Stochastic Tree Models for Macroevolution

Development, Validation and Application

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Diplom Informatikerin Stephanie Keller-Schmidt
geboren am 19. August 1982 in Berlin-Kaulsdorf

Die Annahme der Dissertation wurde empfohlen von

1. Prof. Dr. Kimmo Kaski (Aalto University, Finnland)
2. Prof. Dr. Peter F. Stadler (Universität Leipzig, Deutschland)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 06.09.2012 mit dem Gesamtprädikat cum laude.

Contents

Abstract

Phylogenetic trees capture the relationships between species and can be investigated by morphological and/or molecular data. When focusing on macroevolution, one considers the large-scale history of life with evolutionary changes affecting a single species of the entire clade leading to the enormous diversity of species obtained today. One major problem of biology is the explanation of this biodiversity. Therefore, one may ask which kind of macroevolutionary processes have given rise to observable tree shapes or patterns of species distribution which refers to the appearance of branching orders and time periods. Thus, with an increasing number of known species in the context of phylogenetic studies, testing hypotheses about evolution by analyzing the tree shape of the resulting phylogenetic trees became matter of particular interest. The attention of using those reconstructed phylogenies for studying evolutionary processes increased during the last decades. Many paleontologists (Raup *et al.*, 1973; Gould *et al.*, 1977; Gilinsky and Good, 1989; Nee, 2004) tried to describe such patterns of macroevolution by using models for growing trees. Those models describe stochastic processes to generate phylogenetic trees. Yule (1925) was the first who introduced such a model, the Equal Rate Markov (ERM) model, in the context of biological branching based on a continuous-time, uneven branching process. In the last decades, further dynamical models were proposed (Yule, 1925; Aldous, 1996; Nee, 2006; Rosen, 1978; Ford, 2005; Hernández-García *et al.*, 2010) to address the investigation of tree shapes and hence, capture the rules of macroevolutionary forces. A common model, is the Aldous' Branching (AB) model, which is known for generating trees with a similar structure of "real" trees. To infer those macroevolutionary forces structures, estimated trees are analyzed and compared to simulated trees generatey by models. There are a few drawbacks on recent models such as a missing biological motivation or the generated tree shape does not fit well to one observed in empirical trees.

The central aim of this thesis is the development and study of new biologically motivated approaches which might help to better understand or even discover biological forces which lead to the huge diversity of organisms.

The first approach, called age model, can be defined as a stochastic procedure which de-

scribes the growth of binary trees by an iterative stochastic attachment of leaves, similar to the ERM model. At difference with the latter, the branching rate at each clade is no longer constant, but decreasing in time, i.e., with the age. Thus, species involved in recent speciation events have a tendency to speciate again. The second introduced model, is a branching process which mimics the evolution of species driven by innovations. The process involves a separation of time scales. Rare innovation events trigger rapid cascades of diversification where a feature combines with previously existing features. The model is called innovation model. Three data sets of estimated phylogenetic trees are used to analyze and compare the produced tree shape of the new growth models. A tree shape statistic considering a variety of imbalance measurements is performed. Results show that simulated trees of both growth models fit well to the tree shape observed in real trees. In a further study, a likelihood analysis is performed in order to rank models with respect to their ability to explain observed tree shapes. Results show that the likelihoods of the age model and the AB model are clearly correlated under the trees in the databases when considering small and medium-sized trees with up to 19 leaves. For a data set, representing of phylogenetic trees of protein families, the age model outperforms the AB model. But for another data set, representing phylogenetic trees of species, the AB model performs slightly better. To support this observation a further analysis using larger trees is necessary. But an exact computation of likelihoods for large trees implies a huge computational effort. Therefore, an efficient method for likelihood estimation is proposed and compared to the estimation using a naive sampling strategy. Nevertheless, both models describe the tree generation process in a way which is easy to interpret biologically.

Another interesting field of research in biology is the coevolution between species. This is the interaction of species across groups such that the evolution of a species from one group can be triggered by a species from another group. Most prominent examples are systems of host species and their associated parasites. One problem is the reconciliation of the common history of both groups of species and to predict the associations between ancestral hosts and their parasites. To solve this problem some algorithmic methods have been developed in recent years. But only a few host parasite systems have been analyzed in sufficient detail which makes an evaluation of these methods complex. Within the scope of coevolution, the proposed age model is applied to the generation of cophylogenies to evaluate such host parasite reconciliation methods.

The presented age model as well as the innovation model produce tree shapes which are similar to obtained tree structures of estimated trees. Both models describe an evolutionary dynamics and might provide a further opportunity to infer macroevolutionary processes which lead to the biodiversity which can be obtained today. Furthermore with the application of the age model in the context of coevolution by generating a useful benchmark set of cophylogenies is a first step towards systematic studies on evaluating reconciliation methods.

Für meine Omi Johanne

Acknowledgments

The memorable experience of writing this thesis could not have possible without the help and effort of many people I met along the way.

First of all I want to thank Konstantin Klemm, muchas gracias – Peter F. Stadler, dankescheen – and Martin Middendorf, xiéxie – for your remarkable support, sharing your knowledge, and for giving me the opportunity to work on this thesis.

I thank my office mates Alex, Edith and Wolfi for the great time and helpful hints.

I would also like to thank Petra for helping out with bureaucratic obstacles and Jens for keeping the computers running.

Next, I would like to thank all my colleagues from the bioinf group, especially Berni, Corinna, David, Joe, Lydia, Maribel, Mario, Steve for the welcome distractions. And of course, I would like to thank my colleagues from the pacosy group Kai and Konrad, especially Matthias and Nic for your conscientious proofreading. I really enjoyed the time in both groups.

Good friends, good times and happiness are valuable things in life. Therefore, I would like to thank my family and friends for their love and encouragement.
Thank you, Doreen and Kathrin, for the long-lasting and newly formed friendship, for proofreading and for the conversation among girls. It's always fun with you.
Thank you, Kev, for your showing me the world through different eyes.
Thank you, