and $\sum_{v \in V} d_v = 2 \cdot m$, if and only if T is a binary tree. The number of edges in T can be counted by summation of the outdegree d_v for all vertices v, hence $|E| = \sum_{v \in V} d_v$. In total this yields

$$m + n - 1 = |E| = \sum_{v \in V} d_v \geq = 2m,$$

which completes the proof. □

Theorem 3.7 (Felsenstein's recurrence relation). *The number $r_{m,n}$ of rooted phylogenetic trees with m (unlabeled) internal vertices and $n = |X|$ (labeled) leaves satisfies $r_{m,n} = 0$ for $m < 1$ or $m > n - 1$, and for $n \geq 2$* and *$1 \leq m \leq n - 1$ the number $r_{m,n}$ satisfies the following recurrence relation*

$$r_{m,n} = m \cdot r_{m,n-1} + (n + m - 2) \cdot r_{m-1,n-1}. \tag{3.6}$$

Proof. For $m < 1$ or $m > n - 1$, we have $r_{m,n} = 0$ since there is no such tree (see Lemma 3.6). Any rooted phylogenetic tree with $n \geq 2$ leaves has at least one internal node (namely its root) and at most $n - 1$ internal nodes (see Lemma 3.6). There is obviously only one rooted tree with two leaves and one internal node, so we have $r_{1,2} = 1$. Now, for any rooted phylogenetic tree with $n - 1$ leaves there are two different ways to add a new n-th leaf. Either as a new child of an existing internal node or as a child of a new internal node (the new internal node can be added separating an existing edge or as parent of the former root node). Note that in this way each such tree with n leaves can be constructed by exactly one such tree with $n - 1$ leaves—if this is not clear one can imagine the reverse operation of removing the n-th leaf and the parent internal node if there is only one other child node. In other words, we have described a bijection

between two sets of trees. The first set contains the rooted phylogenetic trees with $n-1$ leaves and m internal nodes, where every tree is counted m times because an internal node has to be chosen, and it contains the trees with $n-1$ leaves and $m-1$ internal nodes, where every tree is counted $(n-1)+(m-1)$ times, because an edge or the root node has to be selected. Thus the claimed recursive formula for $r_{m,n}$ is established. □

Table 3.1 shows the values r_n for $n = 1, \ldots, 8$ and the values for $r_{m,n}$ needed for the computation.

Explicit formula. In the following we will express the sequence r_n explicitly by combining two results for the *associated Stirling numbers of the second kind*, listed as sequence A008299 in [64] and usually denoted by $S_2(n,k)$. On the one hand there is an explicit formula due to D. Wasserman (see A059022 in [64]) for $S_2(n,k)$ as we will show in Lemma 3.11. On the other hand there is an identity (see [11, p. 224]) between the numbers $r_{m,n}$ and $S_2(n,k)$, see (3.9) in Theorem 3.12. However, this explicit formula is not helpful to compute values for a large n as we will explain in Remark 3.14.

The notation $S_2(n,k)$ refers to its generalization, the *r-associated Stirling numbers of the second kind* (see A059022 in [64]), denoted by $S_r(n,k)$. For $r=1$ one gets the (ordinary) *Stirling numbers of the second kind* (for details see [11, p. 221f.] and [10]). Furthermore, for each of these sequences we have a corresponding sequence *of the first kind*, and for all of them there are several applications and combinatorial interpretations. None of these are important for our purpose—the generating function for $S_2(n,k)$ and two of its properties, presented in the following lemmata, will suffice.

Definition 3.8 (Associated Stirling numbers of the second kind). The sequence $S_2(.,k)$ is defined by means of its EGF $\sum_{n\geq 0} \frac{1}{n!} S_2(n,k) z^n = \frac{1}{k!}(e^z - z - 1)^k$ for all $k \geq 0$ (see [45, chapt. 3])

Remark 3.9. A BGF for the associated Stirling numbers of the second kind is given by (see [11, p. 221])

$$\sum_{n,k\geq 0} S_2(n,k)\frac{z^n}{n!}u^k = e^{u(e^z-z-1)}.$$

Lemma 3.10. *The associated Stirling numbers $S_2(n,k)$ satisfy $S_2(0,0)=1$, $S_2(n,k)=0$ for $n<2k$ and $S_2(n,0)=0$ for $n>0$ and the following recurrence relation for $n\geq 1, k\geq 1$*

$$S_2(n+1,k) = kS_2(n,k) + nS_2(n-1,k-1). \tag{3.7}$$

k \ n	0	1	2	3	4	5	6	7	8	9	10	11	12
0	1	0	0	0	0	0	0	0	0	0	0	0	0
1			1	1	1	1	1	1	1	1	1	1	1
2					3	10	25	56	119	246	501	1012	2035
3							15	105	490	1918	6825	22935	74316
4									105	1260	9450	56980	302995
5											945	17325	190575
6													10395

Table 3.2: Values for the associated Stirling numbers $S_2(n,k)$ for $0 \leq n \leq 12$

Proof. The boundary values follow directly from the EGF for the given k. To prove (3.7) we make use of the formal integral operator $\int$ (see e.g. [48, p. 12]). For a generating function $A(z) = \sum_{n\geq 0} a_n z^n$ formal integration and formal derivation is defined by $A'(z) = \sum_{n\geq 0}(n+1)a_{n+1}z^n$ and $\int A(z) = \sum_{n\geq 0} \frac{1}{(n+1)} a_n z^{n+1}$ (that implies $\left[z^0\right] \int A(z) = 0$). Note that $\left(\int A(z)\right)' = A(z)$, but the converse does not hold—in general $\int (A'(z)) \neq A(z)$. This is caused by the derivative's property that the constant term a_0 vanishes. But the other coefficients are not affected, so $[z^n] \int (A'(z)) = [z^n]A(z)$ holds for $n \geq 1$.

Now consider for a fixed $k \geq 1$

$$\begin{aligned}
\left(\sum_{n\geq 0} S_2(n,k)\frac{z^n}{n!}\right)' &= \left(\frac{1}{k!}(e^z - z - 1)^k\right)' \\
&= \frac{1}{(k-1)!}(e^z - z - 1)^{k-1}(e^z - 1) \\
&= k \cdot \frac{1}{k!}(e^z - z - 1)^k + z\frac{1}{(k-1)!}(e^z - z - 1)^{k-1} \\
&= k\sum_{n\geq 0}\frac{1}{n!}S_2(n,k)z^n + z \cdot \sum_{n\geq 0}\frac{1}{n!}S_2(n,k-1)z^n \\
&= \sum_{n\geq 0}\frac{1}{n!}k \cdot S_2(n,k)\left(\frac{z^{n+1}}{n+1}\right)' + \sum_{n\geq 0}\frac{1}{n!}S_2(n,k-1)\left(\frac{z^{n+2}}{n+2}\right)' \\
&= \left(\sum_{n\geq 1} kS_2(n-1,k)\frac{z^n}{n!} + \sum_{n\geq 2}(n-1)\cdot S_2(n-2,k-1)\frac{z^n}{n!}\right)'
\end{aligned}$$

Taking the formal integral at both sides of the equation and comparing the coefficients of z^n

for $n \geq 2$ yields the stated recurrence relation. □

Formula (3.8) from the following lemma can be found without proof in the entry for sequence A008299 in [64] and is due to D. Wasserman.

Lemma 3.11. *The associated Stirling numbers can be expressed explicitly by*

$$S_2(n,k) = \sum_{i=0}^{k} (-1)^i \cdot \binom{n}{i} \sum_{j=0}^{k-i} (-1)^j \cdot \frac{(k-i-j)^{n-i}}{j! \cdot (k-i-j)!} \tag{3.8}$$

Proof. By applying the binomial theorem twice to the EGF of $S_2(.,k)$ and by use of the exponentials series expansion one gets

$$\begin{aligned}
S_2(n,k) &= [z^n]\frac{n!}{k!}(e^z - z - 1)^k \\
&= \frac{n!}{k!}[z^n] \sum_{j=0}^{k} \binom{k}{j} (-1)^j (e^z - z)^{k-j} \\
&= \frac{n!}{k!}[z^n] \sum_{j=0}^{k} \binom{k}{j} (-1)^j \sum_{i=0}^{k-j} \binom{k-j}{i} (-1)^i z^i e^{z\cdot(k-i-j)} \\
&= \sum_{j=0}^{k} (-1)^j \sum_{i=0}^{k-j} (-1)^i \binom{k-j}{i} \cdot \binom{k}{j} \cdot \frac{n! \cdot (k-i-j)^{n-i}}{k! \cdot (n-i)!} \\
&= \sum_{j=0}^{k} (-1)^j \sum_{i=0}^{k-j} (-1)^i \binom{n}{i} \cdot \frac{(k-i-j)^{n-i}}{j! \cdot (k-i-j)!}
\end{aligned}$$

Exchanging the order of summation leads to the desired result. □

The identity $r_n = \sum_{m=1}^{n-1} S_2(n+m-1, m)$ can be found (without proof) in [11, p. 224], but no statement about an explicit expression is made there. In the following theorem this identity is proved using Felsenstein's recurrence relation from Theorem 3.7.

Theorem 3.12. *The following identity holds between the associated Stirling numbers $S_2(n,k)$ and the number $r_{m,n}$ of rooted phylogenetic trees with n leaves and m internal nodes*

$$r_{m,n} = S_2(n+m-1, m) \quad \textit{for all } n \geq 2, m \geq 1, \tag{3.9}$$

and therefore the number r_n of rooted phylogenetic trees can be expressed explicitly in the follow-

ing way:

$$r_n = \sum_{m=1}^{n-1} \sum_{i=0}^{m} (-1)^i \cdot \binom{n+m-1}{i} \sum_{j=0}^{m-i} (-1)^j \cdot \frac{(m-i-j)^{n+m-1-i}}{j! \cdot (m-i-j)!}. \tag{3.10}$$

Proof. To prove (3.9) first verify $r_{1,2} = S_2(3-1,1)$ (see Table 3.2) and $S_2(n+m-1,m) = 0$ for $m > n-1$ and $n \geq 2$ (because of $2m > 2n-2 \geq n$ and Lemma 3.10). Since $r_{m,n} = 0$ for $m > n-1$ (see Theorem 3.7) we are done in this case. Then again, by use of Lemma 3.10 one sees that

$$S_2(n+m-1,m) = mS_2(n+m-2,m) + (n+m-2)S_2(n+m-3,m-1),$$

which equals the recurrence relation (3.6) for $r_{n,m}$, and therefore (3.9) is established.

The second statement of the theorem then follows immediately from (3.9) and Lemma 3.11 by summing $r_{m,n}$ over all possible values for m. □

Remark 3.13. Note that the combinatorial class of rooted binary phylogenetic trees, covered in Section 3.1.1, is a subclass of the multifurcating rooted phylogenetic trees discussed here. A rooted phylogenetic tree with n leaves is binary if and only it has $n-1$ internal nodes (see Lemma 3.6). Hence, we have $b_n = r_{n-1,n} = S_2(2n-2,n-1)$ (see also [45, sect. 3]), and we can immediately establish the recurrence relation $b_n = (2n-3)b_{n-1}$ by use of Lemma 3.10 and therefore $b_n = (2n-3)!!$ as already stated in Theorem 3.2.

Remark 3.14. Felsenstein [23, p. 27] states that there is no closed-form formula for r_n. Nevertheless, his recurrence relation can be used to derive an explicit expression, which does not depend on the r_i for $i < n$ (see Theorem 3.12). Strictly speaking, this expression is not closed-form since the number of operations depends on n. This is an important detail—one might think that an explicit formula makes it easier to calculate specific values of the sequence because it is not necessary to know also all the previous values of the sequence. This might be true sometimes, but in this case it is still more efficient to use Felsenstein's recurrence given in (3.6) and to calculate also all r_i for $i = 2, \ldots, n-1$ in order to get r_n than to use the explicit formula in (3.10). The triple sum in (3.10) leads to a total number of $\frac{n^3}{6} + \frac{n^2}{2} + \frac{n}{3} - 1 = O(n^3)$ summands, and for each of them a calculation involving many operations is necessary. On the other hand, in order to compute r_n with (3.6), only $\frac{n^2-n}{2}$ values of the double sequence $r_{m,n}$ are necessary and each of them is computed by at most two multiplications and two additions. Finally, r_n results from summing $n-1$ values, so in total only $O(n^2)$ summations are necessary.

These considerations can be confirmed by measuring the run times. We compared both methods implemented in Mathematica (see Section A.1). Note that both implementations do not use any parallelization—we want to compare the algorithms and not the most efficient code. The computation of r_{200} takes 36 seconds by use of the code in Section A.1.1 and only 0.04 seconds by use of the algorithm in Section A.1.2. On the same machine[2] the computation takes 0.01716 seconds and 0.00066 seconds respectively for r_{22}. This was the highest value calculated by Felsenstein in 1978 using a Fortran program.

Asymptotic results. We can derive asymptotic results for r_n directly by use of the implicit equation in (3.5) even without having an explicit formula for the generating function. In [27, p. 472f.] this is done by use of a theorem, which the authors call *smooth implicit-function schema.*

Theorem 3.15 (Smooth implicit-function schema). *Let $y(z) = \sum_{n\geq 0} y_n z^n$ be a function analytic at 0 with $y_0 = 0$ and $y_n \geq 0$. Furthermore let y belong to the smooth implicit-function schema, meaning that there exists a bivariate function $G(z, w)$ such that*

$$y(z) = G(z, y(z)),$$

where $G(z, w)$ satisfies:

(i) $G(z, w) = \sum_{m,n\geq 0} g_{m,n} z^m w^n$ is analytic in a domain $|z| < S_1$ and $|w| < S_2$, for some $S_i > 0, i = 1, 2$.

(ii) The coefficients of G satisfy

$$\begin{aligned} g_{0,0} &= 0 \\ g_{0,1} &\neq 1 \\ g_{m,n} &\geq 0 \\ g_{m,n} &> 0 \qquad \text{for some } m \text{ and for some } n \geq 2 \end{aligned}$$

(iii) There exist two numbers s_1, s_2 such that $0 < s_i < S_i$ for $i = 1, 2$, satisfying the system of

[2] A (currently) average desktop computer with an Intel® Core™2 Duo E8400 processor (6M Cache, 3.00 GHz, 1333 MHz FSB) and Mathematica 8 (for GNU/Linux 64bit) were used.

equations,

$$\begin{aligned} G(s_1, s_2) &= s_2 \\ G_w(s_1, s_2) &= 1, \end{aligned}$$

which is called the characteristic system.

Then, $y(z)$ converges at $z = s_1$ where it has a square-root singularity

$$y(z) \underset{z \to s_1}{=} s_2 - \gamma\sqrt{1 - \frac{z}{s_1}} + O\left(1 - \frac{z}{s_1}\right)$$

with

$$\gamma := \sqrt{\frac{2s_1 G_z(s_1, s_2)}{G_{ww}(s_1, s_2)}}$$

and the expansion being valid in a Δ-domain. If, in addition, $y(z)$ is aperiodic, then s_1 is the unique dominant singularity of y and the coefficients satisfy

$$[z^n]y(z) \underset{n \to \infty}{=} \frac{\gamma}{2\sqrt{\pi n^3}} s_1^{-n}\big(1 + O\big(n^{-1}\big)\big).$$

A proof can be found in [27, p. 472f.] using an analytic version of the *Implicit Function Theorem* and singularity analysis (see also Section 2.4).

This theorem is directly applicable to our situation. From (3.5) we get $G(z, w) = z + e^w - 1 - w = z^1 \cdot w^0 + \sum_{n \geq 2} \frac{z^0 w^n}{n!}$ and therefore $g_{0,0} = 0$, $g_{0,1} = 0 \neq 1$, $g_{m,n} \geq 0$ and $g_{0,2} = \frac{1}{2} > 0$. Furthermore is $s_2 = \ln 2$ because of the equation $G_w(s_1, s_2) = e^{s_2} - 1 = 1$ in the characteristic system and $s_1 = 2\ln 2 - 1$ because of the equation $G(s_1, s_2) = s_1 + 1 - \ln 2 = \ln 2$. Hence, the conditions for Theorem 3.15 are fulfilled and we have

$$\frac{1}{n!} \cdot r_n \sim \frac{1}{2\sqrt{\pi n^3}} (2\ln 2 - 1)^{-n + \frac{1}{2}}.$$

3.1.3 Unrooted phylogenetic trees (binary and multifurcating)

In the following, results for the number of unrooted phylogenetic trees will be obtained, both for all phylogenetic trees and for the subclass of binary phylogenetic trees (this was done also in [63, prop. 2.2.3, p. 20] and [23, p. 24]). This can be accomplished by using the results for the number of rooted phylogenetic trees in the previous two sections and by providing a appropriate bijective map between rooted and unrooted trees.

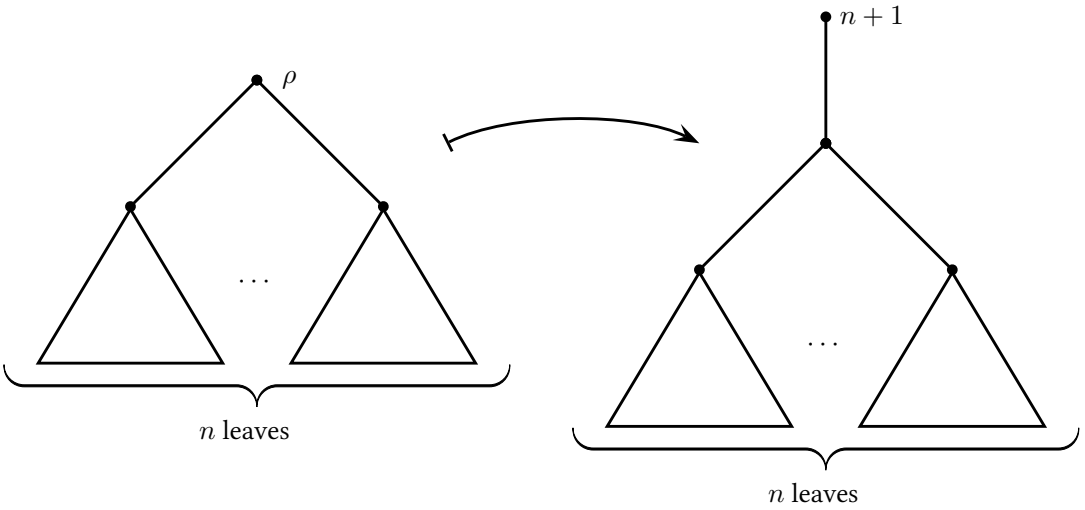


Figure 3.4: Mapping a rooted phylogenetic tree with n leaves to an unrooted phylogenetic tree with $n+1$ leaves by attaching a edge with a new leaf at the root.

For $n \geq 2$ an unrooted phylogenetic tree with $n+1$ leaves can be obtained from a rooted phylogenetic tree with n leaves by adding a new leaf labeled with $n+1$ and an edge between the former root and the new leaf (see Figure 3.4). If the original rooted tree was a binary phylogenetic tree, also the resulting tree will be a binary phylogenetic tree. In the same way, if the original rooted tree was a phylogenetic tree (not necessarily being a binary tree), also the resulting tree will be a multifurcating phylogenetic tree. Therefore we have described two functions, one from the set of rooted phylogenetic trees with n leaves to the set of unrooted phylogenetic trees with $n+1$ leaves and another one from the set of rooted binary phylogenetic trees with n leaves to the set of rooted binary phylogenetic trees with $n+1$ leaves. An inverse map for each of these two functions can be described simply by removing the leaf with the label $n+1$ and its incident edge and rooting the resulting tree at the internal node which was connected to the leaf with the label $n+1$ before. Hence, the two functions are bijective. In that way the number of trees can be deduced immediately from the results in Section 3.1.2 and Section 3.1.1, respectively.

Proposition 3.16. *For $n \geq 2$ there exists a bijective function from the set of rooted phylogenetic trees with n leaves to the set of unrooted phylogenetic trees with $n+1$ leaves. Likewise there exists a bijective function from the set of rooted binary phylogenetic trees with n leaves to the set of unrooted binary phylogenetic trees with $n+1$ leaves.*

Therefore the number of unrooted phylogenetic trees with n leaves is given by r_{n-1} for $n \geq 3$ and the number of unrooted binary phylogenetic trees with n leaves is given by b_{n-1} for $n \geq 3$.

This basic result is also described in [63, prop. 2.2.3, p. 20] and [23, p. 24].

3.1.4 X-trees

In this section the number u_n of X-trees for a set X with $|X| = n$ will be determined as it was done by Foulds and Robinson [33]. In order to do so, we are going to use the number $u_{n,m}$ of X-trees with $|X| = n$ and m vertices and several other classes of X-trees which are rooted in a certain way: *planted*, *point-rooted* and *line-rooted X-trees*. While planted trees are used also elsewhere (see e.g. [79]), point-rooted and line-rooted trees do not seem to be very common terms, but they turn out to be useful to determine u_n.

In addition, the number $\overline{r}_n$ of rooted X-trees in the sense of Definition 2.2 and the mean μ_n and the variance σ_n^2 of the number of vertices of a random X-tree with $n = |X|$ under uniform distribution will be determined. Recursive formulas, allowing to compute values for these sequences and implicit equations for the BGFs and the EGFs, will be provided. We will end this section with asymptotic results, deduced by means of the smooth implicit-function schema presented in Theorem 3.15 in Section 3.1.2.

Planted X-trees. A *planted X-tree* $\mathcal{T} = (T, \phi)$ is similar to a rooted X-tree, but the root vertex ρ has always degree 1 and no label is assigned to ρ. The tree in Figure 2.1d is an X-tree, but it can be perceived also as a planted X-tree. In case of an X-tree ϕ is a label map $\phi : X \to V$. For the tree in Figure 2.1d $\phi(X) \subseteq V \setminus \{\rho\}$ holds, and therefore the label map can be considered to be a function $\phi : X \to V\setminus\{\rho\}$. In this case the tree in Figure 2.1d is a planted X-tree as becomes clear from the following definition.

Definition 3.17 (Planted X-tree). Let X be a finite set, $T = (V, E)$ a tree, $\rho \in V$ a vertex of degree 1, and $\phi : X \to V\setminus\{\rho\}$ a map with $v \in \phi(X)$ for every vertex $v \in V \setminus \{\rho\}$ of degree 1 or 2. The pair $\mathcal{T} = (T, \phi)$ is called *planted X-tree*.

Figure 3.5 illustrates three types of planted X-trees, which will be used later. Throughout the whole section we will stick to the following notation. Note that the root vertex ρ is not counted as vertex in the case of planted X-trees.

Definition 3.18. The number of planted X-trees with $n = |X|$ is denoted by p_n and by $p_{n,m}$ the number of planted X-trees with m vertices where we do not count the vertex ρ, i.e. $m = |V \setminus \{\rho\}|$. The according EGF and BGF are denoted by $P(x) = \sum_{n\geq 1} p_n \frac{x^n}{n!}$ and $P(x, y) = \sum_{n\geq 1}\sum_{m\geq 1} p_{n,m}\frac{x^n y^m}{n!}$, respectively.

Remark 3.19. We consider only X-trees with nonempty label sets $X \neq \emptyset$, and therefore $p_0 = 0$ and $p_{0,m} = 0$ for all $m \geq 0$ and also $p_{n,0} = 0$ for all $n > 0$, because if there are no vertices, the labels cannot be assigned anywhere.

Foulds and Robinson [33] mention the following statement[3] without providing a detailed proof. It allows to write the BGF for planted trees as $P(x,y) = \sum_{n\geq 1} \sum_{m=1}^{2n-1} p_{n,m} \frac{x^n y^m}{n!}$.

Lemma 3.20. *Let $\mathcal{T} = (T, \phi)$ be an X-tree and $\mathcal{T}'$ a planted X-tree with $X \neq \emptyset$. Then the following inequalities for the number of vertices hold*

$$\begin{aligned} 1 &\leq |V(\mathcal{T})| \leq 2 \cdot |X| - 2 \quad \text{if } |X| \geq 2, \\ 1 &\leq \left|V(\mathcal{T}') \setminus \{\rho\}\right| \leq 2 \cdot |X| - 1. \end{aligned}$$

Proof. Clearly $1 \leq |V(\mathcal{T})|$ and $1 \leq |V(\mathcal{T}')|$ holds because $X \neq \emptyset$ and it remains only to prove the other two inequalities. First consider only $\mathcal{T}$. Let $\widetilde{X}$ be the label set obtained from X by removing all labels which are attached to vertices with degree more than 2 and by removing for each vertex with degree 1 or 2 all of its labels but one. The $\widetilde{X}$-tree with label map $\phi|_{\widetilde{X}}$ obtained from $\mathcal{T}$ shall be denoted by $\widetilde{\mathcal{T}} = \left(T, \phi|_{\widetilde{X}}\right)$. If we prove the lemma for $\widetilde{\mathcal{T}}$, it clearly holds also for the original X-tree $\mathcal{T}$, because $\widetilde{X} \subseteq X$ and $\left|\widetilde{X}\right| \leq |X|$ and $\left|V\left(\widetilde{\mathcal{T}}\right)\right| = |V(\mathcal{T})|$ (we did not change the vertex set). The constructed $\widetilde{X}$-tree contains only vertices of degree 1 or 2 with exactly one label and unlabeled vertices of higher degree. If one vertex of degree 2 in $\widetilde{\mathcal{T}}$ is suppressed and its label is removed, we obtain a tree with a vertex set of size $\left|V\left(\widetilde{\mathcal{T}}\right)\right| - 1$ and a label set of size $\left|\widetilde{X}\right| - 1$. Because obviously

$$\left|V\left(\widetilde{\mathcal{T}}\right)\right| - 1 \leq 2\left|\widetilde{X}\right| - 4 \Rightarrow \left|V\left(\widetilde{\mathcal{T}}\right)\right| \leq 2\left|\widetilde{X}\right| - 2$$

we can suppress all nodes of degree 2 in $\widetilde{\mathcal{T}}$ and prove the lemma for the obtained tree. Therefore, by the previous considerations we can assume w.l.o.g. that $\mathcal{T}$ with label set X is a phylogenetic tree. Now, we add a new vertex ρ dividing an arbitrary edge of $\mathcal{T}$ to obtain a rooted phylogenetic tree with $|V(\mathcal{T})| + 1$ vertices and $|X|$ leaves. Lemma 3.6 then implies for this rooted tree $|X| - 1 \geq |V(\mathcal{T})| + 1 - |X|$, which completes the proof for X-trees. To prove the statement for planted X-trees one can follow the same reasoning and apply Lemma 3.6 to the rooted phylogenetic tree obtained from $\mathcal{T}'$ by using the child node of ρ as root vertex and removing the vertex ρ and its incident edge. □

[3] The condition $|X| \geq 2$ is missing in [33]. For an X-tree with $|X| = 1$, of course, $1 = |V(\mathcal{T})| \not\leq 2 \cdot |X| - 2 = 0$.

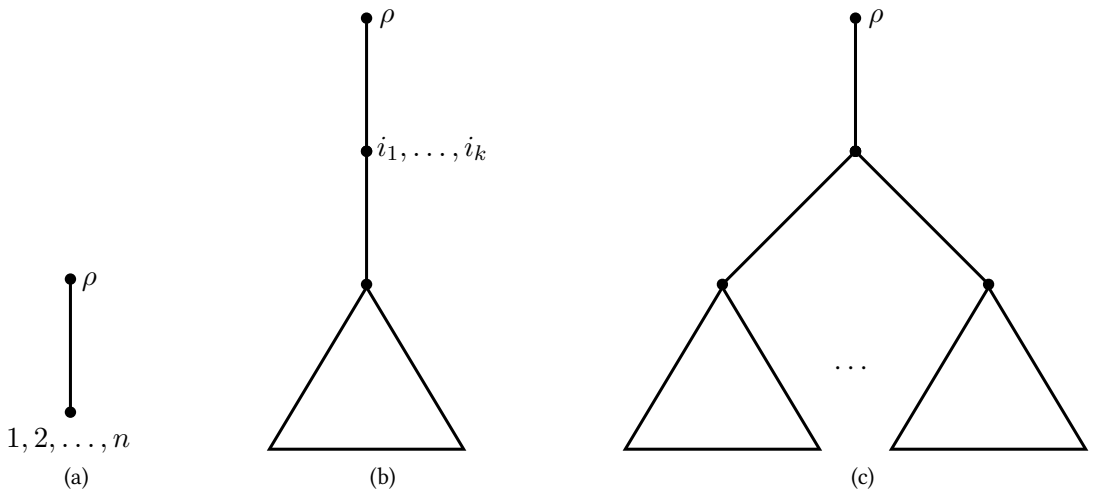


Figure 3.5: Three types of planted X-trees.

Theorem 3.21. *The number $p_{n,m}$ of planted X-trees with $|X| = n$ and m vertices has the BGF $P(x,y) = \sum_{n\geq 1}\sum_{m\geq 1} p_{n,m}\frac{x^n y^m}{n!}$ and the EGF $P(x) = \sum_{n\geq 1} p_n \frac{x^n}{n!}$, for which the following equations hold*

$$\begin{aligned} P(x,y) &= ye^{x+P(x,y)} - yP(x,y) - y \\ P(x) &= e^{x+P(x)} - P(x) - 1. \end{aligned} \tag{3.11}$$

Furthermore $p_0 = 0$, $p_1 = 1$, and for $n \geq 2$ the following recurrence relation for p_n is satisfied

$$p_n = 2p_{n-1} + \sum_{k=1}^{n-1} \binom{n}{k} p_k p_{n-k}. \tag{3.12}$$

Proof. First, we are going to determine implicit expressions for the generating functions $P(x)$ and $P(x,y)$ in order to derive secondly a recursive formula for the number p_n. Each planted X-tree is of one of the three types illustrated in Figure 3.5. Trees of the first type (Figure 3.5a) are called *trivial planted X-trees*. Besides of the root ρ they consist of only one vertex, hence, we have $m = 1$. For every $n \geq 1$ there is only one trivial planted X-tree. Thus, the BGF for these trees is

$$\sum_{n\geq 1} \frac{x^n y}{n!} = (e^x - 1)y. \tag{3.13}$$

For every non-trivial planted X-tree the vertex adjacent to ρ has degree 2 or more—if it has

degree 2 the tree is of the second type (Figure 3.5b) and otherwise of the third type (Figure 3.5c). The trees of the second type (see Figure 3.5b) correspond to planted X-trees, where ρ is replaced by a trivial planted X-tree. The old root is not counted at all, hence we can make use of the labeled product (see Section (2.4) and [27, thm. III.2, p. 175]) to obtain the BGF

$$(e^x - 1)yP(x, y). \tag{3.14}$$

In the third case (Figure 3.5c) the vertex adjacent to ρ does not necessarily have to be labeled, because it has a degree of more than 2. Therefore such trees correspond to a (possible empty) set of labels, one internal node, and a set of at least two planted X-trees. By use of the labeled product and the operators SET and $\text{SET}_{\geq 2}$ this translates to the BGF

$$e^x y \sum_{k\geq 2} \frac{P(x, y)^k}{k!} = e^x y\Big(e^{P(x,y)} - 1 - P(x, y)\Big). \tag{3.15}$$

In total by summation of (3.13), (3.14), and (3.15) this yields

$$P(x, y) = \Big(e^{x+P(x,y)} - P(x, y) - 1\Big)y. \tag{3.16}$$

Since $P(x, 1) = P(x)$ we get for the EGF for planted X-trees

$$P(x) = e^{x+P(x)} - P(x) - 1. \tag{3.17}$$

Now we will establish (3.12) by extracting the coefficients of these equations. Differentiation of (3.17) and using the form $e^{x+P(x)} = 2P(x) + 1$ of (3.17) to eliminate the exponential function leads to

$$P'(x) = 1 + 2P(x) + 2P(x)P'(x) = 1 + 2P(x) + \Big(P(x)^2\Big)'. \tag{3.18}$$

Now by comparing coefficients at both sides we get a recursive formula for p_n as follows. For $n \geq 1$ we have on the left side $(n-1)!\big[x^{n-1}\big]P'(x) = (n-1)!\big[x^{n-1}\big]\sum_{n\geq 1} n \cdot p_n \cdot \frac{x^{n-1}}{n!} = p_n$

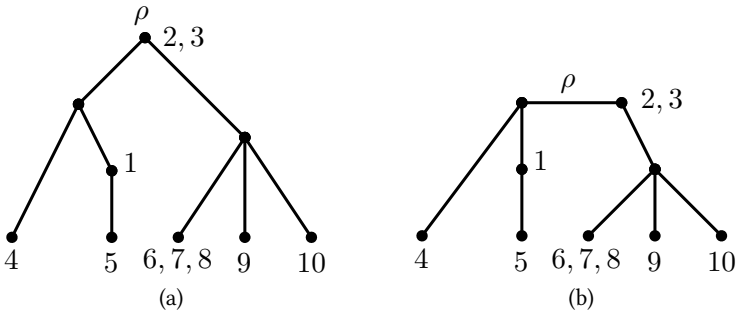


Figure 3.6: Example of a point rooted X-tree and a line-rooted X-tree with $X = \{1, 2, \ldots, 10\}$.

and on the right side $(n-1)!\big[x^{n-1}\big]2P(x) = 2p_{n-1}$ and[4]

$$\begin{aligned}(n-1)!\big[x^{n-1}\big]\Big(P(x)^2\Big)' &= (n-1)!\big[x^{n-1}\big]\left(\sum_{n\geq 1}\left(\sum_{k=0}^{n}\frac{p_k}{k!}\frac{p_{n-k}}{(n-k)!}\right)x^n\right)' \\ &= \sum_{k=0}^{n}\binom{n}{k}p_k p_{n-k}.\end{aligned}$$

As mentioned above $p_0 = 0$ and therefore $p_1 = 1$ and for $n \geq 2$ the claimed recurrence relation (3.12) holds. □

The formula in (3.12) can be used to calculate values for p_n. Values for $n = 1, \ldots, 20$ are given in Table 3.3 on page 72.

Counting X-trees by using planted X-trees. Planted X-trees can be used to determine the number of X-trees. But before doing so, we need two further types of trees.

Definition 3.22. An X-tree is called *point-rooted* if a vertex is distinguished as root ρ (without changing the assignment of the labels in X) and *line-rooted* if an edge is distinguished as root ρ.

Examples in Figure 3.6 illustrate point-rooted and line-rooted X-trees. In the proof of Theorem 3.25 we will count the number of X-trees by subtracting the number of line-rooted X-trees from the number of point-rooted X-trees. The following lemma will prove useful.

[4] Recall that for the product of two generating functions we have $\left(\sum_{n\geq 0} a_n x^n\right) \cdot \left(\sum_{n\geq 0} b_n x^n\right) = \sum_{n\geq 0}\sum_{k=0}^{n} a_n b_{n-k} x^n$ (see [83, p. 36]).

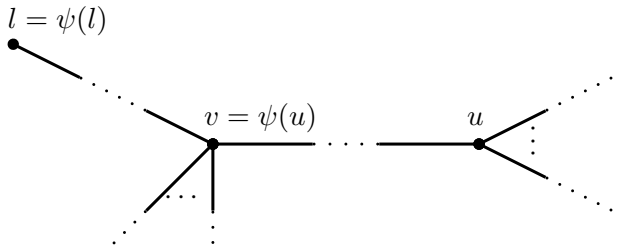


Figure 3.7: A function $\psi : V \to V$ with $\psi(v) = v$ for all leaves v and $\psi(u) = v$ for some vertices $u, v \in V$ with $u \neq v$ can not be a graph automorphism.

Lemma 3.23. *The only graph automorphism*[5] ψ *of an* X*-tree* $\mathcal{T} = (T, \phi)$ *with* $T = (V, E)$ *preserving also its labeling, that is* $\phi(x) = v \Leftrightarrow \phi(x) = \psi(v)$ *for all* $x \in X$*, is the identity map* $\psi(v) = v$ *for all* $v \in V$*.*

Proof. For any vertex with at least one label and therefore in particular for all leaves, we have $\psi(v) = v$. Assume there are internal vertices $u, v \in V$ with $u \neq v$ and $\psi(u) = v$ (illustrated in Figure 3.7). Now consider the unique path from u to v denoted by p_1 and a path p_2 from v to some leaf l. We can choose p_2 to be edge-disjoint with p_1, because v is not a leaf and therefore there exists an edge directed away from u. Following this edge and then any arbitrary path until a leaf is reached, is a valid construction for p_2. We now have two paths of different lengths: p_1 concatenated with p_2 results in a path between l and u, and p_2 is a path between $\psi(l) = l$ and $\psi(u) = v$. This is a contradiction to ψ being an isomorphic map between trees because a path between two vertices in a tree is unique. □

Definition 3.24. The set of (unrooted) X-trees with $n = |X|$ is denoted by $\mathcal{U}_n$ and $u_n := |\mathcal{U}_n|$ number of these trees. By $u_{n,m}$ we denote the number of X-trees with $m = |V|$ vertices. The according EGF and BGF are denoted by

$$U(x) = \sum_{n \geq 1} u_n \frac{x^n}{n!}$$

and

$$U(x, y) = \sum_{n \geq 1} \sum_{m \geq 1} u_{n,m} \frac{x^n y^m}{n!},$$

respectively.

[5] A *graph automorphism* of a graph $G = (V, E)$ is a bijective map $\psi : V \to V$ preserving the graph structure, that is $\{v, u\} \in E \Leftrightarrow \{\psi(v), \psi(u)\} \in E$ for all $v, u \in V$.

Theorem 3.25. *The number $u_{n,m}$ of X-trees with $|X| = n$ and m vertices has the BGF $U(x,y) = \sum_{n\geq 1}\sum_{m\geq 1} u_{n,m}\frac{x^n y^m}{n!}$ and the EGF $U(x) = \sum_{n\geq 1} u_n \frac{x^n}{n!}$, for which the following equations hold*

$$\begin{aligned} U(x,y) &= P(x,y) - (1+y)\frac{P(x,y)^2}{2} \\ U(x) &= P(x) - P(x)^2. \end{aligned} \tag{3.19}$$

Furthermore $u_1 = 1$ and for $n \geq 2$

$$u_n = 2p_{n-1}. \tag{3.20}$$

Proof. Consider an arbitrary (unrooted) X-tree $\mathcal{T}$ with m vertices and $n = |X|$. m point-rooted X-trees can be obtained from $\mathcal{T}$, because there are m possible vertices in $\mathcal{T}$ to place ρ. Lemma 3.23 implies that these m point-rooted X-trees are pairwise distinct[6]. In the same way $m-1$ line-rooted X-trees can be obtained from $\mathcal{T}$, because there are $m-1$ edges to place ρ, and again they are distinct because of Lemma 3.23. This allows us to count the number of unrooted X-trees by counting the number of point-rooted X-trees and then subtracting the number of line-rooted X-trees.

Each point-rooted X-tree corresponds to exactly one planted X-tree except the case where the root of the point-rooted X-tree has degree 2. In this case the root of the point-rooted X-tree has to be labeled, but the vertex adjacent to the root of the corresponding planted X-tree does not have to be labeled necessarily. Hence, in order to count the number of point-rooted X-trees, we have to count planted X-trees and remove the number of planted X-trees, where the vertex adjacent to the root has no label and has exactly 2 child vertices, which leads to the generating function

$$P(x,y) - y\frac{P(x,y)^2}{2}.$$

Each line-rooted X-tree corresponds to two planted X-trees connected at their roots. Thus, the number of line-rooted X-trees is given by the generating function

$$\frac{P(x,y)^2}{2}.$$

In total this yields

$$U(x,y) = P(x,y) - (1+y)\frac{P(x,y)^2}{2}.$$

[6] Recall that *counting trees* actually means to determine the number of isomorphism classes of trees of the considered type. Therefore *distinct* is used synonymously with *not isomorphic*.

By setting $y = 1$ we get

$$U(x) = P(x) - P(x)^2, \tag{3.21}$$

where $U(x) = \sum_{n\geq 1} u_n \frac{x^n}{n!}$ is the EGF for the number u_n of X-trees with $|X| = n$. By differentiation and use of (3.18) one gets

$$U'(x) = 1 + 2P(x).$$

Extracting the coefficient of $\frac{x^{n-1}}{(n-1)!}$ yields $u_1 = 1$ and for $n \geq 2$

$$u_n = 2 \cdot p_{n-1}.$$

□

Together with (3.12) in Theorem 3.21 this enables us to compute values for u_n (see Table 3.3 on page 72 for values for $n = 1, \ldots, 20$).

The number of rooted X-trees. Foulds and Robinson [33] focus on counting the number of unrooted X-trees and introduce planted, point-rooted, and line-rooted X-trees only to use them as tools. Nevertheless, in the same way we can derive easily the number $\overline{r}_n$ of rooted X-trees in the sense of Definition 2.2. (This case was not covered by Foulds and Robinson [33].)

Definition 3.26. The number of rooted X-trees with $|X| = n$ and m vertices is denoted by $\overline{r}_{n,m}$, the BGF by $\overline{R}(x, y) = \sum_{n\geq 0} \sum_{m\geq 0} \overline{r}_{n,m} \frac{x^n y^m}{n!}$ and the EGF by $\overline{R}(x) = \sum_{n\geq 0} \overline{r}_n \frac{x^n}{n!}$.

Theorem 3.27. *The BGF $\overline{R}(x, y)$ and the EGF $\overline{R}(x)$ satisfy the following equations*

$$\begin{aligned} \overline{R}(x, y) &= P(x, y) + yP(x, y), \\ \overline{R}(x) &= 2 \cdot P(x). \end{aligned}$$

Furthermore for $n \geq 1$

$$\overline{r}_n = 2 \cdot p_n = u_{n+1}.$$

Proof. Each rooted X-tree can be constructed either from a planted X-tree by merging ρ and the vertex adjacent to ρ (assigning also its labels) or from a planted X-tree and an additional vertex without labels as root of the rooted X-tree. The latter case corresponds to rooted X-trees where ρ has degree 1 and no labels from X are assigned to it (note that we defined for rooted

phylogenetic trees that ρ has to have at least degree 2, but in the case of X-trees ρ can have an arbitrary degree and does not need to be labeled). Figure 3.5 makes this correspondence more clear. In all three cases a rooted X-tree can be obtained, but in Figure 3.5b the vertex adjacent to the root of the planted X-tree has to be labeled while the root vertex in the corresponding X-tree does not need a label necessarily. So we have to add trees of the type illustrated in 3.5b where the vertex adjacent to the root is not labeled.

The number of planted X-trees has the BGF $P(x, y)$ while a single vertex without labels has the BGF $x^0 y$. Thus, in total we have

$$\begin{aligned} \overline{R}(x, y) &= P(x, y) + yP(x, y) \\ \overline{R}(x) &= 2 \cdot P(x) \end{aligned}$$

and therefore using (3.20) in Theorem 3.25 for $n \geq 1$

$$\overline{r}_n = 2 \cdot p_n = u_{n+1}.$$

□

Remark 3.28. The relation between rooted and unrooted X-trees can be explained also combinatorially, following a similar way as in Section 3.1.3. Given any unrooted X'-tree with $X' = \{1, 2, \ldots, n, n+1\}$ a rooted X-tree with $X = \{1, 2, \ldots, n\}$ is constructed by rooting the tree at the vertex $\phi(n+1)$ and removing the label $n+1$ (note that the root does not need to be labeled necessarily also if it has degree less than 3). This describes a map from unrooted X'-trees with $|X'| = n+1$ to the rooted X-trees with $|X| = n$. Clearly, there exists an inverse map and therefore the map is bijective and we have $\overline{r}_n = u_{n+1}$.

Mean and variance. In the following section for each $n \geq 1$ the mean μ_n and the variance σ_n^2 for the number of vertices of a random X-tree with a label set of size n will be determined.

Theorem 3.29. *The mean μ_n of the number of vertices in a random X-tree with $|X| = n$ under uniform distribution on $\mathcal{U}_n$ is given by $\mu_1 = 1$ and for $n \geq 2$*

$$\mu_n = \frac{\frac{1}{2}p_n + p_{n-1}}{u_n}.$$

The variance σ_n^2 of the number of vertices in a random X-tree with $|X| = n$ is given for all $n \geq 1$

by

$$\sigma_n^2 = \frac{a_n}{u_n} + \mu_n - \mu_n^2,$$

where $a_0 = 0$, $a_1 = 0$ *and for* $n \geq 2$

$$a_n = p_n - 2p_{n-1} + \sum_{k=1}^{n-1} \binom{n}{k} p_n a_{n-k}.$$

Proof. It is a well-known fact that the mean and variance can be found by use of the bivariate generating function and its derivatives as follows (see also Section 2.4 and [27, p. 158f.])

$$\mu_n = \frac{[x^n]U_y(x,1)}{[x^n]U(x,1)} \tag{3.22}$$

$$\sigma_n^2 = \frac{[x^n]U_{yy}(x,1) + U_y(x,1)}{[x^n]U(x,1)} - \mu_n^2, \tag{3.23}$$

where $[x^n]U(x,1) = [x^n]U(x) = \frac{1}{n!} \cdot u_n$. In the following we will deduce recursive formulas from the generating functions $U(x,y)$ and $P(x,y)$ in order to compute these quantities.

Differentiation of (3.19) and (3.11) with respect to y and using (3.11) again to eliminate the exponential yields

$$\begin{aligned} U_y(x,y) &= P_y(x,y) - \frac{P(x,y)^2}{2} - (1+y)P(x,y) \cdot P_y(x,y) \\ P_y(x,y) &= \underbrace{e^{x+P(x,y)} - P(x,y) - 1}_{=\frac{P(x,y)}{y}} + \underbrace{y \cdot \left(e^{x+P(x,y)} P_y(x,y) - P_y(x,y)\right)}_{=P_y(x,y)\cdot P(x,y)\cdot(y+1)} \\ &= \frac{P(x,y)}{y} + (y+1) \cdot P(x,y) \cdot P_y(x,y) \end{aligned} \tag{3.24}$$

and these two equations together simplify to

$$U_y(x,y) = \frac{P(x,y)}{y} - \frac{P(x,y)^2}{2}. \tag{3.25}$$

Now again, differentiating with respect to y and using (3.24) in the form

$$y \cdot P_y(x,y) = \frac{P(x,y)}{1 - (y+1)P(x,y)}$$

yields

$$\begin{aligned} U_{yy}(x,y) &= \frac{P_y(x,y)y - P(x,y)}{y^2} - P(x,y)P_y(x,y) \\ &= \frac{1}{y^2}\cdot\left(\frac{P(x,y)}{1-(y+1)P(x,y)} - P(x,y) - \frac{yP(x,y)}{1-(y+1)P(x,y)}\right) \\ &= \frac{P(x,y)^2}{y^2(1-(y+1)P(x,y))}. \end{aligned} \tag{3.26}$$

The derivatives of $U(x,y)$ in (3.25) and (3.26) allow us to determine the mean and the variance. By setting $y=1$ in (3.25) and by using (3.12) in the form $\frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}p_k p_{n-k} = \frac{1}{2}p_n - p_{n-1}$ we finally have a formula for the mean μ_n from (3.22), namely $\mu_1 = 1$ and for $n \geq 2$

$$\mu_n = \frac{n!\cdot[x^n]P(x) - \frac{1}{2}P(x)^2}{u_n} = \frac{p_n - \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}p_k p_{n-k}}{u_n} = \frac{\frac{1}{2}p_n + p_{n-1}}{u_n}. \tag{3.27}$$

In order to determine a formula for the variance σ_n^2, set $y=1$ in (3.26) and denote the resulting EGF by $A(x)$ which yields

$$U_{yy}(x,1) = \frac{P(x)^2}{1-2\cdot P(x)} =: A(x).$$

Further denote its coefficients by $a_n := n![x^n]A(x)$. Using the form $A(x) = P(x)^2 + 2P(x)A(x)$ and again (3.12) as previously, yields $a_0 = 0$, $a_1 = 0$ and for $n \geq 2$

$$a_n = p_n - 2p_{n-1} + 2\sum_{k=1}^{n-1}\binom{n}{k}p_n a_{n-k}.$$

From (3.23) we get for the variance σ_n^2 gives for $n \geq 1$

$$\sigma_n^2 = \frac{a_n}{u_n} + \mu_n - \mu_n^2, \tag{3.28}$$

where a_n, u_n and μ_n can be computed as previously stated. □

Asymptotic analysis. Asymptotic results for the sequences under consideration can be determined by applying Theorem 3.15 to the equation for the EGF $P(x)$ in (3.17). With the terminol-

n	p_n	u_n	μ_n	σ_n^2
1	1	1	1.	0.
2	4	2	1.5	0.25
3	32	8	2.5	0.75
4	416	64	3.75	1.1875
5	7552	832	5.03846	1.57544
6	176128	15104	6.33051	1.95856
7	5018624	352256	7.62355	2.34084
8	168968192	10037248	8.91706	2.72267
9	6563282944	337936384	10.2108	3.10424
10	288909131776	13126565888	11.5047	3.48565
11	14212910809088	577818263552	12.7988	3.86697
12	772776684683264	28425821618176	14.0929	4.24822
13	46017323176296448	1545553369366528	15.387	4.62942
14	2978458881388183552	92034646352592896	16.6812	5.01058
15	208198894960190160896	5956917762776367104	17.9754	5.39172
16	15631251601179130462208	416397789920380321792	19.2696	5.77283
17	1254492810303112820555776	3126250320235826092441 6	20.5639	6.15393
18	10717440394145143468746342 4	2508985620606225641111552	21.8581	6.53502
19	97110224589894382553000837 12	21434880788290286937492 6848	23.1524	6.91609
20	93018622400042824880715569 5616	19422044917978876510600167424	24.4467	7.29715

Table 3.3: The number of planted X-trees p_n, the number of X-trees u_n and the mean μ_n and the variance σ_n^2 of the number of vertices of an X-tree where $n = |X|$. The values were computed with Mathematica by means of the deduced recursive formulas (see Section A.2). Note that the second column contains also values for $\overline{r}_n$ because $r_{n-1} = u_n$ for $n \geq 2$.

ogy of Theorem 3.15 we have

$$\begin{aligned} G(z,w) &= e^{z+w} - w - 1 \\ &= -w + \sum_{n\geq 1} \frac{(z+w)^n}{n!} \\ &= -w + \sum_{n\geq 1} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} z^k w^{n-k}. \end{aligned}$$

Hence, $g_{0,0} = 0$, $g_{0,1} = 0 \neq 1$, $g_{0,2} = \frac{1}{2} > 0$ and $g_{m,n} \geq 0$. Furthermore, from $G_w(s_1, s_2) = e^{s_1+s_2} - 1 = 1$ follows $\ln 2 = s_1 + s_2$, and therefore with $G(s_1, s_2) = e^{s_1+s_2} - s_2 - 1 = s_2$ we have $s_2 = \frac{1}{2}$ and $s_1 = \ln 2 - \frac{1}{2}$. Thus, the conditions of Theorem 3.15 are fulfilled and the

following asymptotic expression for the number p_n of planted trees holds

$$\frac{p_n}{n!} \sim \sqrt{\frac{\ln 2 - \frac{1}{2}}{2\pi}} \cdot n^{-\frac{3}{2}} \cdot \left(\ln 2 - \frac{1}{2}\right)^{-n}.$$

By use of (3.20) we have

$$\frac{u_n}{n!} = \frac{2}{n}\frac{p_{n-1}}{(n-1)!} \sim 2 \cdot \sqrt{\frac{\ln 2 - \frac{1}{2}}{2\pi}} \cdot n^{-\frac{5}{2}} \cdot \left(\ln 2 - \frac{1}{2}\right)^{-n+1}$$

and

$$\frac{\overline{r}_n}{n!} = \frac{2 \cdot p_n}{n!} \sim 2 \cdot \sqrt{\frac{\ln 2 - \frac{1}{2}}{2\pi}} \cdot n^{-\frac{3}{2}} \cdot \left(\ln 2 - \frac{1}{2}\right)^{-n}.$$

These results coincide with the findings in [33, p. 116ff.] where Theorem 3.15 was not used.

3.2 Expected parsimony score

In Section 2.2 the parsimony score of a character $\chi : X \to C$ on a phylogenetic tree with label set X was introduced as the changing number of a minimum character extension for χ

$$l(\chi, \mathcal{T}) = \min_{\overline{\chi}} \mathrm{ch}(\overline{\chi}).$$

In this section the expected parsimony score for a random character on a fixed tree and the expected parsimony score for a fixed character on a random tree will be determined as it is done in [63, chap. 5.6].

Random characters on a fixed tree. In Section 2.3 Markov models on phylogenetic trees were presented. This is one natural way to define a probability distribution on the set of possible character extensions for a given rooted phylogenetic tree. At the same time also a probability distribution on the set of characters is induced, since a character can be considered as a equivalence class of character extensions. In this section a different way will be followed to obtain a probability distribution for random characters, namely simply a uniform distribution on the set of all $|C|^{|X|}$ characters. We follow here the approach of Semple and Steel [63, chap. 5.6].

Definition 3.30. Let $\mathcal{T} = (T, \phi)$ be a phylogenetic tree with label set X and $\chi : X \to C$ a random character on $\mathcal{T}$, where the values $\chi(x) \in C$ for $x \in X$ are independent and uniformly

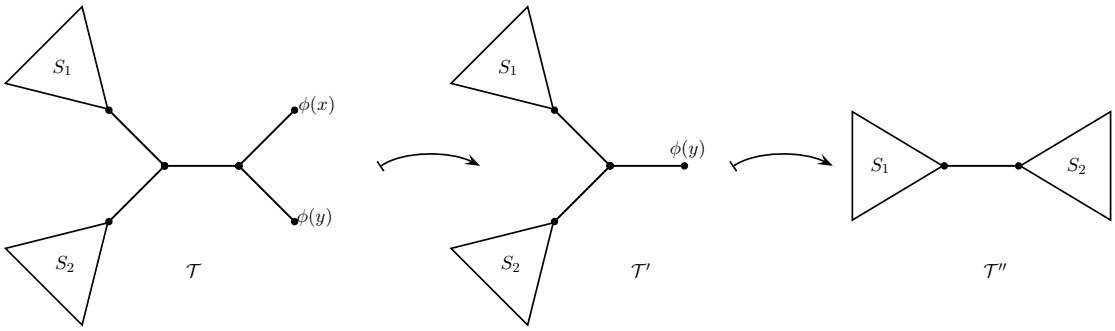


Figure 3.8: $\mathcal{T}'$ and $\mathcal{T}''$ are obtained from $\mathcal{T}$ as described in Definition 3.31. This figure illustrates the case $|X| \geq 4$ (the figure is inspired by [63, fig. 5.9,p. 103]).

distributed. The random variable given by the parsimony score of χ is denoted by $L_r(\mathcal{T})$ for $r = |C|$.

Note that $L_r(\mathcal{T})$ only depends on the number of states $|C| = r$ but not on C itself. For binary phylogenetic trees and $r = 2$ a closed form expression for $L_r(\mathcal{T})$ can be established. Surprisingly, this expression is independent from T and depends only on the number of labels $|X|$ as will be shown in Theorem 3.33. To prove the theorem another definition and a lemma are helpful.

Definition 3.31. Let $\mathcal{T} = (T, \phi)$ be a binary phylogenetic tree with label set X and $|X| \geq 3$. The trees $\mathcal{T}'$ and $\mathcal{T}''$ are obtained from $\mathcal{T}$ in the following way (Figure 3.8 illustrates this construction). Let $x, y \in X$ such that $\phi(x)$ and $\phi(y)$ form a cherry in T. To obtain the tree T' remove the vertex $\phi(x)$ from T and its attached edge and suppress all vertices of degree 2. $\mathcal{T}'$ is the phylogenetic tree $\big(T', \phi|_{X\setminus\{x\}}\big)$ with label set $X \setminus \{x\}$. $\mathcal{T}''$ is obtained from $\mathcal{T}'$ in the same way be removing $\phi(y)$ and its attached edge and by suppressing all vertices of degree 2.

Note that for any phylogenetic tree $\mathcal{T} = (T, \phi)$ with label set X and $|X| \geq 3$ there exist $x, y \in X$ such that $\phi(x)$ and $\phi(y)$ form a cherry in T (see [63, prop. 1.2.5, p. 8]).

Lemma 3.32. *For a binary phylogenetic tree $\mathcal{T}$ with label set X and $|X| \geq 3$ and for $\mathcal{T}'$ and $\mathcal{T}''$ as in Definition 3.31 the probability distribution of $L_2(\mathcal{T})$ satisfies the recursive formula*

$$\mathbb{P}(L_2(\mathcal{T}) = k) = \frac{1}{2}\mathbb{P}\big(L_2\big(\mathcal{T}'\big) = k\big) + \frac{1}{2}\mathbb{P}\big(L_2\big(\mathcal{T}''\big) = k - 1\big).$$

Proof. Let $x, y \in X$ as in Definition 3.31 and let χ be a random character as in Definition 3.30 with $r = |C| = 2$. By applying the law of total probability we get

$$\mathbb{P}(L_2(\mathcal{T}) = k) = \mathbb{P}(E) \cdot \mathbb{P}(L_2(\mathcal{T}) = k|E) + \mathbb{P}(\neg E) \cdot \mathbb{P}(L_2(\mathcal{T}) = k|\neg E),$$

where E denotes the probability event $\{\chi(x) = \chi(y)\}$. Since the values of χ are independent and uniformly distributed, we have $\mathbb{P}(E) = \mathbb{P}(\neg E) = \frac{1}{2}$. Furthermore, given that E occurs, we have $l(\chi, \mathcal{T}) = l\big(\chi|_{X\setminus\{x\}}, \mathcal{T}'\big)$, and given that $\neg E$ occurs, we have $l(\chi, \mathcal{T}) - 1 = l\big(\chi|_{X\setminus\{y,x\}}, \mathcal{T}''\big)$. Therefore $\mathbb{P}(L_2(\mathcal{T}) = k|E) = \mathbb{P}(L_2(\mathcal{T}') = k)$ and $\mathbb{P}(L_2(\mathcal{T}) = k|\neg E) = \mathbb{P}(L_2(\mathcal{T}'') = k - 1)$, which completes the proof. □

This lemma allows to express the probability distribution of $L_2(\mathcal{T})$ for binary phylogenetic trees explicitly and to determine the mean and the variance.

Theorem 3.33. *Let $\mathcal{T}$ be a phylogenetic binary tree with label set $X \neq \emptyset$ and $n = |X|$ the number of labels.*

(i) The probability distribution of the random variable $L_2(\mathcal{T})$ depends only on n and is given by

$$\mathbb{P}(L_2(\mathcal{T}) = k) = \begin{cases} \frac{2n-3k}{k}\binom{n-k-1}{k-1}2^{k-n}, & \text{if } 1 \leq k \leq \frac{n}{2}, \\ 2^{1-n}, & \text{if } k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) The mean $\mu_n = \mathbb{E}(L_2(\mathcal{T}))$ and the variance $\sigma_n^2 = \mathbb{V}(L_2(\mathcal{T}))$ are given by

$$\mu_n = \frac{1}{9}\left(3n - 2 - \left(-\frac{1}{2}\right)^{n-1}\right) \sim \frac{n}{3}$$

and

$$\sigma_n^2 = \frac{1}{81}\left(6n + 2 - (6n+1)\cdot\left(-\frac{1}{2}\right)^{n-1} - \left(-\frac{1}{2}\right)^{2n-2}\right) \sim \frac{2n}{27}.$$

Proof. First we want to prove that $L_2(\mathcal{T})$ depends only on n by induction. If $n \leq 2$, the random variable $L_2(\mathcal{T})$ is obviously independent of T since there is only one possible tree T with n leaves for each $n = 1, 2$. Now assume $L_2(\mathcal{T}_1) = L_2(\mathcal{T}_2)$ for any two trees with $m < n$ leaves. Then, by Lemma 3.32 we have also $L_2(\mathcal{T}_1) = L_2(\mathcal{T}_2)$ for two trees with n leaves. Hence, $L_2(\mathcal{T})$ depends only on n and it makes sense to define $l_{n,k} := \mathbb{P}(L_2(\mathcal{T}) = k)$ for all $k \geq 0$ and some arbitrary representative $\mathcal{T}$ with n labels. Furthermore denote by

$$L(x, y) := \sum_{n\geq 1}\sum_{k\geq 0} l_{n,k}x^n y^k$$

the according BGF and by $L_3(x, y) := \sum_{n\geq 3}\sum_{k\geq 0} l_{n,k}x^n y^k$ the BGF for $n \geq 3$. If $n = 1$

clearly $l(\chi, \mathcal{T}) = 0$ and if $n = 2$ we have $l(\chi, \mathcal{T}) = 0$ or $l(\chi, \mathcal{T}) = 1$, each with probability $\frac{1}{2}$. Hence, we have

$$L(x, y) = x + \frac{1}{2}x^2 + \frac{1}{2}x^2 y + L_3(x, y). \tag{3.29}$$

From Lemma 3.32 follows

$$l_{n,k} = \frac{1}{2}l_{n-1,k} + \frac{1}{2}l_{n-2,k-1}$$

and therefore

$$L_3(x, y) = \frac{1}{2}x(L(x, y) - x) + \frac{1}{2}x^2 y L(x, y).$$

Together with (3.29) this yields

$$L(x, y) = x + \frac{1}{2}x^2 y + \frac{1}{2}xL(x, y) + \frac{1}{2}x^2 y L(x, y).$$

Solving this equation for $L(x, y)$ gives an explicit expression for the BGF of $l_{n,k}$

$$L(x, y) = \frac{x + \frac{1}{2}x^2 y}{1 - \frac{1}{2}x - \frac{1}{2}x^2 y}.$$

To get the coefficients $l_{n,k} = \left[x^n y^k\right] L(x, y)$, first determine the coefficients of the denominator of $L(x, y)$ with the help of the identity $(1 - A)^{-1} = \sum_{i \geq 0} A^i$ and the binomial theorem

$$\begin{aligned}
\left[x^n y^k\right] \left(1 - \left(\frac{1}{2}x + \frac{1}{2}x^2 y\right)\right)^{-1} &= \sum_{i \geq 0} \left[x^n y^k\right] \frac{1}{2^i} x^i (1 + xy)^i = \\
&= \sum_{i \geq 0} \left[x^n y^k\right] \frac{1}{2^i} x^i \sum_{j=0}^{i} \binom{i}{j} x^j y^j \\
&= \binom{n-k}{k} \cdot \frac{1}{2^{n-k}}.
\end{aligned} \tag{3.30}$$

Then one has for $n \geq 2$ and $k \geq 1$

$$\begin{aligned}
\left[x^n y^k\right] L(x, y) &= \left[x^n y^k\right] \frac{x}{1 - \frac{1}{2}x(1 + xy)} + \left[x^n y^k\right] \frac{1}{2} \frac{x^2 y}{1 - \frac{1}{2}x(1 + xy)} \\
&= \left[x^{n-1} y^k\right] \frac{1}{1 - \frac{1}{2}x(1 + xy)} + \frac{1}{2}\left[x^{n-2} y^{k-1}\right] \frac{1}{1 - \frac{1}{2}x(1 + xy)}.
\end{aligned}$$

With (3.30) this yields

$$\begin{aligned}\left[x^n y^k\right] L(x,y) &= \binom{n-k-1}{k}\frac{1}{2^{n-1-k}} + \frac{1}{2}\binom{n-k-1}{k-1}\frac{1}{2^{n-1-k}} \\ &= \frac{1}{2^{n-k}}\binom{n-k-1}{k-1}\left(\frac{2(n-2k)}{k}+1\right).\end{aligned}$$

If $k = 0$, all states of χ are equal and therefore $\mathbb{P}(L_2(\mathcal{T}) = 0) = \frac{2}{2^n} = 2^{1-n}$, which completes the proof of (i).

In oder to determine μ_n and σ_n^2 we make use of (see Section 2.4 and [27, p. 158f.])

$$\begin{aligned}\mu_n &= [x^n]L_y(x,1) \\ \sigma_n^2 &= [x^n]L_{yy}(x,1) + \mu_n - \mu_n^2.\end{aligned}$$

Note that it is not necessary to divide through $L(x,1)$, because the coefficients of $L(x,y)$ are probabilities and $L(x,y)$ is not a generating function of a counting sequence. The claimed results for μ_n and σ_n^2 then can be established in a similar way (e.g. by partial fraction decomposition and the generalized binomial theorem).

□

A fixed character on a random tree. If $\chi : X \to C$ is a character on X with state set C, $|C| = r$ and $k \in \mathbb{N}$, in general the enumeration problem, how many trees $\mathcal{T}$ with $l(\chi, \mathcal{T}) = k$ exist, is unsolved (see [75]). However, there are results by Carter et al. [9] for two special cases (see also [75] and [63, p. 105ff.]). Results for other special cases are summarized in [75].

Definition 3.34. Let $C = \{\alpha_1, \ldots, \alpha_r\}$ be a set of character states with $r \geq 2$, X a set of labels and $\chi : X \to C$ a character on X. For $i \in \{1, 2, \ldots, r\}$ let denote $a_i := \left|\chi^{-1}(\alpha_i)\right|$ the number of labels with state α_i. Then $p_l(a_1, a_2, \ldots, a_r)$ denotes the probability that $l(\chi, \mathcal{T}) = l$, where $\mathcal{T}$ is a (unrooted) binary phylogenetic tree selected uniformly from all binary phylogenetic trees with label set X.

In the first solved special case the character is binary, that is $r = 2$, and in the second case we have $l = r - 1$.

Theorem 3.35. *Let $r = 2$ and let χ be a character as in the previous definition. The probability*

that $l(\chi, \mathcal{T}) = l$ *is given by*

$$p_l(a_1, a_2) = 2^l \cdot \frac{(2n-3l)(2a_1-l-1)!(2a_2-l-1)!(n-l)!}{(a_1-l)!(a_2-l)(l-1)!(2n-2l)!}.$$

A proof can be found in Carter et al. [9]. A different proof was given by Steel [73] and refined by Erdős and Székely [20].

Theorem 3.36. *For* $a_1, \ldots, a_r$ *as in the previous definition, the probability that* $l(\chi, \mathcal{T}) = r - 1$ *is given by*

$$p_{r-1}(a_1, \ldots, a_r) = \frac{1}{b_{n-r+1}} \prod_{i=1}^{r} b_{a_i},$$

where b_n *is the number of rooted binary phylogenetic trees with* n *leaves (see Theorem 3.2).*

A proof is given in [9] and in [63, p. 104f.].

3.3 Isomorphism between phylogenetic trees

In this section the probability p_n, that two random rooted binary phylogenetic trees are isomorphic, will be determined as it was done by Bóna and Flajolet [6]. We will define isomorphism between phylogenetic trees by ignoring their labels and comparing only their tree shapes.

Definition 3.37. Two phylogenetic trees $\mathcal{T}_1 = (T_1, \phi_1)$ and $\mathcal{T}_2 = (T_2, \phi_2)$ are *isomorphic*, denoted by $\mathcal{T}_1 \cong \mathcal{T}_2$, if they share the same tree shape, i.e. if T_1 and T_2 are equal[7].

Hence, each isomorphism class $[\mathcal{T}]_{\cong}$ of a phylogenetic tree $\mathcal{T}$ corresponds to an unlabeled unordered tree T. If the phylogenetic trees $\mathcal{T}$ are rooted and binary, also the corresponding trees T are rooted and binary, also called Otter trees.

Definition 3.38 (Otter trees). The set of unlabeled and unordered rooted binary trees with n leaves is denoted by $\mathcal{O}_n$ and $\mathcal{O} := \cup_{n\geq 1}\mathcal{O}_n$. Furthermore let $o_n := |\mathcal{O}_n|$ be their counting sequence and $O(z) = \sum_{n\geq 1} o_n z^n$ the according OGF[8].

The sequence o_n is also referred to as the *Wedderburn-Etherington numbers* and listed as A001190 in [64]. The OGF $O(z)$ can be obtained by the symbolic method in the following way.

[7] Recall that we speak of equal or identical graphs if they are isomorphic. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic*, if there is a bijective map $\psi : V_1 \to V_2$, with $\{u, v\} \in E_1 \Leftrightarrow \{\psi(u), \psi(v)\} \in E_2$ and in the case of rooted trees $\psi(\rho_1) = \rho_2$ for the root ρ_1 of G_1 and the root ρ_2 of G_2.

[8] $O(z)$ is the OGF for Otter trees and should not to be confused with the Landau notation $O(.)$, which will not be used in this section.

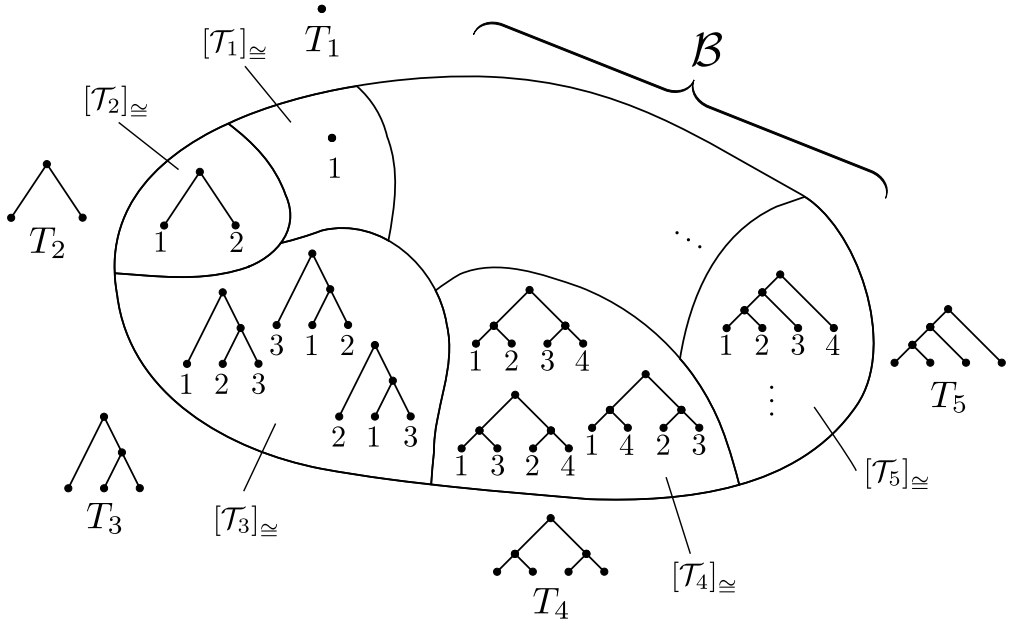


Figure 3.9: Illustration of the isomorphism classes $[\mathcal{T}_i]_{\cong}$ of $\mathcal{B}$ and their correspondence to Otter trees $T_i \in \mathcal{O}$, where $i = 1, \ldots, 5$.

Every Otter tree has either only one vertex or can be constructed by connecting two Otter trees to a new vertex, which is the root vertex of the resulting tree. Since Otter trees are unordered, each such tree corresponds to a multiset of two of its kind

$$\mathcal{O} = \mathcal{Z} + \textsc{MSet}_2(\mathcal{O}).$$

The unlabeled $\textsc{MSet}_2$-operator translates in the following way to a OGF (see Section e(2.4) and [27, fig. I.18, p. 93])

$$O(z) = z + \frac{1}{2}\Big(O(z)^2 + O\big(z^2\big)\Big). \tag{3.31}$$

A bivariate generating function. The desired probability p_n now can be described as the probability that two rooted binary phylogenetic trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathcal{B}_n$ are both in the same isomorphism class. Since we know already b_n (see Section 3.1.1), we need to determine the size of each isomorphism class. To do so, we first introduce the term *symmetry vertex* (see [63, sect. 2.4]) and then express the probability p_n in terms of the coefficients of a BGF.

Definition 3.39. Let T be an Otter tree. An internal vertex $v \in V(T)$ is called *symmetry vertex*, if the two subtrees of T rooted at v are identical. $\mathrm{sym}(T)$ denotes the number of symmetry

vertices $v \in V(T)$.

The next lemma describes the number of labelings of the leaves of an Otter tree.

Lemma 3.40. *Let T be an Otter tree with n leaves. The number of different rooted (binary) phylogenetic trees $\mathcal{T} = (T, \phi)$ for the fixed tree T and $\phi : \{1, \dots, n\} \to V(T)$ being a label map, i.e. the number of leaf labelings for the tree T, is given by*

$$w(T) = \frac{n!}{2^{\mathrm{sym}(T)}}. \tag{3.32}$$

$w(T)$ equals also the size of the isomorphism class of $\mathcal{B}$ corresponding to T (see Figure 3.9). Furthermore summation over all Otter trees with n leaves, gives the number of rooted binary phylogenetic trees

$$\sum_{T \in \mathcal{O}_n} \frac{n!}{2^{\mathrm{sym}(T)}} = \sum_{T \in \mathcal{O}_n} w(T) = b_n = (2n-3)!!. \tag{3.33}$$

While (3.32) is stated without proof in [6], in [63, sect. 2.4] it is shown with the help of the Burnside's Lemma. But for this simple case it might be easier to follow a more direct reasoning.

Proof of Lemma 3.40. Let T be an Otter tree with n leaves. Clearly, for $n = 1$ the equality in (3.32) is true. Assuming (3.32) holds for all Otter trees T' with less than n leaves, we conclude that (3.32) holds for T in the following way. T has subtrees T_1 and T_2 rooted at the root of T, because $n > 1$. If $k \geq 1$ is the number leaves in T_1, we have

$$w(T) = \frac{1}{2^{s_\rho}} \binom{n}{k} w(T_1) w(T_2) = \frac{n!}{2^{\mathrm{sym}(T_1)+\mathrm{sym}(T_1)+s_\rho}}$$

and $\mathrm{sym}(T_1) + \mathrm{sym}(T_2) + s_\rho = \mathrm{sym}(T)$, where $s_\rho = 1$ if $T_1 = T_2$ and $s_\rho = 0$ otherwise.

The second equality in (3.33) follows from Definition 3.1 and is illustrated in Figure 3.9. The third equality in (3.33) is Theorem 3.2. □

Definition 3.41. The BGF of all Otter trees counting their leaves and their symmetry vertices is denoted by

$$F(z, u) := \sum_{T \in \mathcal{O}} u^{\mathrm{sym}(T)} z^{|T|},$$

where $|T|$ denotes the number of leaves in T.

Lemma 3.42. *The BGF $F(z, u)$ is given implicitly by*

$$F(z, u) = z + \frac{1}{2} F(z, u)^2 + \left(u - \frac{1}{2}\right) F\left(z^2, u^2\right).$$

Proof. Let $T \in \mathcal{O}_n$. If $n = 1$, there is only one Otter tree T and $\mathrm{sym}(T) = 0$, hence

$$F(z,u) = z + \text{higher terms.}$$

If $n \geq 2$, T consists of a root vertex and the subtrees T_1 and T_2. As already used in the proof of the previous lemma, the number of symmetry vertices is then given by

$$\mathrm{sym}(T) = \begin{cases} \mathrm{sym}(T_1) + \mathrm{sym}(T_2) + 1, & \text{if } T_1 = T_2, \\ \mathrm{sym}(T_1) + \mathrm{sym}(T_2), & \text{otherwise.} \end{cases}$$

T is constructed by its two subtrees T_1 and T_2. However, the MSET$_2$-operator is not directly applicable[9] as it was possible for the OGF $O(z)$, but a similar result is obtained in the following way. Consider first the number of ordered pairs (T_1, T_2), where the two trees T_1 and T_2 have n leaves and k symmetry vertices in total, which is given by $\left[u^k z^n\right] F(z,u)^2$. We first count the number of Otter trees with different subtrees, so we have to subtract the number of ordered pairs (T_1, T_2) with $T_1 = T_2$. The number of ordered pairs (T_1, T_1) with n leaves and k symmetry vertices in total is given by $\left[u^k z^n\right] F\left(z^2, u^2\right)$. Hence,

$$\frac{1}{2}\Big(F(z,u)^2 - F\left(z^2, u^2\right)\Big) \tag{3.34}$$

is the BGF of Otter trees with two different subtrees $T_1 \neq T_2$, because the trees are unordered and therefore we get every tree twice when counting the ordered pairs (T_1, T_2), where T_1 and T_2 are different. It remains to count the Otter trees with identical subtrees. If T has the subtrees $T_1 = T_2$ and T_i has n leaves and k symmetry vertices, where $i = 1, 2$, then T has $2n$ leaves and $2k + 1$ symmetry vertices, i.e. the BGF to count such trees is given by

$$uF\left(z^2, u^2\right). \tag{3.35}$$

Summation of (3.34) and (3.35) yields the claimed result. □

$F(z,1)$ equals the OGF $O(z)$ for Otter trees already given in (3.31). If one sets $u = \frac{1}{2}$, one

[9] The parameter $\mathrm{sym}(.)$ is not inherited in the sense of [27, sect. III.3.2], because in general $\mathrm{sym}(T) = \mathrm{sym}(T_1) + \mathrm{sym}(T_2)$ does not hold as we have seen.

gets the EGF $B(z)$ as already stated in (3.2) on page 49

$$F\left(z,\frac{1}{2}\right)=\sum_{T\in\mathcal{O}}\frac{1}{2^{\mathrm{sym}(T)}}z^{|T|}=\sum_{n\geq 1}\sum_{T\in\mathcal{O}_n}\frac{1}{2^{\mathrm{sym}(T)}}z^n=\sum_{n\geq 1}\frac{b_n}{n!}z^n=B(z),$$

where (3.33) is used for the last equality. Surprisingly $F\left(z,\frac{1}{4}\right)$ can be used to determine the probability p_n.

Theorem 3.43. *For $n\geq 2$ the probability that two rooted binary phylogenetic trees selected uniformly from $\mathcal{B}_n$ are isomorphic, is given by*

$$p_n=\sum_{T\in\mathcal{O}_n}\frac{w(T)^2}{b_n^2}=\left(\frac{n!}{(2n-3)!!}\right)^2\cdot[z^n]F\left(z,\frac{1}{4}\right).$$

Proof. There are b_n^2 possibilities to choose a ordered pair $(\mathcal{T},\mathcal{T}')$ of trees $\mathcal{T},\mathcal{T}'\in\mathcal{B}_n$. In $w(T)^2$ of these possibilities both trees $\mathcal{T}$ and $\mathcal{T}'$ are of shape T, i.e. $\mathcal{T}=(T,\phi)$ and $\mathcal{T}'=(T,\phi')$, because $w(T)$ is the size of the isomorphism class corresponding to T (see Figure 3.9 and Lemma 3.40), hence, there are $w(T)^2$ pairs $(\mathcal{T},\mathcal{T}')$ with $\mathcal{T}$ and $\mathcal{T}'$ being elements of this isomorphism class. Therefore the probability p_n is given by

$$p_n=\sum_{T\in\mathcal{O}_n}\frac{w(T)^2}{b_n^2}=\left(\frac{n!}{(2n-3)!!}\right)^2\cdot\sum_{T\in\mathcal{O}_n}\frac{1}{4^{\mathrm{sym}(T)}},$$

where Lemma 3.40 is used for the second equality. And finally, by Definition 3.41 we have $\sum_{T\in\mathcal{O}_n}\frac{1}{4^{\mathrm{sym}(T)}}=[z^n]F\left(z,\frac{1}{4}\right)$, which completes the proof. □

Asymptotic results. In Theorem 3.43 the probability p_n was expressed in terms of the coefficients of z^n in the power series

$$G(z):=F\left(z,\frac{1}{4}\right).\tag{3.36}$$

Solving the quadratic equation in Lemma 3.42 yields with $F(0,0)=0$

$$F(z,u)=1-\sqrt{1-2z-(2u-1)F(z^2,u^2)}$$

and therefore

$$G(z)=1-\sqrt{1-2z-\frac{1}{2}F\left(z^2,\frac{1}{16}\right)}.\tag{3.37}$$

We are now going to determine the growth rate of the coefficients of $G(z)$ to establish asymptotic results for the sequence p_n in Theorem 3.47. For this purpose we need to determine the location, type and number of the dominant singularities (see Section 2.4 and [27, sect. IV.4]).

Lemma 3.44. *Let r be the radius of convergence of the power series expansion of $G(z)$ in (3.36) centered at the origin. r satisfies the following inequalities*

$$0.4 < r < 0.625.$$

Proof. To prove the lower bound $0.4 < r$ first note that the OGF of the Otter trees is a majorant series, i.e.

$$[z^n]G(z) = [z^n]F\left(z, \frac{1}{4}\right) < [z^n]F(z,1) = [z^n]O(z) = o_n.$$

$O(z)$ is known to be convergent for all $|z| < 0.40269\ldots =: r_o$ (according to [6] a proof can be found in [55]). Hence, also $G(z)$ converges in a disc of radius 0.4 and $0.4 < r$.

Now we want to prove the upper bound $r < 0.625$. From Theorem 3.43 we know, that for all $n \geq 2$

$$p_n = \sum_{T \in \mathcal{O}_n} \frac{w(T)^2}{b_n^2} = \frac{\sum_{T\in\mathcal{O}_n} w(T)^2}{\left(\sum_{T\in\mathcal{O}_n} w(T)\right)^2}. \tag{3.38}$$

Recall that $o_n = |\mathcal{O}_n|$ and denote the summands of $\sum_{T\in\mathcal{O}_n} w(T)$ by $a_1 + \cdots + a_{o_n}$. Now the Cauchy-Schwarz inequality (see e.g. [1, lem. 6.4.9, p. 246]) can be applied to $(a_1, \ldots, a_{o_n}) \in \mathbb{R}^{o_n}$ and $(1, \ldots, 1) \in \mathbb{R}^{o_n}$, which yields

$$\underbrace{(1 + \cdots + 1)}_{=o_n} \cdot \left(a_1^2 + \cdots + a_{o_n}^2\right) > (1 \cdot a_1 + \cdots + 1 \cdot a_{o_n})^2$$

and therefore with (3.38)

$$p_n = \frac{a_1^2 + \cdots + a_{o_n}^2}{(1 \cdot a_1 + \cdots + 1 \cdot a_{o_n})^2} > \frac{1}{o_n}. \tag{3.39}$$

As mentioned the radius of convergence r_o of the power series $O(z) = \sum_n o_n z^n$ is greater than 0.4, i.e. $r_o > 0.4$. The radius of convergence $r_{1/o}$ of the power series $\sum_n \frac{1}{o_n} z^n$ satisfies

$r_{1/o} < 2.5$, because

$$r_{1/o} = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{\frac{1}{o_n}}} = \liminf_{n\to\infty} \sqrt[n]{o_n} \leq \limsup_{n\to\infty} \sqrt[n]{o_n} = \frac{1}{r_0} < \frac{1}{0.4} = 2.5.$$

Hence, (3.39) implies that also the radius of convergence r_p of the power series $\sum_n p_n x^n$ satisfies $r_p < 2.5$. Furthermore with Theorem 3.43 and with

$$\frac{n!}{(2n-3)!!} = \frac{2^{n-2} \cdot (n-2)! \cdot n!}{(2n-3)!} \leq 2^n$$

we have

$$p_n = \left(\frac{n!}{(2n-3)!!}\right)^2 \cdot [z^n]F\left(z, \frac{1}{4}\right) \leq 4^n [z^n]F\left(z, \frac{1}{4}\right).$$

Therefore is $r < \frac{r_p}{4} < \frac{2.5}{4} = 0.625$. □

Lemma 3.45. *The dominant singularities of $G(z)$ are isolated and of square-root type.*

Proof. The radius of convergence r of the power series $G(z)$ satisfies $r < 1$ as shown in Lemma 3.44, and therefore also $r < \sqrt{r}$. This implies that $G'(z) := F\left(z^2, \frac{1}{16}\right)$ converges at least for all z with $|z| < \sqrt{r}$ because $G'(z)$ is a majorant series of $F\left(z, \frac{1}{4}\right)$, i.e.

$$[z^n]G'(z) < [z^n]F\left(z^2, \frac{1}{4}\right)$$

and $F\left(z^2, \frac{1}{4}\right)$ converges for all $|z| < \sqrt{r}$ since by definition the radius of convergence of $F\left(z, \frac{1}{4}\right) = G(z)$ is r. Thus, the function

$$G''(z) = 1 - 2z + \frac{1}{2}G'(z) = 1 - 2z + \frac{1}{2}F\left(z^2, \frac{1}{16}\right)$$

is analytic in the interior of a disc with radius $\sqrt{r}$, and considering the expression for $G(z)$ in (3.37) we conclude that all dominant singularities of $G(z)$ are zeros of $G''(z)$ with modulus r and therefore of square-root type. These zeros are isolated because $G''(z)$ is analytic in a disc with radius $\sqrt{r}$ and if there were not isolated zeros of $G''(z)$, analytic continuation would lead to $G''(z) \equiv 0$ for all z with $|z| < \sqrt{r}$. □

Lemma 3.46. *$r \in \mathbb{R}^+$ is a dominant singularity of $G(z)$ and it is the only dominant singularity.*

Pringsheim's Theorem (see e.g. [27, p. 240ff.]) implies that r is a dominant singularity of $G(z)$.

A proof of the fact that r is the only dominant singularity is given in [6]. Singularity analysis (see [27, chapt. VI]) and in particular the *O-transfer* (see [27, p. 390]) is used to prove that, as a power series, $G(z)$ converges for all z with $\rho = |z|$. Then using the triangle inequality one can conclude that there cannot be a second dominant singularity.

Now by the previous lemmata the conditions for singularity analysis (as summarized in [27, chapt. VI.4] and Section 2.4) are satisfied, and the following results can be established.

Theorem 3.47. *The probability p_n that two rooted binary phylogenetic trees with label set X and $n = |X|$ are isomorphic is asymptotically equivalent to*

$$p_n \sim a \cdot (4r)^{-n} \cdot n^{\frac{3}{2}} \left(1 + \sum_k \frac{c_k}{n^k}\right),$$

where $a = 3.17508\ldots$, $4r = 2.35967\ldots$ and the c_k are computable constants.

3.4 Other enumeration problems

In Section 3.1 we discussed the enumeration of X-trees, phylogenetic trees, and binary phylogenetic trees—all of them either rooted or unrooted. These are the types of trees commonly used (e.g. in the book *Phylogenetics* by Semple and Steel [63]). Nevertheless, there are several additional classes of trees which are related to phylogenetics.

In five articles Foulds and Robinson [29, 30, 31, 32, 33] published a complete list of results for the size of certain classes of phylogenetic trees. They considered twelve different types of trees, which are obtained by restricting vertex degree, labeling and number of labels per leaf. Each class either contains only trees with vertices of degree 3 or 1 (binary trees), or of degree 1 or greater than 2 (e. g. phylogenetic trees), or the degree is unrestricted (X-trees). Furthermore, either only leaves are labeled, or internal nodes are labeled too. And last, either multiple labels per vertex are allowed, or labeling is restricted to at most one label per vertex. In Section 3.1.4 the enumeration of X-trees was discussed following the way presented in [33]. The results for the other classes of trees, discussed by Foulds and Robinson [29, 30, 31, 32, 33], are obtained using similar methods. The enumeration of binary phylogenetic trees was discussed in Section 3.1.1 and can be found also in [30], the enumeration of phylogenetic trees (Section 3.1.2) is discussed also in [29], but we followed a slightly different approach in these two sections.

In [63, sect. 2.4] the number of so-called *tree shapes* is discussed, that is the number of unlabeled trees T. A tree shape is obtained by ignoring the labels of a phylogenetic tree. The number

of tree shapes can be interesting for problems concerning isomorphisms of phylogenetic trees as studied in Section 3.3. As already mentioned, in the case of rooted binary phylogenetic trees this is the class of unlabeled and unordered rooted binary trees, called *Otter trees*. The number of such trees was first studied by Otter [55]; Harding [42] discussed this class of trees particularly with regard to phylogenetics. The number of tree shapes of rooted multifurcating phylogenetic trees, i.e. the number of rooted unlabeled and unordered trees without vertices of degree two, is determined in [27, p. 479, VII.23]. Trees of this type are called *unlabeled hierarchies*. Tree shapes of X-trees are the rooted unordered and unlabeled trees. The number of these trees is discussed in [27, sect. I.5.2].

Rooted phylogenetic trees define implicitly a chronological order of the speciation events represented by the internal vertices. If a vertex u is on the path from v to the root, the speciation event represented by v happened after the speciation event represented by u. But this is only a partial order since two internal nodes u, v are unrelated with respect to this order if u and v are in different subtrees of some vertex. Therefore it makes sense to extend the concept of phylogenetic trees and to define a linear order on the set of internal vertices, which respects the given partial order. Felsenstein [21, p. 32] wrote in 1978

> “Of particular interest would be the number of different rooted trees (bifurcating or multifurcating) which are consistent with a set of fossil species ordered in time, plus a certain number of contemporary species.”

Such trees are called *ranked phylogenetic tree* in [63, sect. 2.3] or *labeled histories* in [23, p. 35f.]. In [63, sect. 2.3] and [23, p. 35f.] the number of different ranked phylogenetic trees for a fixed label set X with $n = |X|$ is given as

$$\frac{n!(n-1)!}{2^{n-1}}.$$

Similar to the first problem considered in Section 3.2, there are also other enumeration problems on a fixed phylogenetic tree. For example the number of convex characters $\chi : X \to C$ on a fixed binary phylogenetic tree $\mathcal{T}$ with label set X is given by

$$\frac{c!}{(c-r)!}\binom{2n-r-1}{r-1},$$

where $r = |\chi(X)|$, $c = |C|$ and $n = |X|$ (see [63, p. 68]). Surprisingly, this number does not depend on the shape of $\mathcal{T}$.

As already mentioned, neighborhoods play a role for hill-climbing algorithms. The number of

trees in such neighborhoods are studied in [2, 66].

These are only a few examples—there are many more enumeration problems concerning biomathematics. Apart from that, phylogenetics is an ongoing field of research, and new concepts are developed. For instance, Sanderson et al. [60] introduce so-called *terraces* as certain subsets of the tree space, which leads to new enumeration problems.

Chapter 4

Maximum parsimony on subtrees

In this Chapter we want to illuminate a question related much closer to biology than the problems considered in the previous chapter. Anyhow, the used mathematical methods are quite similar and based on the same models as introduced in Chapter 2. The selection of topics in this chapter can be considered as example how discrete mathematics is applied in the field of evolutionary biology.

According to Fischer and Thatte [24] simulations by Salisbury and Kim [59] and Zhang and Nei [85] suggest that the reconstruction accuracy of parent character states is increased when more species are considered. But Li et al. [49] and Fischer and Thatte [24] presented counter examples of the following form. The probability that the Fitch-Hartigan algorithm correctly reconstructs the character state of the root of a certain phylogenetic tree for a random character under the N_r-model is higher if the algorithm is applied only on a subset $Y \subseteq X$ of the label set and its induced subtree. We will discuss these examples in detail in Section 4.2. In Section 4.3 a positive result for the reconstruction accuracy of the Fitch-Hartigan algorithm under much stronger constraints will be presented: for ultrametric binary trees the reconstruction accuracy is at least better than guessing the character state of the root as the character state of some of the leaves of the tree.

In section 4.4 we present some considerations and initial results concerning the generalization of the statement from Section 4.2 for characters with more than two states. We tried to solve this question as part of the research work for this thesis, but it turned out to be more complicated than expected. Nevertheless, we can present some partial results, derived with help of a computer algebra system and sketch some ideas for future work.

4.1 Reconstruction accuracy of MP for a given tree

Li et al. [49, p. 648] introduced the *reconstruction accuracy* for a random character under the N_r-model with two states. We will generalize this definition for characters with $r = |C| \geq 2$ states.

We start with some naming conventions and other definitions to simplify the notation, which we will use throughout the whole chapter. $C = \{\alpha, \beta, \gamma, \ldots\}$ will always denote a set of r distinct character states and the state α will play a special role for the definition of reconstruction accuracy. Nevertheless, we will consider characters always under the N_r-model in this chapter, and due to the symmetry of the N_r-model we could have chosen any other element of C for this special role.

Definition 4.1. Let $\mathcal{T} = (T, \phi)$ be a rooted phylogenetic tree and $\{\xi_v | v \in V(\mathcal{T})\}$ a Markov process on $\mathcal{T}$ and with label set X and $Y \subseteq X$ a subset of labels. Furthermore, denote by y the root of the subtree T_Y of T induced by the leaves with labels in Y (see Figure 4.1) and χ a random character on $(T_Y, \phi|_Y)$ induced by the Markov process, i.e. $\chi(x) = \xi_{\phi(x)}$ for all $x \in Y$. Then $\mathrm{MP}(Y; \mathcal{T})$ denotes the set of character states reconstructed by the Fitch-Hartigan algorithm for the vertex y from the character χ on $(T_Y, \phi|_Y)$. With the notation of Theorem 2.9 this is

$$\psi(y) =: \mathrm{MP}(Y; \mathcal{T}).$$

$\mathrm{MP}(Y; \mathcal{T})$ is a random variable, which depends on the random variables of the Markov process. We are mostly interested in the probability event $\{\mathrm{MP}(Y; \mathcal{T}) = A\}$ for some nonempty set $A \subseteq C$ of character states.

For $\mathrm{MP}(Y; \mathcal{T})$ we write shorter $\mathrm{MP}(Y)$ if it cannot be mistaken. Furthermore, we define for the fixed state $\alpha \in C$

$$P_A^{\mathcal{T}}(Y) := \mathbb{P}(\mathrm{MP}(Y; \mathcal{T}) = A | \xi_y = \alpha)$$

and

$$R_A^{\mathcal{T}}(Y) := \mathbb{P}(\mathrm{MP}(Y; \mathcal{T}) = A | \xi_\rho = \alpha)$$

and write occasionally shorter $P_A(Y)$ instead of $P_A^{\mathcal{T}}(Y)$ and $R_A(Y)$ instead of $R_A^{\mathcal{T}}(Y)$. For states $\alpha, \beta, \ldots$, we will write $P_{\alpha\beta\ldots}(Y)$ instead of $P_{\{\alpha,\beta,\ldots\}}(Y)$. Note that y depends on $Y \subseteq X$ and $P_A^{\mathcal{T}}(Y)$ depends only on the structure of T_Y and the substitution probabilities of edges $e \in E(T_Y)$. In addition, $P_A^{\mathcal{T}}(X) = R_A^{\mathcal{T}}(X)$, because ρ is the root of $T_X = T$ and therefore $y = \rho$.

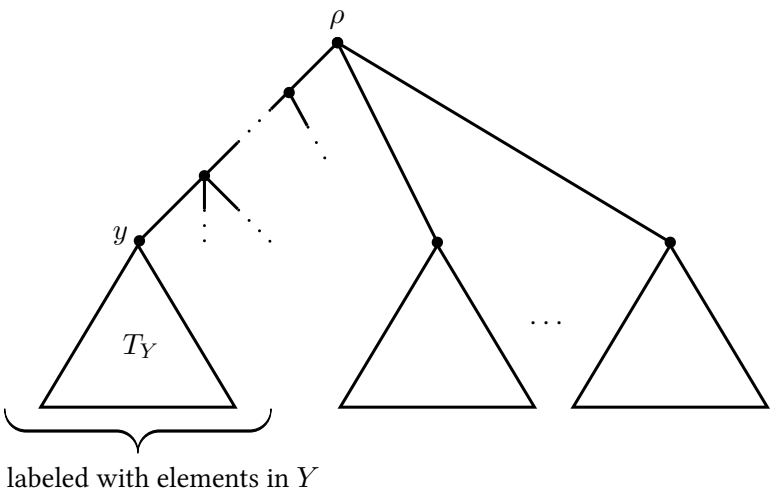


Figure 4.1: The subtree T_Y induced by a subset $Y \subseteq X$ as in Definition 4.1.

Throughout the whole chapter we will consider only binary phylogenetic trees (although the definitions make sense also for not binary phylogenetic trees). Therefore the simplification for the Fitch-Hartigan algorithm explained in Remark 2.10 on page 27 can be applied.

Definition 4.2. The *reconstruction accuracy* for a rooted phylogenetic tree $\mathcal{T} = (T, \phi)$ under the N_r-model with label set X and $Y \subseteq X$ is defined by

$$\begin{aligned} RA(Y;\mathcal{T}) &:= \sum_{A\subseteq C, A\ni\alpha} \frac{1}{|A|} \cdot \mathbb{P}(\mathrm{MP}(Y;\mathcal{T})) = A|\xi_\rho = \alpha) \\ &= \sum_{A\subseteq C, A\ni\alpha} \frac{1}{|A|} \cdot R_A^{\mathcal{T}}(Y), \end{aligned}$$

where the summation is over all subsets of character states A with $\alpha \in A$ and $A \subseteq C$.

Note that $\mathrm{RA}(Y;\mathcal{T})$ is independent of the selection of the state $\alpha \in C$ due to the symmetry of the N_r-model. Moreover, by the law of total probability it becomes clear that $RA(Y;\mathcal{T})$ is the probability of the event $\{\xi_\rho \in \mathrm{MP}(Y;\mathcal{T})\} \cap B$, where B is the event that the correct state

ξ_ρ is selected from the set $\mathrm{MP}(Y;\mathcal{T})$ with uniform distribution. In detail this yields

$$\begin{aligned} & \frac{1}{|\mathrm{MP}(Y;\mathcal{T})|} \cdot \mathbb{P}(\xi_\rho \in \mathrm{MP}(Y;\mathcal{T})) \quad = \\ = \sum_{c\in C} \underbrace{\mathbb{P}(\xi_\rho = c)}_{=\frac{1}{r}} \cdot & \sum_{A\subseteq C, A\ni c} \frac{1}{|A|} \cdot \mathbb{P}(\mathrm{MP}(Y;\mathcal{T}) = A | \xi_\rho = c) \quad = \\ & \sum_{A\subseteq C, A\ni \alpha} \frac{1}{|A|} \cdot R_A^{\mathcal{T}}(Y) \quad = \quad RA(Y;\mathcal{T}). \end{aligned}$$

(We write "$A \subseteq C, A \ni c$" to indicate a summation over all sets A with $c \in A$ and $A \subseteq C$ for a fixed c and "$c \in C$" to indicate a summation over all states c.)

4.2 Misleading information

Li et al. [49] quote Crisp and Cook [13, p. 127] to emphasize that intuitively one would expect reconstruction to be easier if there are more species and therefore also more character data available:

> "If ancestral features are to be inferred from a phylogeny, a method that optimizes character states over the whole tree should be used."

However, it is possible to construct trees where this is not true as we will show in this section. At first glance it might be counterintuitive that leaving out information improves the results of reconstruction. On the other hand, it is not surprising that it may help if a misleading species $z \in X$ is excluded, e.g. if z is very far away from ρ compared with the species in $X \setminus \{z\}$. This means that the substitution probability from ρ to z is very high and therefore it is very unlikely that one can guess the state of ρ correctly by choosing the state of z. We will prove that such a misleading species $z \in X$ can be even arbitrarily close to ρ, that is $\mathbb{P}(\xi_\rho \neq \xi_z)$ can be arbitrarily small. Also Li et al. [49] gave an example where $\mathrm{RA}(Y) > \mathrm{RA}(X)$ with $Y \subseteq X$, but without the property that all species in $X \setminus Y$ are close to the root. The following theorem and its proof is due to Fischer and Thatte [24].

Theorem 4.3. *Let X be a set of labels. For any p_z with $0 < p_z < \frac{1}{2}$ there exists a rooted binary phylogenetic tree $\mathcal{T}$ under the N_r-model with character state set $C = \{\alpha, \beta\}$ and label set X, z a leaf of $\mathcal{T}$ and $\mathbb{P}(\xi_\rho \neq \xi_z) = p_z$, such that z is closer to the root ρ than any other leaf (see Figure 4.2), i.e.*

$$p_z < \mathbb{P}(\xi_\rho \neq \xi_v) \tag{4.1}$$

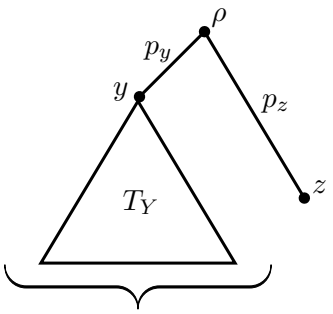


Figure 4.2: Illustration of a tree with $\mathrm{RA}\big(X \setminus \{\phi^{-1}\{z\}\}\big) > \mathrm{RA}(X)$ as in Theorem 4.3.

for any leaf $v \in V(\mathcal{T})$ *with* $v \neq z$ *and*

$$\mathrm{RA}\big(X \setminus \{\phi^{-1}(z)\}\big) > \mathrm{RA}(X). \tag{4.2}$$

To prove the theorem Fischer and Thatte [24] use the following lemma (further discussion of the convergence of these probabilities can be found in [70], a proof for the lemma was given by Steel [72]).

Lemma 4.4. *Let* $q \in \mathbb{R}$ *with* $0 < q < \frac{1}{8}$*. For all* $n \geq 2$ *let* $\mathcal{T}_n$ *be a balanced*[1] *rooted binary tree under the* N_r*-model with character state set* $C = \{\alpha, \beta\}$ *and label set* X_n*, where* $\mathcal{T}_n$ *is of depth* n *and the substitution probability is* q *for each edge* $e \in E(\mathcal{T}_n)$*. Then* $P_\alpha^{\mathcal{T}_n}(X_n)$ *approaches*

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

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

Proof sketch of Theorem 4.3. We choose $\mathcal{T}$ to be the rooted phylogenetic binary tree with a single vertex z as first subtree of ρ and the tree T_Y as second subtree as illustrated in Figure 4.2. Furthermore, we choose T_Y to be a balanced tree of height n for some $n \geq 2$ as in Lemma 4.4, i.e. $(T_Y, \phi|_Y) := \mathcal{T}_n$ and therefore $Y := X_n$ and substitution probability q for all edges $e \in E(T_Y)$. By $p_y := \mathbb{P}(\xi_\rho \neq \xi_y)$ we denote the substitution probability from ρ to y.

Now $\mathrm{RA}(X;\mathcal{T})$ and $\mathrm{RA}(Y;\mathcal{T})$ can be expressed in terms of p_z, p_y, $P_\alpha^{\mathcal{T}}(Y)$, $P_\beta^{\mathcal{T}}(Y)$ and $P_{\alpha\beta}^{\mathcal{T}}(Y)$ (a similar task will be done detailed in the proof of Theorem 4.5 and in Remark 4.9). With

[1] A *balanced tree* of depth n is a rooted tree with exactly n edges on the path from the root to each leaf (see [24, p. 291]).

the resulting expressions, we can express the condition in (4.2), i.e. $\mathrm{RA}(Y;\mathcal{T}) > \mathrm{RA}(X;\mathcal{T})$, as

$$(p_z - p_y)P_\alpha^{\mathcal{T}}(Y) > (1-2p_z)P_{\alpha\beta}^{\mathcal{T}}(Y) + (1-p_z-p_y)P_\beta^{\mathcal{T}}(Y). \tag{4.3}$$

With Lemma 4.4 we can conclude that the left side of the inequality approaches $p_z - p_y$ and the right side 0 as $n \to \infty$ and $q \to 0$. Therefore, if we choose p_y to satisfy $p_y < p_z$, then there exist a $\frac{1}{8} > q > 0$ and a $N \in \mathbb{N}$ such that (4.3) is fulfilled for all $n \geq N$. Furthermore (4.1) can be expressed with help of Lemma 2.15 (details omitted)

$$(1-2q)^n < \frac{1-2p_z}{1-2p_y}. \tag{4.4}$$

The left side and the right side of (4.4) are both between 0 and 1. Now choose n sufficiently large such that (4.4) holds and additionally $n \geq N$. Then all claimed properties of $\mathcal{T}$ are fulfilled. □

4.3 A lower bound for the reconstruction accuracy

Despite the undesired and surprising results of the previous section, under much stronger conditions we can prove some positive results. As already conjectured by Li et al. [49], with some additional conditions the reconstruction accuracy is at least better than the conservation probability from the root to any leaf. This corresponds to the case where the subset $Y \subseteq X$ in Definition 4.2 consists only of a single leaf $Y = \{x\}$. This result shows also that the height of the tree provides a lower bound for the reconstruction accuracy of the tree. The conjecture was formally proven in [24] for characters with two states. In this section we present this proof, in the next section the generalized statement for characters with two or more states will be discussed.

Theorem 4.5. *Let $\mathcal{T}$ be an ultrametric rooted binary phylogenetic tree under the N_r-model with label set X and $\{\xi_v | v \in V(\mathcal{T})\}$ the according Markov process on $\mathcal{T}$ with state set $C = \{\alpha, \beta\}$ and p the height of $\mathcal{T}$. Then the reconstruction accuracy is not less than the conservation probability $1 - p = \mathbb{P}(\xi_\rho = \xi_v)$ from the root to any leaf v*

$$\mathrm{RA}(X;\mathcal{T}) \geq 1 - p.$$

Proof. Let $\mathcal{T} = (T, \phi)$ be a rooted binary phylogenetic tree as in the conditions of the theorem. In addition, denote the two child nodes of ρ by y_1 and y_2 respectively, the subtrees of T rooted at

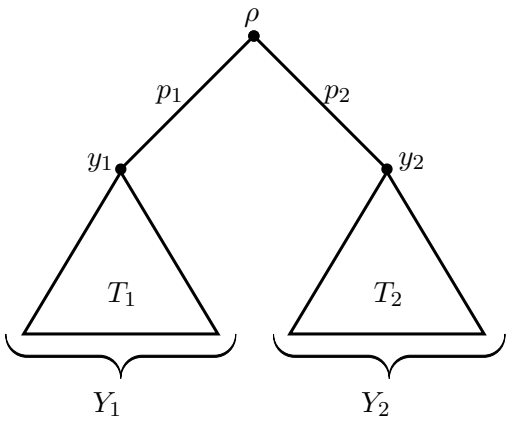


Figure 4.3: A rooted binary ultrametric phylogenetic tree illustrates the notation used in the proof of Theorem 4.5.

y_i by T_i and their label sets by $Y_i \subseteq X$ for $i = 1, 2$ (see Figure 4.3). The substitution probability from ρ to y_i shall be denoted by p_i for $i = 1, 2$.

We will express $\mathrm{RA}(X; \mathcal{T})$ recursively by means of p_1 and p_2 and probabilities of the random variables of the nodes in T_1 and T_2 and then prove the claimed statement by induction. First note that

$$\mathrm{RA}(X; \mathcal{T}) = P_\alpha(X) + \frac{1}{2} P_{\alpha\beta}(X).$$

We will express both $P_\alpha(X)$ and $P_{\alpha\beta}(X)$ separately by means of $P_\alpha(Y_i)$, $P_\beta(Y_i)$, $P_{\alpha\beta}(Y_i)$ and p_i for $i = 1, 2$. Inspecting Step (ii) in the Fitch-Hartigan algorithm presented in Theorem 2.9 (see also Remark 2.10) makes clear, that we have $\mathrm{MP}(X; \mathcal{T}) = \{\alpha\}$ if and only if one of the following three cases occurs:

$$\mathrm{MP}(Y_1; \mathcal{T}) = \{\alpha\} \text{ and } \mathrm{MP}(Y_2; \mathcal{T}) = \{\alpha\}$$

or

$$\mathrm{MP}(Y_1; \mathcal{T}) = \{\alpha, \beta\} \text{ and } \mathrm{MP}(Y_2; \mathcal{T}) = \{\alpha\}$$

or

$$\mathrm{MP}(Y_1; \mathcal{T}) = \{\alpha\} \text{ and } \mathrm{MP}(Y_2; \mathcal{T}) = \{\alpha, \beta\}.$$

Therefore the event $\{\mathrm{MP}(X; \mathcal{T}) = \{\alpha\}\}$ can be expressed by a disjoint union of intersections of events of the form $\{\mathrm{MP}(Y_i; \mathcal{T}) = \{\alpha\}\}$ and $\{\mathrm{MP}(Y_i; \mathcal{T}) = \{\alpha, \beta\}\}$ and its probability by a

sum of a product of the corresponding probabilities. These considerations lead to

$$\begin{aligned} P_\alpha(X) &= R_\alpha(Y_1) \cdot R_\alpha(Y_2) \\ &+ R_{\alpha\beta}(Y_1) \cdot R_\alpha(Y_2) \\ &+ R_\alpha(Y_1) \cdot R_{\alpha\beta}(Y_2). \end{aligned}$$

By the law of total probability $R_\alpha(Y_i)$ can be expressed as sum

$$\begin{aligned} R_\alpha(Y_i) &= \mathbb{P}(\xi_\rho = \xi_{y_i})\mathbb{P}(\mathrm{MP}(Y_i;\mathcal{T}) = \{\alpha\}|\xi_\rho = \alpha, \xi_\rho = \xi_{y_i}) \\ &+ \mathbb{P}(\xi_\rho \neq \xi_{y_i})\mathbb{P}(\mathrm{MP}(Y_i;\mathcal{T}) = \{\alpha\}|\xi_\rho = \alpha, \xi_\rho \neq \xi_{y_i}) \\ &= (1-p_i) \cdot P_\alpha(Y_i) + p_i \cdot P_\beta(Y_i). \end{aligned}$$

For the second summand we used the symmetry of the N_r-model to derive the equality

$$\mathbb{P}(\mathrm{MP}(Y_i;\mathcal{T}) = \{\alpha\}|\xi_{y_i} = \beta) = \mathbb{P}(\mathrm{MP}(Y_i;\mathcal{T}) = \{\beta\}|\xi_{y_i} = \alpha) = P_\beta(Y_i).$$

In an analogous way we find the expressions $R_\beta(Y_i) = (1-p_i) \cdot P_\beta(Y_i) + p_i \cdot P_\alpha(Y_i)$ and $R_{\alpha\beta}(Y_i) = P_{\alpha\beta}(Y_i)$. In total this yields

$$\begin{aligned} P_\alpha(X) &= ((1-p_1) \cdot P_\alpha(Y_1) + p_1 \cdot P_\beta(Y_1)) \cdot ((1-p_2) \cdot P_\alpha(Y_2) + p_2 \cdot P_\beta(Y_2)) \\ &+ P_{\alpha\beta}(Y_1) \cdot ((1-p_2) \cdot P_\alpha(Y_2) + p_2 \cdot P_\beta(Y_2)) \\ &+ ((1-p_1) \cdot P_\alpha(Y_1) + p_1 \cdot P_\beta(Y_1)) \cdot P_{\alpha\beta}(Y_2). \end{aligned}$$

An analogous expression for $P_{\alpha\beta}(X)$ can be determined in the same way. Now the state sets reconstructed for y_1 and y_2 equal either both $\{\alpha, \beta\}$ or one is $\{\alpha\}$ and the other one $\{\beta\}$. In total we get

$$\begin{aligned} P_{\alpha\beta}(X) &= P_{\alpha\beta}(Y_1) \cdot P_{\alpha\beta}(Y_2) \\ &+ ((1-p_1) \cdot P_\alpha(Y_1) + p_1 \cdot P_\beta(Y_1)) \cdot ((1-p_2) \cdot P_\beta(Y_2) + p_2 \cdot P_\alpha(Y_2)) \\ &+ ((1-p_1) \cdot P_\beta(Y_1) + p_1 \cdot P_\alpha(Y_1)) \cdot ((1-p_2) \cdot P_\alpha(Y_2) + p_2 \cdot P_\beta(Y_2)). \end{aligned}$$

In order to simplify the expression $\mathrm{RA}(X;\mathcal{T}) = P_\alpha(X) + \frac{1}{2}P_{\alpha\beta}(X)$ a computer algebra system is helpful. We define as abbreviation $D(Y) := \mathrm{RA}(Y;\mathcal{T}_Y) - \mathbb{P}(\xi_v \neq \xi_y)$, where $\mathcal{T}_Y$ is the subtree of $\mathcal{T}$ induced by $Y \subseteq X$, y is its root, and v a leaf of $\mathcal{T}_Y$. Furthermore, we apply the

transformation $P_i := 1 - 2p_i$ and $P := 1 - 2p$, which simplifies the expressions (we used the same transformation previously in Lemma 2.15). It remains to show $D(X) \geq 0$.

With two further equations for each $i = 1, 2$ it is possible to simplify the expression $D(X)$. Clearly $P_\alpha(Y_i)$, $P_\beta(Y_i)$ and $P_{\alpha\beta}(Y_i)$ sum up to 1 because the probability events form a partition of the set of all possible outcomes, hence we have $P_\beta(Y_i) = 1 - P_\alpha(Y_i) - P_{\alpha\beta}(Y_i)$ and it is possible to eliminate the unknown $P_\beta(Y_i)$. On the other hand, with Lemma 2.15 the substitution probability from y_i to a leaf $v \in V(T_i)$ can be obtained

$$\mathbb{P}(\xi_{y_i} \neq \xi_v) = \frac{1}{2}\left(1 - \frac{P}{P_i}\right).$$

Therefore the conservation probability in $D(Y_i)$ can be expressed explicitly and the unknown $P_\alpha(Y_i)$ can be eliminated using the equation

$$P_\alpha(Y_i) = D(Y_i) - \frac{1}{2}P_{\alpha\beta}(Y_i) + \frac{1}{2}\left(1 - \frac{P}{P_i}\right).$$

The Mathematica notebook in Section A.3 consists of the considerations made so far and a final `FullSimplify` command yields

$$\begin{aligned} 4D(X) &= 2 \cdot D(Y_2)P_2(1 + P_{\alpha\beta}(Y_1)) + 2 \cdot D(Y_1)P_1(1 + P_{\alpha\beta}(Y_2)) \\ &+ P \cdot P_{\alpha\beta}(Y_1) + P \cdot P_{\alpha\beta}(Y_2). \end{aligned} \tag{4.5}$$

If T consists only of one single vertex ρ, clearly $P_\alpha(X) = 1$, $P_{\alpha\beta}(X) = 0$ and $\mathbb{P}(\xi_v \neq \xi_\rho) = 0$ for a leaf $v \in V(\mathcal{T})$ and therefore $D(X) \geq 0$ in this case. Now assume by induction that $D(Y_i) \geq 0$ for $i = 1, 2$. Since all summands in (4.5) are nonnegative, we conclude $D(X) \geq 0$. □

4.4 Characters with more than two states

Especially the cases with character state sets C, where $r = |C| = 4$ or $r = |C| = 20$, are applicable in evolutionary biology because there are four nucleobases in the DNA and twenty amino acids occur in nature (see e.g. [28, p. 43f.]). In view of this fact it would be very interesting to generalize Theorem 4.5 to have an analogous statement also for characters allowing more than two states. This is formulated in the following conjecture (originally due to Li et al. [49], and partially proven by Fischer and Thatte [24] as described in the previous section).

Conjecture 4.6. *Let $\mathcal{T}$ be an ultrametric rooted binary phylogenetic tree under the N_r-model with label set X and $\{\xi_v | v \in V(\mathcal{T})\}$ the according Markov process on $\mathcal{T}$ with state set C and p the height of $\mathcal{T}$. Also if $|C| = r > 2$, the reconstruction accuracy is not less than the conservation probability $1 - p = \mathbb{P}(\xi_\rho = \xi_v)$ from the root to a leaf v*

$$\mathrm{RA}(X; \mathcal{T}) \geq 1 - p.$$

In this section we show that a mere transfer of the proof idea of Theorem 4.5 to the case $r > 2$ fails to do its job. Furthermore we will present alternative approaches, which might help solving the problem.

In the following paragraphs we will analyze the proof of Theorem 4.5 and use its notation to denote the occurring expressions.

A possible dead end street? It suggests itself, to generalize the proof of Theorem 4.5 directly for some fixed $r > 2$. This would not solve the question for all $r \geq 2$, but still could solve the problem for interesting special cases, such as $r = 4$ or $r = 20$, by using the same arguments as in the proof of the binary case. Unfortunately, this does not work out as expected. In the following we describe how to generalize the proof idea of Theorem 4.5 and why we do not consider this approach as a promising one.

The short answer is, already in the case $r = 3$ and $r = 4$ not all summands in the recursion formula for $D(X)$ are nonnegative. So we cannot easily conclude $D(X) \geq 0$. Moreover, we have actually found values for the unknowns such that $D(X)$ is negative. But neither does this disprove the conjecture nor gives it hope that a counter example can be constructed easily.

Definition 4.7. Let $C \neq \emptyset$ be an arbitrary finite set. We say that the arbitrarily chosen real numbers $P_A(Y_i)$, P and P_i for all nonempty $A \subseteq C$ and $i = 1, 2$, *occur as probabilities* (or shortly *occur*), if there exists a rooted binary ultrametric phylogenetic tree $\mathcal{T}$ under the N_r-model, such that $P_A(Y_i)$, P and P_i are the probabilities of $\mathcal{T}$ as defined in the proof of Theorem 4.5.

Example 4.8. Let $C = \{\alpha, \beta\}$, $P_\alpha(Y_1) := 42$ and all other values arbitrarily $P_A(Y_i) \in \mathbb{R}$, $P \in \mathbb{R}$, $P_i \in \mathbb{R}$. Also without knowledge of the other values, obviously the probabilities $P_A(Y_i)$, P and P_i do not occur, because $P_\alpha(Y_1) = 42$ cannot be a probability.

Note that P, P_1 and P_2 are actually not probabilities, because the transformation $P := 1 - \frac{r}{r-1}p$ and $P_i := 1 - \frac{r}{r-1}p_i$ was applied (see Lemma 2.15), but still $0 < P, P_i < 1$. This example is a simple illustration, but in general it is not easy to say, if some values occur as probabilities

or not. In the following we will demonstrate, why this is important for us. With $r = 2$ it was possible to prove

$$P_A(Y_i) \geq 0 \quad \forall A, P \geq 0, P_i \geq 0 \Rightarrow D(X) \geq 0. \tag{4.6}$$

But if $r = 3$ or $r = 4$ it turns out that

$$1 \geq P_A(Y_i) \geq 0 \quad \forall A, P \geq 0, P_i \geq 0 \not\Rightarrow D(X) \geq 0. \tag{4.7}$$

We found values for $P_A(Y_i)$, P and P_i, such that $D(X) < 0$. But this does not mean, that these values occur as probabilities. If the implication in (4.6) does not hold for values which occur, we can conclude that Conjecture 4.6 is false. Otherwise if it holds for all values which occur, the conjecture is true.

In the following we will explain in detail, why the implication in (4.6) does not hold for the case $r = 3$ and $r = 4$. We start with a general description how $D(X)$ can be expressed for an arbitrary $r \geq 2$ in terms of the probabilities $P_A(Y_i)$ of the subtrees of ρ and the probabilities p_i for $i = 1, 2$, following the same way as in the proof of Theorem 4.5. We will determine $D(X)$ for $r = 3$ and $r = 4$ and then explain (4.7) in detail. However, for $r = 3$ and $r = 4$ it would be a challenging task to compute $D(X)$ by hand—after a `FullSimplify` command in Mathematica and some further simplifications, $D(X)$ is a sum with 107 summands for $r = 3$ (see Figure 4.4) and 143 summands for $r = 4$. Therefore, the procedure in Remark 4.9 is implemented in Mathematica in order to compute $D(X)$ automatically for any given $r \geq 2$ (the source code can be found in Section A.4).

Remark 4.9 (A cooking recipe for the expression $D(X)$ for $r \geq 2$). Let $\mathcal{T}$, Y_1, Y_2, y_1, y_2, p_1, p_2 as in the proof of Theorem 4.5 (see Figure 4.3), but the number of character states $r \geq 2$. First, note that for any set of character states $\emptyset \neq A \subseteq C$ by Remark 2.10 we have the equivalence

$$\mathrm{MP}(X) = A \Leftrightarrow \begin{cases} \mathrm{MP}(Y_1) \cap \mathrm{MP}(Y_2) = A, & \text{if } \mathrm{MP}(Y_1) \cap \mathrm{MP}(Y_2) \neq \emptyset \\ \mathrm{MP}(Y_1) \dot{\cup} \mathrm{MP}(Y_2) = A, & \text{otherwise.} \end{cases}$$

Recall that this equivalence is only valid for a binary phylogenetic tree, but for any number of character states $r \geq 2$. So, we can express the probability event $\{\mathrm{MP}(X) = A\}$ by the disjoint union of events $\{\mathrm{MP}(Y_1) = A_1\} \cap \{\mathrm{MP}(Y_2) = A_2\}$ for all $A_1, A_2 \subseteq C$ with $A_1 \cap A_2 = A$ or $A_1 \dot{\cup} A_2 = A$. Because of the Markov property $\{\mathrm{MP}(Y_1) = A_1\}$ and $\{\mathrm{MP}(Y_2) = A_2\}$ are

independent[2]. Therefore in total we get

$$\begin{aligned} P_A^{\mathcal{T}}(X) &= \sum_{A_1 \cap A_2 = A} \mathbb{P}(\mathrm{MP}(Y_1) = A_1 | \xi_\rho = \alpha) \cdot \mathbb{P}(\mathrm{MP}(Y_2) = A_2 | \xi_\rho = \alpha) \\ &+ \sum_{A_1 \dot\cup A_2 = A} \underbrace{\mathbb{P}(\mathrm{MP}(Y_1) = A_1 | \xi_\rho = \alpha)}_{= R_{A_1}^{\mathcal{T}}(Y_1)} \cdot \underbrace{\mathbb{P}(\mathrm{MP}(Y_2) = A_2 | \xi_\rho = \alpha)}_{= R_{A_2}^{\mathcal{T}}(Y_2)}. \end{aligned}$$

We proceed with the decomposition of the summands. For $i = 1, 2$ by the law of total probability one gets

$$\begin{aligned} R_{A_i}^{\mathcal{T}}(Y_i) &= \mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_\rho = \alpha) = \\ &= \sum_{c \in C} \mathbb{P}(\xi_{y_i} = c | \xi_\rho = \alpha) \cdot \mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_\rho = \alpha, \xi_{y_i} = c) \\ &= \underbrace{\mathbb{P}(\xi_{y_i} = \alpha | \xi_\rho = \alpha)}_{= 1 - p_i} \cdot \underbrace{\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_\rho = \alpha, \xi_{y_i} = \alpha)}_{= P_{A_i}^T(Y_i)} \\ &+ \sum_{c \in C \setminus \{\alpha\}} \underbrace{\mathbb{P}(\xi_{y_i} = c | \xi_\rho = \alpha)}_{= \frac{p_i}{r-1}} \cdot \mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_\rho = \alpha, \xi_{y_i} = c). \end{aligned}$$

Using the symmetry of the N_r-model, for $c \in C \setminus \{\alpha\}$ the probability

$$\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_\rho = \alpha, \xi_{y_i} = c)$$

can be expressed my means of $P_{A'}^{\mathcal{T}}(Y_i)$ for an appropriate set $A' \subseteq C$. We distinguish between four cases:

Case 1. If $\alpha \in A_i$ and $c \in A_i$, we can simply exchange the roles of α and c using the symmetry of the N_r-model, which yields

$$\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = c) = \mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = \alpha) = P_{A_i}(Y_i).$$

Case 2. If $\alpha \notin A_i$ and $c \notin A_i$, again as previously the roles of α and c can be exchanged

$$\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = c) = \mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = \alpha) = P_{A_i}(Y_i).$$

[2]To be precise, we need independence with respect to the probability measure $\mathbb{P}_{\xi_\rho = \alpha}(B) := \mathbb{P}(B | \xi_\rho = \alpha)$.

$$
\begin{aligned}
81P_1P_2D(X) = &-8P^2P_1 + 2PP_1^2 - 12D(Y_1)PP_1^2 - 8P^2P_2 + 22PP_1P_2 - 12D(Y_1)PP_1P_2 - 12D(Y_2)PP_1P_2 + 28P^2P_1P_2 - 5P_1^2P_2\\
&+ 57D(Y_1)P_1^2P_2 + 3D(Y_2)P_1^2P_2 - 18D(Y_1)D(Y_2)P_1^2P_2 - 16PP_1^2P_2 + 42D(Y_1)PP_1^2P_2 + 2PP_2^2 - 12D(Y_2)PP_2^2\\
&- 5P_1P_2^2 + 3D(Y_1)P_1P_2^2 + 57D(Y_2)P_1P_2^2 - 18D(Y_1)D(Y_2)P_1P_2^2 - 16PP_1P_2^2 + 42D(Y_2)PP_1P_2^2 + 4P_1^2P_2^2\\
&- 24D(Y_1)P_1^2P_2^2 - 24D(Y_2)P_1^2P_2^2 + 63D(Y_1)D(Y_2)P_1^2P_2^2 - 6PP_1^2P_{\alpha\beta}(Y_1) + 12PP_1P_2P_{\alpha\beta}(Y_1) + 15P_1^2P_2P_{\alpha\beta}(Y_1)\\
&- 9D(Y_2)P_1^2P_2P_{\alpha\beta}(Y_1) + 48PP_1^2P_2P_{\alpha\beta}(Y_1) - 3P_1P_2^2P_{\alpha\beta}(Y_1) + 18D(Y_2)P_1P_2^2P_{\alpha\beta}(Y_1) - 12P_1^2P_2^2P_{\alpha\beta}(Y_1)\\
&+ 72D(Y_2)P_1^2P_2^2P_{\alpha\beta}(Y_1) - 2PP_1^2P_{\alpha\beta\gamma}(Y_1) + 16PP_1P_2P_{\alpha\beta\gamma}(Y_1) + 5P_1^2P_2P_{\alpha\beta\gamma}(Y_1) - 3D(Y_2)P_1^2P_2P_{\alpha\beta\gamma}(Y_1)\\
&+ 16PP_1^2P_2P_{\alpha\beta\gamma}(Y_1) + 5P_1P_2^2P_{\alpha\beta\gamma}(Y_1) + 24D(Y_2)P_1P_2^2P_{\alpha\beta\gamma}(Y_1) - 4P_1^2P_2^2P_{\alpha\beta\gamma}(Y_1) + 24D(Y_2)P_1^2P_2^2P_{\alpha\beta\gamma}(Y_1)\\
&+ 12PP_1P_2P_{\alpha\beta}(Y_2) - 3P_1^2P_2P_{\alpha\beta}(Y_2) + 18D(Y_1)P_1^2P_2P_{\alpha\beta}(Y_2) - 6PP_2^2P_{\alpha\beta}(Y_2) + 15P_1P_2^2P_{\alpha\beta}(Y_2) - 9D(Y_1)P_1P_2^2P_{\alpha\beta}(Y_2)\\
&+ 48PP_1P_2^2P_{\alpha\beta}(Y_2) - 12P_1^2P_2^2P_{\alpha\beta}(Y_2) + 72D(Y_1)P_1^2P_2^2P_{\alpha\beta}(Y_2) + 9P_1^2P_2P_{\alpha\beta}(Y_1)P_{\alpha\beta}(Y_2) + 9P_1P_2^2P_{\alpha\beta}(Y_1)P_{\alpha\beta}(Y_2)\\
&+ 36P_1^2P_2^2P_{\alpha\beta}(Y_1)P_{\alpha\beta}(Y_2) + 3P_1^2P_2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta}(Y_2) - 15P_1P_2^2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta}(Y_2) + 12P_1^2P_2^2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta}(Y_2)\\
&+ 16PP_1P_2P_{\alpha\beta\gamma}(Y_2) + 5P_1^2P_2P_{\alpha\beta\gamma}(Y_2) + 24D(Y_1)P_1^2P_2P_{\alpha\beta\gamma}(Y_2) - 2PP_2^2P_{\alpha\beta\gamma}(Y_2) + 5P_1P_2^2P_{\alpha\beta\gamma}(Y_2)\\
&- 3D(Y_1)P_1P_2^2P_{\alpha\beta\gamma}(Y_2) + 16PP_1P_2^2P_{\alpha\beta\gamma}(Y_2) - 4P_1^2P_2^2P_{\alpha\beta\gamma}(Y_2) + 24D(Y_1)P_1^2P_2^2P_{\alpha\beta\gamma}(Y_2) - 15P_1^2P_2P_{\alpha\beta}(Y_1)P_{\alpha\beta\gamma}(Y_2)\\
&+ 3P_1P_2^2P_{\alpha\beta}(Y_1)P_{\alpha\beta\gamma}(Y_2) + 12P_1^2P_2^2P_{\alpha\beta}(Y_1)P_{\alpha\beta\gamma}(Y_2) - 5P_1^2P_2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta\gamma}(Y_2) - 5P_1P_2^2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta\gamma}(Y_2)\\
&+ 4P_1^2P_2^2P_{\alpha\beta\gamma}(Y_1)P_{\alpha\beta\gamma}(Y_2) - 6PP_1^2P_\beta(Y_1) - 24PP_1P_2P_\beta(Y_1) + 15P_1^2P_2P_\beta(Y_1) - 9D(Y_2)P_1^2P_2P_\beta(Y_1) + 48PP_1^2P_2P_\beta(Y_1)\\
&+ 6P_1P_2^2P_\beta(Y_1) - 36D(Y_2)P_1P_2^2P_\beta(Y_1) - 12P_1^2P_2^2P_\beta(Y_1) + 72D(Y_2)P_1^2P_2^2P_\beta(Y_1) + 9P_1^2P_2P_{\alpha\beta}(Y_2)P_\beta(Y_1)\\
&- 18P_1P_2^2P_{\alpha\beta}(Y_2)P_\beta(Y_1) + 36P_1^2P_2^2P_{\alpha\beta}(Y_2)P_\beta(Y_1) - 15P_1^2P_2P_{\alpha\beta\gamma}(Y_2)P_\beta(Y_1) - 6P_1P_2^2P_{\alpha\beta\gamma}(Y_2)P_\beta(Y_1)\\
&+ 12P_1^2P_2^2P_{\alpha\beta\gamma}(Y_2)P_\beta(Y_1) - 24PP_1P_2P_\beta(Y_2) + 6P_1^2P_2P_\beta(Y_2) - 36D(Y_1)P_1^2P_2P_\beta(Y_2) - 6PP_2^2P_\beta(Y_2) + 15P_1P_2^2P_\beta(Y_2)\\
&- 9D(Y_1)P_1P_2^2P_\beta(Y_2) + 48PP_1P_2^2P_\beta(Y_2) - 12P_1^2P_2^2P_\beta(Y_2) + 72D(Y_1)P_1^2P_2^2P_\beta(Y_2) - 18P_1^2P_2P_{\alpha\beta}(Y_1)P_\beta(Y_2)\\
&+ 9P_1P_2^2P_{\alpha\beta}(Y_1)P_\beta(Y_2) + 36P_1^2P_2^2P_{\alpha\beta}(Y_1)P_\beta(Y_2) - 6P_1^2P_2P_{\alpha\beta\gamma}(Y_1)P_\beta(Y_2) - 15P_1P_2^2P_{\alpha\beta\gamma}(Y_1)P_\beta(Y_2)\\
&+ 12P_1^2P_2^2P_{\alpha\beta\gamma}(Y_1)P_\beta(Y_2) - 18P_1^2P_2P_\beta(Y_1)P_\beta(Y_2) - 18P_1P_2^2P_\beta(Y_1)P_\beta(Y_2) + 36P_1^2P_2^2P_\beta(Y_1)P_\beta(Y_2)
\end{aligned}
$$

Figure 4.4: An expression for $D(X)$ in the case $r = 3$ analogous to equation (4.5), computed with the Mathematica notebook in Section A.4.

Case 3. If $\alpha \notin A_i$ and $c \in A_i$, we can use the set $A' = (A_i \backslash \{c\}) \cup \{\alpha\}$ to express the probability

$$
\begin{aligned}
\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = c) &= \mathbb{P}(\mathrm{MP}(Y_i) = (A_i \backslash \{c\}) \cup \{\alpha\} | \xi_{y_i} = \alpha)\\
&= P_{(A_i \backslash \{c\}) \cup \{\alpha\}}(Y_i).
\end{aligned}
$$

Case 4. If $\alpha \in A_i$ and $c \notin A_i$, we can use the set $A' = (A_i \backslash \{\alpha\}) \cup \{c\}$ to express the probability

$$
\begin{aligned}
\mathbb{P}(\mathrm{MP}(Y_i) = A_i | \xi_{y_i} = c) &= \mathbb{P}(\mathrm{MP}(Y_i) = (A_i \setminus \{\alpha\}) \cup \{c\} | \xi_{y_i} = \alpha)\\
&= P_{(A_i \backslash \{\alpha\}) \cup \{c\}}(Y_i).
\end{aligned}
$$

This allows $R^{\mathcal{T}}_{A_i}(Y_i)$ to be expressed as sum of probabilities $P_{A'}(Y_i)$ for appropriate sets $A' \subseteq C$ multiplied with $1 - p_i$ or $\frac{p_i}{r-1}$. In total $P_A(X)$ can be expressed as sum of products of $P^T_{A'}(Y_i)$, $1 - p_i$ or $\frac{p_i}{r-1}$ for $i = 1$ or $i = 2$. This completes the cooking recipe for the expression $D(X)$.

The Mathematica notebook in Section A.4 computes $D(X)$ as explained in the previous remark. The resulting expression for $r = 3$ is displayed in Figure 4.4. In the binary case we then concluded, that $D(X) \geq 0$ because all summands are nonnegative. However, if $r = 3$ or $r = 4$ there are negative summands (see Figure 4.4 for the case $r = 3$) and in addition Table 4.1 gives values for $r = 3$, where $D(X) < 0$. But we do not know if there is a phylogenetic tree with such probabilities, i.e. if these values do occur in the sense of Definition 4.7, and if they do how

to construct such a phylogenetic tree. The following lemma lists further properties of occurring probabilities as we did already in Example 4.8 for the obvious property, that for all occurring probabilities $0 \leq P_\alpha(Y_i) \leq 1$ holds.

Lemma 4.10. *Let $\mathcal{T}$ be a rooted binary phylogenetic tree as in Conjecture 4.6 and $P_A(Y_i)$, P, P_i for $\emptyset \neq A \subseteq C$ and $i = 1, 2$ probabilities as in Conjecture 4.6. If the values occur in the sense of Definition 4.7, the following properties are satisfied:*

(i) $P < P_1$ *and* $P < P_2$.

(ii) For $i = 1, 2$

$$\sum_{\substack{A \subseteq C, A \neq \emptyset \\ A \neq \{\alpha\}}} P_A(Y_i) < 1.$$

(iii) For $i \in \{1, 2\}$ *is either* $p_i = p$ and $P_\alpha(Y_i) = 1$, or

$$P_A(Y_i) > 0 \quad \forall \emptyset \neq A \subseteq C, |A| \leq 2.$$

Proof. (i) follows directly from Lemma 2.15, because $0 < P_i < 1$. Furthermore, we have $P_\alpha(Y_i) > 0$, because the conservation probability $\mathbb{P}(\xi_{y_i} = \xi_v)$ from y_i to any leaf $v \in V(T_1)$ satisfies $\mathbb{P}(\xi_{y_i} = \xi_v) > 0$ and therefore the probability that every leaf is in state α is positive. The sum

$$\sum_{A \subseteq C, A \neq \emptyset} P_A(Y_i) \leq 1$$

is a sum of probabilities of disjoint events and therefore less or equal than 1. $P_\alpha(Y_i) > 0$ implies (ii). (iii) describes that either T_i consists only of one vertex and in this case $p_i = p$ and $P_\alpha(Y_i) = 1$, or T_1 consists of at least three vertices and because $p_e > 0$ every combination of states is possible for the leaves in T_1 and this implies $P_A(Y_i) > 0$ for $|A| \leq 2$. □

The values listed in Table 4.1 fulfill this three properties. Since it is unclear, if this list of conditions can be continued or completed, we tried other approaches which might be more promising to prove or disprove Conjecture 4.6.

Proofs for some simple cases with a computer algebra system. In order to examine the conjecture for some specific trees, we developed the Mathematica package `Phylgen` (the source code is completely listed in Section A.5). With its help it is possible to define a tree structure

$D(Y_1)$	$D(Y_2)$	P	P_1	P_2	$P_{\alpha\beta}(Y_1) = P_{\alpha\gamma}(Y_1)$	$P_{\alpha\beta\gamma}(Y_1)$
0.01	0.01	0.01	0.1	0.99	0.1	0.7

$P_{\alpha\beta}(Y_2) = P_{\alpha\gamma}(Y_2)$	$P_{\alpha\beta\gamma}(Y_1)$	$P_{\beta}(Y_1) = P_{\gamma}(Y_1)$	$P_{\beta}(Y_2) = P_{\gamma}(Y_2)$
0.01	0.01	0.01	0.01

Table 4.1: Values where $D(X) < 0$ for $r = 3$

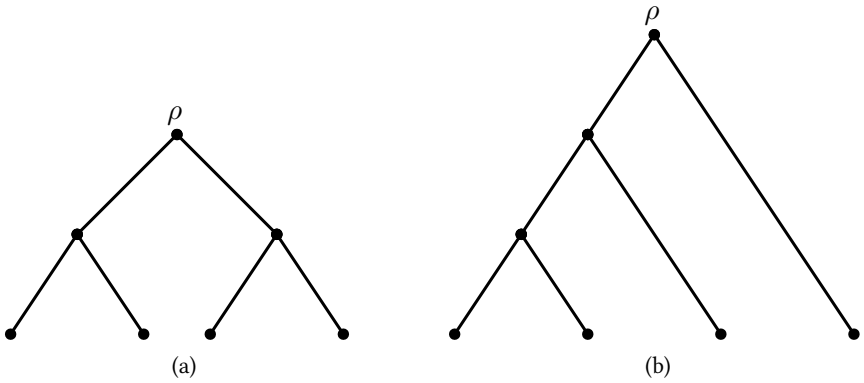


Figure 4.5: Some simple rooted binary trees, where Conjecture 4.6 holds for $r = 4$ and $r = 3$.

and associated edge weights and to calculate its adjacency matrix, the substitution probability for a certain path in the tree, and the reconstruction accuracy for the tree. For the purpose of defining the tree a syntax similar to the *Newick format* (see e.g. [56, p. 10]) is used although it is strongly adopted to the syntax of Mathematica. Note that there might be more efficient methods to compute the reconstruction accuracy, but for our purpose it suffices to follow the approach presented in Remark 4.9.

We used the package `Phylgen` to define the trees in Figure 4.5 in a Mathematica notebook, see Section A.6. Unassigned variables are used as edge weights and therefore the function `ReconstructionAccuracy` from the package `Phylgen` returns the reconstruction accuracy $\mathrm{RA}(X)$ for the trees T in Figure 4.5 as function of the substitution probabilities p_e for $e \in E(T)$. Some of these unknowns p_e can be substituted because the tree has to be ultrametric and therefore all substitution probabilities from the root to the leaves must be equal. Having done that, the Mathematica function `Reduce` outputs true for the implication

$$\forall e \in E(T) : 0 < p_e < \frac{r-1}{r} \Rightarrow \mathrm{RA}(X; \mathcal{T}) \geq 1 - p,$$

where the condition for the substitution probabilities follows from (2.6) on page 33 and $\mathcal{T} =$

(T, ϕ) being one of the trees in Figure 4.5.

This proves Conjecture 4.6 for the trees in Figure 4.5 as expressed in the following remark.

Remark 4.11. Let T be one of the trees displayed in Figure 4.5 and $\mathcal{T} = (T, \phi)$ a rooted binary phylogenetic tree as in Conjecture 4.6 with $r = 3$ or $r = 4$. Then the inequality

$$\mathrm{RA}(X; \mathcal{T}) \geq 1 - p$$

holds. In addition this inequality holds if T is a tree with 2 or 3 leaves as can be shown by hand or with help of Mathematica.

Unfortunately it is not easy possible to examine this statement in the same way for trees with more than 4 leaves. To compute the reconstruction accuracy using the package `Phylgen`, takes several hours, even for a tree with only 5 leaves and unknown edge weights. It might be possible to improve the efficiency of the algorithm, which would allow to extend the statement to trees with more leaves.

Of course, we cannot draw any conclusions from this remark to a statement for all trees. It is only possible to exclude the examined trees as possible counter examples. However, the proof sketch of Theorem 4.3 showed that such a counter example could be a tree with a huge number of vertices.

Further ideas. Finally, we want to present an idea, which could lead to a proof of Conjecture 4.6 with an approach, which is different from the one Fischer and Thatte [24] used for the binary case.

Consider a sequence of phylogenetic trees under the N_r-model, converging to a cherry as illustrated in Figure 4.6 and enunciated formally in the Conjecture 4.12. The length of the two edges attached to the root converges to the height of the tree while the length of all other edges converges from above to 0. We conjecture that there is something like monotony and continuity for the reconstruction accuracy $\mathrm{RA}(\mathcal{T}_n; X)$ for an appropriate sequence $(\mathcal{T}_n)_{n\geq 0}$ of phylogenetic trees, namely

$$\mathrm{RA}(\mathcal{T}_n; X) \geq \mathrm{RA}(\mathcal{T}_{n+1}; X)$$

and

$$\mathrm{RA}(\mathcal{T}_n; X) \to \mathrm{RA}(\mathcal{T}_\infty; X) \quad \text{as } n \to \infty,$$

where $\mathcal{T}_\infty$ denotes the cherry with height p, i.e. the rooted phylogenetic tree with two leaves and

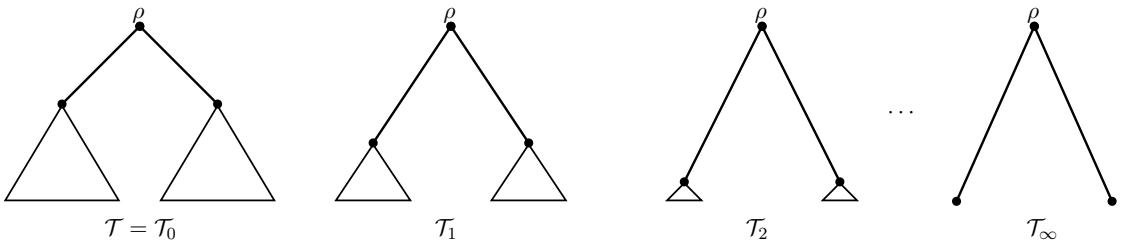


Figure 4.6: A sequence of rooted phylogenetic trees converging to a cherry.

height p. That would be an even stronger statement than Conjecture 4.6 because for any $r \geq 2$

$$\mathrm{RA}(\mathcal{T}_\infty; X) = (1-p)^2 + \frac{1}{2}p(1-p) + \frac{1}{2}(1-p)p = 1-p.$$

Conjecture 4.12. *Let $\mathcal{T} = (T, \phi, p)$ be a binary rooted ultrametric phylogenetic tree under the N_r-model of height p, i.e. $p = \mathbb{P}(\xi_\rho \neq \xi_v)$ for any leaf $v \in V(T)$. Then there exists a sequence $(\mathcal{T}_n)_{n\in\mathbb{N}}$ with $\mathcal{T}_n = (T, \phi, p_n)$ of binary rooted ultrametric phylogenetic trees under the N_r-model of equal height with the following properties:*

(i) $\mathcal{T} = \mathcal{T}_0$, i.e. $p(e) = p_0(e)$ for all edges $e \in E(T)$.

(ii) For $e = (\rho, v)$ the edge weight $p_n(e)$ converges to p from below, that is

$$p_n(e) \to p \quad \text{as } n \to \infty$$

and

$$p_{n+1}(e) \leq p_n(e) \quad \forall n \geq 0.$$

(iii) $\mathrm{RA}(X; \mathcal{T}_n)$ converges to $\mathrm{RA}(X; \mathcal{T}_\infty)$ from above, that is

$$\mathrm{RA}(X; \mathcal{T}_n) \to \mathrm{RA}(X; \mathcal{T}_\infty) \quad \text{as } n \to \infty$$

and

$$\mathrm{RA}(X; \mathcal{T}_n) \geq \mathrm{RA}(X; \mathcal{T}_{n+1}) \quad \forall n \geq 0.$$

If the height of the subtrees T_1 and T_2 is very small, it is very unlikely that $\overline{\chi}(u) \neq \overline{\chi}(v)$ for an edge $(u, v) \in V(T_i)$. Therefore with high probability we have $\overline{\chi}(y_i) = \overline{\chi}(v)$ where v is a leaf $v \in V(T_i)$ and y_i the root of T_i as illustrated in Figure 4.3. In addition, the substitution probability between ρ and y_i equals approximately p. Hence, the reconstruction probability

equals approximately the reconstruction accuracy of the tree $\mathcal{T}_\infty$ with two leaves. If it is assumed that Conjecture 4.6 is true, these informal considerations suggest that there might be a sequence $(\mathcal{T}_n)_{n\geq 0}$ of rooted phylogenetic trees with the desired properties. In addition, Conjecture 4.12 is supported by experiments with some specific trees—we could not find any tree, where it does not hold. But it is unclear if an appropriate sequence of rooted phylogenetic trees can be constructed easily in general.

It might be possible to use Remark 4.9 and its implementation in Mathematica in Section A.4 to prove the monotony, if one uses $P_\alpha^{\mathcal{T}_n}(Y_i) \approx 1$ and $P_A^{\mathcal{T}_n}(Y_i) \approx 0$ for $A \neq \{\alpha\}$, $i = 1, 2$ and a large n.

Appendix A

Source code

A.1 Compute the number of rooted phylogenetic trees

A.1.1 Explicit formula

Mathematica code to calculate the number r_n of rooted phylogenetic trees by use of (3.10) in Theorem 3.12. The first part computes the associated Stirling numbers of the second kind using Mathematica code by Jean-François Alcover and a formula by D. Wasserman (see sequence A008299 in [64]).

```
(* Compute associated Stirling number... *)
(* From Jean-François Alcover, Oct 13 2011, after David
   Wasserman *)
s2[n_, k_] := Sum[
    (-1)^i * Binomial[n, i] *
    Sum[
        (-1)^j*(k - i - j)^(n - i)/(j!*(k - i - j)!),
        {j, 0, k - i}
    ],
    {i, 0, k}
];

(* Compute number of rooted phylogenetic trees... *)
(
    n = 200;
    Sum[
```

```
        s2[n + k, k + 1],
        {k, 0, n - 2}
    ]
) // Timing
```

A.1.2 Recursive formula

The following Mathematica code implements Felsenstein's recurrence relation (see (3.6) in Theorem 3.7) to calculate the number r_n of rooted phylogenetic trees.

```
(* Compute number of rooted phylogenetic trees
   using Felsenstein's recurrence relation... *)
(
    n = 200;
    column = ConstantArray[0, n - 1];
    column[[1]] = 1;
    For[j = 3, j <= n, j++,
        column =
            column *
            Flatten[{Range[j - 2], ConstantArray[0, n - j +
                1]}]
              +
            Drop[Prepend[column, 0], -1] *
            Flatten[{Range[j - 1, 2*j - 3], ConstantArray[0, n
                - j]}];
    ];
    Sum[column[[i]], {i, n - 1}]
) // Timing
```

A.2 Compute the number of X-trees

The following Mathematica code was used in order to compute the values in Table 3.3 on page 72.

```
length = 50;
```

```
(* p[[n]] = number of planted X-trees with |X|=n *)
p = ConstantArray[0, length];
p[[1]] = 1;
For[n = 2, n <= length, n++,
  p[[n]] = 2*p[[n - 1]]
    + Sum[Binomial[n, k]*p[[k]]*p[[n - k]], {k, 1, n - 1}]
]
p // MatrixForm

(* u[[n]] = number of unrooted X-trees with |X|=n *)
u = Drop[Prepend[2*p, 1], -1];
u // MatrixForm

(* mu[[n]] = mean of vertices in an X-tree with |X|=n *)
mu = Prepend[
  ((1/2) * Drop[p, 1] + Drop[p, -1])/Drop[u, 1],
  1];
N[mu] // MatrixForm

(* sig[[n]] = mean of vertices in an X-tree with |X|=n *)
a = ConstantArray[0, length];
a[[1]] = 0;
For[n = 2, n <= length, n++,
  a[[n]] = p[[n]] - 2*p[[n - 1]] +
    2*Sum[Binomial[n, k]*p[[n - k]]*a[[k]], {k, 1, n - 1}];
]
sig = a/Drop[u, -1] + mu - mu^2;
N[sig] // MatrixForm
```

A.3 Simplifying an expression for the proof of Theorem 4.5

The following Mathematica code completes the proof of Theorem 4.5.

```
(* Applying law of total probability *)
RaY[i_]  := (1 - p[i])*PaY[i] + p[i]*PbY[i];
RbY[i_]  := (1 - p[i])*PbY[i] + p[i]*PaY[i];
RabY[i_] := PabY[i];

PaX  = RaY[1]*RaY[2] + RabY[1]*RaY[2] + RaY[1]*RabY[2]
PabX = RabY[1]*RabY[2] + RaY[1]*RbY[2] + RbY[1]*RaY[2]

(* Transformation to simplify expressions *)
p[i_] := (1 - P[i])/2;

(* The probabilities sum up to 1 *)
PbY[i_] := 1 - PaY[i] - PabY[i];

(* The conservation probability in T_i after the transformation
    equals (1+(P/P[i]))/2 *)
PaY[i_] := DY[i] - ((PabY[i])/2) + (1 + (P/P[i]))/2

FullSimplify[4*DX]
```

A.4 Automatically computing $D(X)$

The following Mathematica notebook computes an expression for $D(X)$ as described in Section 4.4. In the Mathematica notebook the variables are named in the following way:

`setC`	set C of states
`setA`	subset $A \subseteq C$
`probPY[i, states]`	$\mathbb{P}(\mathrm{MP}(Y_i) = A_i \mid y_i = \alpha)$
`probR[i, states]`	$\mathbb{P}(\mathrm{MP}(Y_i) = A_i \mid \rho = \alpha)$
`simplifiedDX`	simplified expression for $D(X)$, i.e. $D(X) = \mathrm{RA}(X) - (1-p)$
`p0`	substitution probability from ρ to any leaf in $\mathcal{T}$
`p[i]`	substitution probability from ρ to y_i
`probP0`	$P := 1 - \frac{r}{r-1}p$

probP[i]	$P_i := 1 - \frac{r}{r-1} p_i$
q[i]	substitution probability form y_i to leaf

```
(*
For states {a,a1,a2,...,an} = setC this notebook calculates an
    expression for D(X).

Note that one state in setC has to be called a!

We will use lists as sets. We have to assure that lists are
    always sorted and do not contain any duplicates. The
    function Union[list] is a useful tool for this task.
*)

(* Load the package for combinatorics *)
<< Combinatorica`

(* * * * * * Constants * * * * * *)
setC = Union[{a, b, c}]; (* no duplicates allowed! 'a' needs to
     be in the list! *)
(*setC = Union[{a,b,c,d}];*) (* uncomment this for case r=4 *)
r = Length[setC];

(* * * * * * Reconstruction Accuracy RA(X) * * * * * *)
(* The reconstruction accuracy RA(X) is given by the sum over
    all subsets of setA\{a}. Every summand is given by the
    probability that MP equals the union of the given subset and
     state a and that the state a is selected uniformly.
*)
RAX = Sum[
     (1/(Length[states] + 1))*(probPX[Union[states, {a}]]),
     {states, Subsets[Complement[setC, {a}]]}
];
```

```
(* * * * * * Recursive expression of P_A(X) * * * * * *)
(* Expressing the probability probR[i,setAi] through probPY[i,
   setAi].
(This is done by the law of total probability and using the
   symmetry of the N_r-model.)
*)
probR[i_, setAi_] := (
    (1 - p[i])*probPY[i, setAi] + (p[i]/(r - 1))*
    Sum[
        If[MemberQ[setAi, a] == MemberQ[setAi, state],
            probPY[i, setAi],
            If[MemberQ[setAi, a],
                probPY[i, Union[setAi /. a -> state ]],
                probPY[i, Union[setAi /. state -> a]]
            ]
        ],
        {state, Complement[setC, {a}]}
    ]
)

(*
For states a,a[1],a[2],...,a[n]=setA the function probPX[]
   returns an expression for the probability P_setA(X) = Pr (MP
    (X) = setA|rho = a). setA has to be subset of setC!
*)
probPX[setA_] := (
    (* parameter setA has to be a subset of setC without
       duplicates! *)

    (* Step 1 *)
    (* sum1 is the sum over all subsets setA1 and setA2 of setC
        so that the intersection of setA1 and setA2 equals setA
        *)
```

```
tmp = Flatten[
    Table[
        Table[
            {Union[tmp1, setA], Union[tmp2, setA]},
            {tmp2, Subsets[Complement[setC, setA, tmp1]]}
        ],
        {tmp1, Subsets[Complement[setC, setA]]}
    ],
    1
];
sum1 = Sum[
    probR[1, settupel[[1]] ] * probR[2, settupel[[2]] ],
    {settupel, tmp}
];

(* Step 2 *)
(* sum2 is the sum over all subsets setA1 and setA2 of setC
    so that the union of setA1 and setA2 equals setA *)

tmp = KSetPartitions[setA, 2];
sum2 = Sum[
    probR[1, partitionset[[1]]] * probR[2, partitionset
       [[2]]],
    {partitionset, tmp}
] + Sum[
    probR[1, partitionset[[2]]] * probR[2, partitionset
       [[1]]],
    {partitionset, tmp}
];

(* Result of step 1 and step 2 *)
(* P_setA (X) equals the sum of the two partial results *)
sum1 + sum2
```

```
)

(* * * * * * Substitution probability for subtrees * * * * * *)
(* q[i] is the height of the subtree T_i *)
q[1] = (p0 - p[1])/(1 + p[1]*(((r - 2)/(r - 1)) - 2));
q[2] = (p0 - p[2])/(1 + p[2]*(((r - 2)/(r - 1)) - 2));


(* * * * * * Assumptions and variable elimination * * * * * *)
(* Due to symmetry many probPY[i, states] are equal. E.g.
   probPY[i,{b}]==probPY[i,{c}]. Two such variables probPY[i,
   states1] and probpY[i,states2] are equal if a is in states1
   and states2 or neither in both of them and if the size of
   states1 and states2 equals. Therefore we can replace probPY[
   i,states] by a new variable which counts only the number of
   states and if a is included or not in the set of states. *)
Do[
    (
        probPY[1, subset] = If[MemberQ[subset, a],
            probPYcountWitha[1, Length[subset]],
            probPYcountWithouta[1, Length[subset]]
        ];
        probPY[2, subset] = If[MemberQ[subset, a],
            probPYcountWitha[2, Length[subset]],
            probPYcountWithouta[2, Length[subset]]
        ];
    ),
    {subset, Subsets[setC]}
]

(* D1 and D1 are defined by RA(Yi) - (1-q[i]). Note that probPY
   [i,{a}] == probPYcountWitha[i,1]. We eliminate the variable
   probPYcountWitha[i,1]. *)
```

```
probPYcountWitha[1, 1] = D1 - Sum[
    (1/(Length[states] + 1))*(probPY[1, Union[states, {a}]]),
    {states, Complement[Subsets[Complement[setC, {a}]], {{}}]}
] + (1 - q[1]);
probPYcountWitha[2, 1] = D2 - Sum[
    (1/(Length[states] + 1))*(probPY[2, Union[states, {a}]]),
    {states, Complement[Subsets[Complement[setC, {a}]], {{}}]}
] + (1 - q[2]);

(* The sum of all probabilities equals 1. On right side we sum
   over all subsets of setC except {} and except setC\{a}. On
   the left side we express probPY[i, setC\{a}] by probPYcount
   [1,withouta,Length[setC]-1]. *)
probPYcountWithouta[1, Length[setC] - 1] = 1 - Sum[
        probPY[1, states],
        {states, Complement[Subsets[setC], {Complement[setC, {
            a}]}, {{}}]}
    ];
probPYcountWithouta[2, Length[setC] - 1] =  1 - Sum[
        probPY[2, states],
        {states, Complement[Subsets[setC], {Complement[setC, {a
            }]}, {{}}]}
    ];

(* * * * * * Transformation to [0,1] * * * * * * *)
(* A simple transformation makes some expressions much simpler.
    While p[i] and p0 are probabilities between 0 and (r-1)/r,
   probP[i] and probP0 are between 0 and 1. *)
p[1] = (1 - probP[1])/(r/(r - 1));
p[2] = (1 - probP[2])/(r/(r - 1));
p0 = (1 - probP0)/(r/(r - 1));

(* * * * * * The equation D(X) * * * * * * *)
```

```
(* This is the final expression for D(X) *)
simplifiedDX = FullSimplify[(RAX - (1 - p0))];
```

A.5 Mathematica package `Phylgen`

```
(* ::Package:: *)
BeginPackage["Phylgen`", "GraphUtilities`", "Combinatorica`"];

(* ::Input:: *)

(* Run "?Functionname" to get this help texts *)
PlotPhylTree::usage="PlotPhylTree[matrix] Plot a phylogenetic
    tree, matrix ist the adjacency matrix";
getAdjacencyMatrix::usage="getAdjacencyMatrix[newickTree]
    convert a tree  in Newick format, for example {a, 0.3, {b,
    0.2, c, 0.2}, 0.3}, to an adjacency matrix";
ReconstructionAccuracy::usage="ReconstructionAccuracy[
    newickTree,numberSates]";
getPathLength::usage="getPathLength[{prob1,prob2,...},r]";


Begin["`Private`"];
    (* getXY[matrix,{x,y}] returns the entry x,y of the Matrix
       *)
    getXY[matrix_, xy_]:=matrix[[  xy[[1]] ]][[  xy[[2]]  ]];


    (* Plot tree, see usage of function *)
    PlotPhylTree[tree_]:=TreePlot[
        tree,Top,1,
            DirectedEdges->True,
            EdgeLabeling-> True,
            VertexLabeling->False,
```

```
        EdgeRenderingFunction -> ({
            Black,
            Arrow[#1,0],
            Text[getXY[tree,#2],LineScaledCoordinate
                [#1,.5],Background->White]
        }&)
]


(* Converts a tree in pseudo Newick format to an adjacency
   matrix... *)
getAdjacencyMatrix[treeNewick_]:= (
    Module[{matrix1,matrix2,newsize,tree}, (
        If[ListQ[ treeNewick[[1]] ],
            matrix1 = getAdjacencyMatrix[ treeNewick[[1]]
                ];,
            matrix1 = {{0}};
        ];

        If[ListQ[ treeNewick[[3]] ],
            matrix2 = getAdjacencyMatrix[ treeNewick[[3]]
                ];,
            matrix2 = {{0}};
        ];

        newsize = (Length[matrix1] + Length[matrix2]);

        tree = ArrayFlatten[{{
            ConstantArray[0,{newsize+1,1}],
            ArrayFlatten[{
                {ConstantArray[0,{1,newsize}]},
                {ArrayFlatten[
                    {{matrix1,0},
```

```
                    {0, matrix2}}
                ]}
            }]
        }}];

        tree[[1]][[2]] = treeNewick[[2]];
        tree[[1]][[ Length[matrix1]+2 ]] = treeNewick[[4]];

        (* Return the adjacency matrix... *)
        tree
    )]
)


(* Calculates the reconstruction accuracy for a given tree
   and a specific number of states... *)
ReconstructionAccuracy[treeNewick_, numberStates_]:=(
    (* The states are denoted by a[1], a[2], ..., a[
       numberStates] *)
    setC = Map[a,Range[numberStates]]; r=Length[setC];

    Sum[
        (1/(Length[states]+1))*(probP[treeNewick,Union[
            states,{a[1]}]]),
             {states,Subsets[Complement[setC,{a[1]}]]}
    ]
)


(*  returns the probability that the Fitch algorithm will
   reconstruct the state set setA on the tree treeNewick...
    *)
probP[treeNewick_,setA_]:= (
```

```
Module[{sum1,sum2,tmp},(
    (* parameter setA has to be a subset of setC
       without duplicates! *)

    (* Step 1 *)
    (* sum1 is the sum over all subsets setA1 and setA2
        of setC so that the intersection of setA1 and
       setA2 equals setA *)
    tmp = Flatten[
        Table[
            Table[
                {Union[tmp1,setA], Union[tmp2,setA]},
                {tmp2, Subsets[Complement[setC,setA,
                    tmp1]]}
            ],
            {tmp1,Subsets[Complement[setC,setA]]}
        ],
        1
    ];
    sum1 = Sum[
        probR[1,treeNewick,settupel[[1]] ] *    probR
           [2,treeNewick,settupel[[2]] ],
        {settupel, tmp}
    ];


    (* Step 2 *)
    (* sum2 is the sum over all subsets setA1 and setA2
        of setC such that the union of setA1 and setA2
       equals setA *)
    tmp=KSetPartitions[setA,2];
    sum2 = Sum[probR[1,treeNewick,partitionset[[1]] ] *
        probR[2,treeNewick,partitionset[[2]] ],{
```

```
            partitionset,tmp}] +
        Sum[probR[1,treeNewick,partitionset[[2]] ] * probR
            [2,treeNewick,partitionset[[1]] ],{partitionset,
            tmp}];

        (* Result of step 1 and step 2 *)
        (* P_setA (X) equals the sum of the two partial
            results *)
        sum1 + sum2
    )]
)


(* Expressing the probability probR[i,setAi] = Prob(MP(Y_i)
    =setAi|rho=a) through probPY[i, setAi] = Prob (MP(Y_i)=
    setAi|yi=a).
(This is done by the law of total probability and using the
    symmetry of the N_r-model.)  *)
probR[i_,treeNewick_,setAi_]:=(
    (* treeNewick[[2*i]] equates to p[i] *)
    (1-treeNewick[[2*i]])*probPY[i,setAi,treeNewick] +  (
        treeNewick[[2*i]]/(r-1))*
    Sum[
        If[MemberQ[setAi,a[1]] == MemberQ[setAi,state],
            probPY[i,setAi,treeNewick],
                If[MemberQ[setAi,a[1] ],
                    probPY[i,Union[setAi /. a[1] -> state
                        ],treeNewick],
                    probPY[i,Union[setAi /. state -> a[1]],
                        treeNewick]
                ]
            ],
        {state, Complement[setC, {a[1]}]}
```

```
        ]
    )


    (* Returns the probability that the Fitch algorithm
       reconstructs the state set setAi in the i-th subtree of
       treeNewick for the root of this subtree under condition
       root of this subtree is a[1] *)
probPY[i_,setAi_,treeNewick_]:=(
        (* Note: treeNewick[[2*i-1]] is the first subtree, if i
           ==1 or the 2nd subtree if i==2 *)
        If[ListQ[treeNewick[[2*i-1]]],
            probP[treeNewick[[2*i-1]],setAi],
            If[setAi==={a[1]},
                1,
                0
            ]
        ]
    )


    (* Calculates the path length for a list of substitution
       probabilities *)
    getPathLength[probabilities_,r_:r]:=((r-1)/r) *
        (1-Product[(1-(r/(r-1))*prob),{prob,probabilities}])

End[];

EndPackage[]
```

A.6 A Mathematica proof of Conjecture 4.6 for some specific trees

A.6.1 Proof for the tree in Figure 4.5a

```
(* Load the Phylgen package *)
<< Phylgen`
(* Define the tree in a Newick-like format... *)
tree = {{a, p3, b, p3}, p1, {c, p4, d, p4}, p2};

(* Plot the tree... *)
PlotPhylTree[getAdjacencyMatrix[tree]]

(* Define the number of character states... *)
r = 4;
pathlength = FullSimplify[getPathLength[{p1, p3}, r]];

(* The following equation needs to be fulfilled too,
because of the ultrametricity of the tree... *)
sol1 = Solve[pathlength == getPathLength[{p2, p4}, r], p4];
p4 = p4 /. sol1[[1]];
p4 = FullSimplify[p4];

(* Tree is ultrametric *)
Reduce[ pathlength == getPathLength[{p1, p3}, r] ]
Reduce[ pathlength == getPathLength[{p2, p4}, r] ]

(* Find a possible counter example... *)
FindInstance[
    (
        0 < p1 < (r - 1)/r &&
        0 < p2 < (r - 1)/r &&
        0 < p3 < (r - 1)/r &&
        0 < p4 < (r - 1)/r &&
        (1 - pathlength) > recon
    ),
    {p1, p2, p3},
    Reals
```

```
]

Reduce[
    Implies[
        (
            0 < p1 < (r - 1)/r &&
            0 < p2 < (r - 1)/r &&
            0 < p3 < (r - 1)/r &&
            0 < p4 < (r - 1)/r
        ),
        (1 - pathlength) <= recon
    ],
    Reals
]
```

A.6.2 Proof for the tree in Figure 4.5b

```
(* Load the Phylgen package *)
<< Phylgen`

(* Define the tree in a Newick-like format... *)
tree = {{{a, q3, b, q3}, q2, c, p2}, q1, d, p1};

(* Plot the tree... *)
PlotPhylTree[getAdjacencyMatrix[tree]]

(* Define the number of character states... *)
r = 4;

pathlength = p1;

(* The following equation needs to be fulfilled too,
because of the ultrametricity of the tree... *)
sol1 = Solve[pathlength == getPathLength[{q1, p2}, r], p2];
```

```
p2 = p2 /. sol1[[1]];
p2 = FullSimplify[p2];
sol2 = Solve[pathlength == getPathLength[{q1, q2, q3}, r], q3];
q3 = q3 /. sol2[[1]];
q3 = FullSimplify[q3];

(* Tree is ultrametric *)
Reduce[ pathlength == getPathLength[{p1}, r] ]
Reduce[ pathlength == getPathLength[{q1, p2}, r] ]
Reduce[ pathlength == getPathLength[{q1, q2, q3}, r] ]

recon = FullSimplify[ReconstructionAccuracy[tree, r]];

(* Find a possible counter example... *)
FindInstance[
    (
        0 <= p1 < (r - 1)/r &&
        0 <= p2 < (r - 1)/r &&
        0 <= q1 < (r - 1)/r &&
        0 <= q2 < (r - 1)/r &&
        0 <= q3 < (r - 1)/r &&
        (1 - pathlength) > recon
    ),
    {p1, q1, q2},
    Reals
]

Reduce[
    Implies[
        (
            0 <= p1 < (r - 1)/r &&
            0 <= p2 < (r - 1)/r &&
            0 <= p3 < (r - 1)/r &&
```

```
            0 <= q1 < (r - 1)/r &&
            0 <= q2 < (r - 1)/r &&
            0 <= q3 < (r - 1)/r &&
            0 <= q4 < (r - 1)/r
    ),
    (1 - pathlength) <= recon
    ],
    Reals
]
```

Bibliography

[1] C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis: A hitchhiker's guide*, 3rd ed. Berlin: Springer, 2006.

[2] B. L. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Annals of Combinatorics*, vol. 5, no. 1, pp. 1–15, 2001.

[3] E. S. Allman and J. A. Rhodes, "Phylogenetics," in *Modeling and simulation of biological networks*, ser. Proceedings of Symposia in Applied Mathematics. American Mathematical Society, 2007, vol. 64, pp. 21–52.

[4] Q. D. Atkinson and R. D. Gray, "Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513–526, 2005.

[5] P. E. Black. (2008, August) Binary tree. Published electronically. Access: 2012-04-18. [Online]. Available: http://www.nist.gov/dads/HTML/binarytree.html

[6] M. Bóna and P. Flajolet, "Isomorphism and symmetries in random phylogenetic trees," *Journal of Applied Probability*, vol. 46, no. 4, pp. 1005–1019, 2009.

[7] O. P. Buneman, "The recovery of trees from measures of dissimilarity," in *Mathematics in the Archaeological and Historical Sciences*, F. Hodson, D. Kendall, and P. Tautu, Eds. Edinburgh: Edinburgh University Press, 1971, pp. 387–395.

[8] J. H. Camin and R. R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, vol. 19, no. 3, pp. 311–326, 1965.

[9] M. Carter, M. Hendy, D. Penny, L. A. Székely, and N. C. Wormald, "On the distribution of lengths of evolutionary trees," *SIAM Journal on Discrete Mathematics*, vol. 3, no. 1, pp. 38–47, 1990.

[10] C. A. Charalambides, *Combinatorial methods in discrete distributions*, ser. Wiley series in probability and statistics. New Jersey: John Wiley & Sons Inc., 2005.

[11] L. Comtet, *Advanced combinatorics: the art of finite and infinite expansions*, revised and enlared ed. D. Reidel Publishing Co., 1974.

[12] C. Cotta and P. Moscato, "Inferring phylogenetic trees using evolutionary algorithms," in *Parallel Problem Solving from Nature – PPSN VII*, ser. Lecture Notes in Computer Science, J. Guervós, P. Adamidis, H.-G. Beyer, H.-P. Schwefel, and J.-L. Fernández-Villacañas, Eds. Berlin: Springer, 2002, vol. 2439, pp. 720–729.

[13] M. D. Crisp and L. G. Cook, "Do early branching lineages signify ancestral traits?" *Trends in Ecology & Evolution*, vol. 20, no. 3, pp. 122–128, 2005.

[14] C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, 1859.

[15] ——, "Notebook B: Transmutation of species," 1837–1838, Darwin Manuscripts Catalogue, DAR 121, Cambridge University Library. [Online]. Available: http://darwin.amnh.org/viewer.php?history=&eid=73206

[16] C. Darwin and M. T. Ghiselin, *The descent of man.* John Murray, 1874.

[17] P. Davis, R. Hersh, and E. Marchisotto, *The Mathematical Experience*, study ed., ser. Modern Birkhäuser Classics. Birkhäuser Boston, 2011.

[18] E. Demaine, S. Hohenberger, and D. Liben-Nowell, "Tetris is hard, even to approximate," in *Computing and Combinatorics*, ser. Lecture Notes in Computer Science, T. Warnow and B. Zhu, Eds. Berlin: Springer, 2003, vol. 2697, pp. 351–363.

[19] Department of Mathematics, North Dakota State University. Mathematics genealogy project. Published electronically. Access: 2012-04-20. [Online]. Available: http://genealogy.math.ndsu.nodak.edu/

[20] P. L. Erdős and L. A. Székely, "Counting bichromatic evolutionary trees," *Discrete Applied Mathematics*, vol. 47, no. 1, pp. 1–8, 1993.

[21] J. Felsenstein, "The number of evolutionary trees," *Systematic Biology*, vol. 27, no. 1, pp. 27–33, 1978.

[22] ——, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic Biology*, vol. 27, no. 4, pp. 401–410, 1978.

[23] ——, *Inferring phyogenies*. Massachusetts: Sinauer Associates, 2004.

[24] M. Fischer and B. Thatte, "Maximum parsimony on subsets of taxa," *Journal of theoretical biology*, vol. 260, no. 2, pp. 290–293, 2009.

[25] ——, "Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics," *Bulletin of Mathematical Biology*, vol. 72, pp. 208–220, 2010.

[26] W. M. Fitch, "Toward defining the course of evolution: Minimum change for a specific tree topology," *Systematic Zoology*, vol. 20, no. 4, pp. 406–416, 1971.

[27] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*. Cambridge: Cambridge University Press, 2009.

[28] L. R. Foulds and R. L. Graham, "The steiner problem in phylogeny is NP-complete," *Advances in Applied Mathematics*, vol. 3, no. 1, pp. 43–49, 1982.

[29] L. R. Foulds and R. W. Robinson, "Enumeration of phylogenetic trees without points of degree two," *Ars Combinatoria*, vol. 17A, pp. 196–183, 1984.

[30] ——, "Enumeration of binary phylogenetic trees," in *Combinatorial Mathematics VIII*, ser. Lecture Notes in Mathematics, K. McAvaney, Ed. Springer, 1981, vol. 884, pp. 187–202.

[31] ——, "Counting certain classes of evolutionary trees with singleton labels," in *Proceedings of the fifteenth southeastern conference on combinatorics, graph theory and computing*, vol. 44, 1984, pp. 65–88.

[32] ——, "Enumerating phylogenetic trees with multiple labels," *Discrete Mathematics*, vol. 72, no. 1-3, pp. 129–139, 1988.

[33] ——, "Determining the asymptotic number of phylogenetic trees," in *Combinatorial mathematics, VII (Proc. Seventh Australian Conf., Univ. Newcastle, Newcastle, 1979)*, ser. Lecture Notes in Math. Berlin: Springer, 1980, vol. 829, pp. 110–126.

[34] P. H. Fuss, *Correspondance mathématique et physique de quelques célèbres géomètres du XVIIIème siècle*. New York: Johnson Reprint Corp., 1968, access: 2012-04-17. [Online]. Available: http://eulerarchive.maa.org/correspondence/fuss/

[35] M. R. Garey, R. L. Graham, and D. S. Johnson, "The complexity of computing Steiner minimal trees," *SIAM Journal on Applied Mathematics*, vol. 32, no. 4, pp. 835–859, 1977.

[36] O. Gascuel, Ed., *Mathematics of evolution and phylogeny*. New York: Oxford University Press, 2005.

[37] O. Gascuel and M. Steel, *Reconstructing evolution: new mathematical and computational advances*. Oxford: Oxford University Press, 2007.

[38] D. Gusfield, *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge: Cambridge University Press, 1997.

[39] E. Haeckel, *Anthropogenie oder Entwicklungsgeschichte des Menschen: gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes- und Stammes-Geschichte*, 2nd ed. Leipzig: Wilhelm Engelmann, 1874.

[40] ——, *The Evolution of Man: A Popular Scientific Study*, 5th ed. London: Watts & Co., 1910.

[41] F. Harary, *Graph Theory*. Addison-Wesley Publishing Company, Inc., 1969.

[42] E. F. Harding, "The probabilities of rooted tree—shapes generated by random bifurcation," *Advances in Applied Probability*, vol. 3, no. 1, pp. 44–77, 1971.

[43] J. A. Hartigan, "Minimum mutation fits to a given tree," *Biometrics*, vol. 29, no. 1, pp. 53–65, 1973.

[44] P. Henrici, *Applied and computational complex analysis. Vol. 1*, ser. Wiley Classics Library. New York: John Wiley & Sons Inc., 1988.

[45] F. T. Howard, "Explicit formulas for numbers of ramanujan," *Fibonacci Quarterly*, vol. 24, no. 2, pp. 168–175, 1986.

[46] J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine, "Inferring the root of a phylogenetic tree," *Systematic biology*, vol. 51, no. 1, pp. 32–43, 2002.

[47] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian protein metabolism*, M. N. Munro, Ed. Academic Press, 1969, vol. III, pp. 21–132.

[48] S. K. Lando, *Lectures on generating functions*, ser. Student mathematical library. American Mathematical Society, 2003.

[49] G. Li, M. Steel, and L. Zhang, "More taxa are not necessarily better for the reconstruction of ancestral character states," *Systematic biology*, vol. 57, no. 4, pp. 647–653, 2008.

[50] H. G. Liddell and R. Scott, *A Greek-English lexicon*, 8th ed. Oxford: Clarendon Press, 1901.

[51] C. D. Michener and R. R. Sokal, "A quantitative approach to a problem in classification," *Evolution*, pp. 130–162, 1957.

[52] F. Murtagh, "Counting dendrograms: a survey," *Discrete Applied Mathematics*, vol. 7, no. 2, pp. 191–199, 1984.

[53] L. Nakhleh, G. Jin, F. Zhao, and J. Mellor-Crummey, "Reconstructing phylogenetic networks using maximum parsimony," in *Computational Systems Bioinformatics Conference*, 2005, pp. 93–102.

[54] J. Neyman, "Molecular studies of evolution: a source of novel statistical problems," *Statistical decision theory and related topics*, pp. 1–27, 1971.

[55] R. Otter, "The number of trees," *The Annals of Mathematics*, vol. 49, no. 3, pp. 583–599, 1948.

[56] R. D. M. Page, "Visualizing phylogenetic trees using treeview," *Current Protocols in Bioinformatics*, vol. 6.2, 2003.

[57] H. Philippe and C. Douady, "Horizontal gene transfer and phylogenetics," *Current opinion in microbiology*, vol. 6, no. 5, pp. 498–505, 2003.

[58] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[59] B. A. Salisbury and J. Kim, "Ancestral state estimation and taxon sampling density," *Systematic Biology*, vol. 50, no. 4, pp. 557–564, 2001.

[60] M. J. Sanderson, M. M. McMahon, and M. Steel, "Terraces in phylogenetic tree space," *Science*, vol. 333, no. 6041, pp. 448–450, 2011.

[61] W. H. Schikhof, *Ultrametric Calculus: An Introduction to P-Adic Analysis*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2007.

[62] E. Schröder, "Vier combinatorische Probleme," *Z. Math. Phys*, vol. 15, pp. 361–376, 1870.

[63] C. Semple and M. Steel, *Phylogenetics*, ser. Oxford Lecture Series in Mathematics and its Applications. New York: Oxford University Press, 2003, vol. 24.

[64] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences. Published electronically. Access: 2012-04-25. [Online]. Available: http://oeis.org

[65] E. Sober, "Parsimony in systematics: philosophical issues," *Annual Review of Ecology and Systematics*, vol. 14, pp. 335–357, 1983.

[66] Y. S. Song, "On the combinatorics of rooted binary phylogenetic trees," *Annals of Combinatorics*, vol. 7, no. 3, pp. 365–379, 2003.

[67] M. Spencer, E. A. Davidson, A. C. Barbrook, and C. J. Howe, "Phylogenetics of artificial manuscripts," *Journal of Theoretical Biology*, vol. 227, no. 4, pp. 503–511, 2004.

[68] R. P. Stanley, *Enumerative combinatorics. Vol. I*, ser. The Wadsworth & Brooks/Cole Mathematics Series. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1986, with a foreword by Gian-Carlo Rota.

[69] ——, *Enumerative combinatorics. Vol. 2*, ser. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press, 1999, vol. 62, with a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[70] M. Steel and M. Charleston, "Five surprising properties of parsimoniously colored trees," *Bulletin of Mathematical Biology*, vol. 57, no. 2, pp. 367–375, 1995.

[71] M. Steel and D. Penny, "Two further links between MP and ML under the poisson model," *Applied Mathematics Letters*, vol. 17, no. 7, pp. 785–790, 2004.

[72] M. A. Steel, "Distributions on bicolored evolutionary trees," Ph.D. dissertation, Massey University, Palmerston North, New Zealand, 1989.

[73] ——, "Distributions on bicoloured binary trees arising from the principle of parsimony," *Discrete Applied Mathematics*, vol. 41, no. 3, pp. 245–261, 1993.

[74] M. Syvanen, "Horizontal gene transfer: evidence and possible consequences," *Annual review of genetics*, vol. 28, no. 1, pp. 237–261, 1994.

[75] L. A. Székely, P. L. Erdős, and M. A. Steel, "The combinatorics of evolutionary trees—a survey," in *Séminaire Lotharingien de Combinatoire*, ser. Publ. Inst. Rech. Math. Av. Strasbourg: Univ. Louis Pasteur, 1992, vol. 498, pp. 129–143.

[76] C. Tuffley and M. Steel, "Links between maximum likelihood and maximum parsimony under a simple model of site substitution," *Bulletin of Mathematical Biology*, vol. 59, no. 3, pp. 581–607, 1997.

[77] W. Wallace, "Empedocles," in *Encyclopaedia Britannica: A Dictionary of Arts, Sciences, Literature and General Information*, 11th ed., H. Chisholm, Ed. New York: Cambridge University Press, 1911, vol. IX, pp. 344–345.

[78] T. Warnow, "Mathematical approaches to comparative linguistics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6585–6590, 1997.

[79] E. W. Weisstein. Planted tree. From MathWorld—A Wolfram Web Resource. Published electronically. Access: 2012-04-13. [Online]. Available: http://mathworld.wolfram.com/PlantedTree.html

[80] Wikimedia Foundation. Charles Darwin's 1837 sketch. Published electronically. Access: 2011-10-28. [Online]. Available: https://commons.wikimedia.org/wiki/File:Darwin_tree.png

[81] ——. Diagram representing the divergence of species. Published electronically. Access: 2011-11-22. [Online]. Available: https://commons.wikimedia.org/wiki/File:Darwin_divergence.jpg

[82] ——. Haeckel's paleontological tree of vertebrates. Published electronically. Access: 2011-11-01. [Online]. Available: http://en.wikipedia.org/wiki/File:Age-of-Man-wiki.jpg

[83] H. S. Wilf, *generatingfunctionology*, 3rd ed. Boston: Academic Press, Inc., 1990.

[84] P. Winkler. Seven puzzles you think you must not have heard correctly. Published electronically. Access: 2012-04-24. [Online]. Available: http://www.math.dartmouth.edu/~pw/solutions.pdf

[85] J. Zhang and M. Nei, "Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods," *Journal of molecular evolution*, vol. 44, pp. 139–146, 1997.

Commonly used symbols

$a_n \sim b_n$	asymptotic equivalence of sequences a_n and b_n, i.e. $a_n \sim b_n :\Leftrightarrow \lim_{n\to\infty} \frac{a_n}{b_n} = 1$
$[z^n]A(z)$	the coefficient a_n of z^n in $A(z) = \sum_n a_n z^n$
2^A	power set of the set A
$[a_0]_\sim$	if $\sim$ is a equivalence relation, the equivalence classes with respect to $\sim$ are denoted by $[a_0]_\sim := \{a \in A \mid a \sim a_0\}$
$\lvert A\rvert$	cardinality of the set A
$\lvert . \rvert_{\mathcal{A}}$	size function of the combinatorial class $\mathcal{A}$
$A \dot\cup B$	disjoint union of sets, i.e. the union $A \cup B$ for sets A, B with $A \cap B = \emptyset$
$f\lvert_{A'}$	restriction of a map $f : A \to B$ to a subset $A' \subseteq A$, i.e. $f\lvert_{A'} : A' \to B$
$f^{-1}(A)$	preimage of the set A under the function f
$F^{\langle -1\rangle}(z)$	compositional inverse, i.e. $F^{\langle -1\rangle}(F(z)) = F\big(F^{\langle -1\rangle}(z)\big) = z$ for $F(z) = \sum_{n\geq 0} f_n z^n$ with $f_0 = 0$ and $f_1 \neq 0$
$\mathcal{A} + \mathcal{B}$	disjoint union of combinatorial classes $\mathcal{A}$ and $\mathcal{B}$, see Section 2.4
$\mathcal{A} \star \mathcal{B}$	labeled product of combinatorial classes $\mathcal{A}$ and $\mathcal{B}$, see Section 2.4
$\mathcal{A} \times \mathcal{B}$	cartesian product of combinatorial classes $\mathcal{A}$ and $\mathcal{B}$, see Section 2.4
$\mathcal{B}$	combinatorial class of rooted binary phylogenetic trees
$\mathcal{B}_n$	set of rooted binary phylogenetic trees with n labels
b_n	number of rooted binary phylogenetic trees with n labels

C	set of character states, usually $C = \{\alpha, \beta, \gamma, \ldots\}$
$\mathrm{ch}(\overline{\chi})$	changing number of a character extension
χ	character
$\overline{\chi}$	character extension
$E(G)$	set of edges of the graph G
$E(\mathcal{T})$	edge set $E(T)$ of $\mathcal{T} = (T, \phi)$
$\mathbb{E}(\xi)$	mean of the random variable ξ
ϕ	label map of a phylogenetic tree, see Definition 2.1 on page 16
$f(A)$	image of the set A under the function f
L	set of leaves of a tree
$l(\mathcal{C}, \mathcal{T})$	parsimony score of a set of characters
$l(\chi, \mathcal{T})$	parsimony score of a character
μ_n	mean $\mathbb{E}(\xi_n) = \mu_n$ of some random variable ξ_n
$\mathbb{N}$	set of natural numbers $\{0, 1, 2, \ldots\}$ including 0
$\mathbb{N}^{>0}$	set of natural numbers $\{1, 2, \ldots\}$ excluding 0
$\mathbb{P}(A)$	probability of the probability event A
$\mathbb{P}(A\|B)$	conditional probability of A given B defined by $\mathbb{P}(A\|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, where $\mathbb{P}(B) > 0$
$P(e)$	transition matrix of the edge $e \in E(T)$
$p(e)$	substitution probability of the edge $e \in E(T)$
p_n	number of planted X-trees with $\|X\| = n$
$p_{n,m}$	number of planted X-trees with $\|X\| = n$ and m vertices excluding ρ, i.e. $m = \|V \setminus \{\rho\}\|$

r	number of character states
$\overline{R}(x, y)$	BGF for rooted X-trees
$\mathbb{R}^+$	positive real numbers $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x > 0\}$
$\overline{r}_{n,m}$	number of rooted X-trees with $\lvert X\rvert = n$ and m vertices
$\overline{r}_n$	number of rooted X-trees with $\lvert X\rvert = n$
$\mathrm{RA}(Y; \mathcal{T})$	reconstruction accuracy
ρ	root of a tree
$S_2(n, k)$	associated Stirling numbers of second kind
$\textsc{MSet}_2(\mathcal{A})$	multisets with 2 elements in a combinatorial class $\mathcal{A}$ of unlabeled objects, see Section 2.4
$\textsc{Set}_2(\mathcal{A})$	sets of size 2 with elements in a combinatorial class $\mathcal{A}$ of labeled objects, see Section 2.4
$\textsc{Set}_{\geq 2}(\mathcal{A})$	sets with at least 2 elements in a combinatorial class $\mathcal{A}$ of labeled objects, see Section 2.4
σ_n^2	variance $\mathbb{V}(\xi_n) = \sigma_n^2$ of some random variable ξ_n
$\mathrm{sym}(T)$	number of symmetry vertices, see Definition 3.39 on page 79
$\mathcal{T}$	phylogenetic tree or X-tree, see Definition 2.1 and Definition 2.2 on pages 16–17
$\mathcal{T} = (T, \phi, p)$	phylognetic tree under the N_r-model, see Defintion 2.12 on page 33
$U(x)$	EGF of the (unrooted) X-trees
$U(x, y)$	BGF of the (unrooted) X-trees
$\mathcal{U}_n$	set of (unrooted) X-trees with $n = \lvert X\rvert$
$u_{n,m}$	number of unrooted X-trees with $\lvert X\rvert = n$ and m vertices
u_n	number of X-trees with $\lvert X\rvert = n$

$V(G)$	vertex set of the graph G
$V(\mathcal{T})$	vertex set $V(T)$ of $\mathcal{T} = (T, \phi)$
$\mathbb{V}(\xi)$	variance of the random variable ξ
X	set of species or label set, usually $X = \{1, 2, 3, \ldots, n\}$, see Definition 2.1 and Definition 2.2 on pages 16–17

Index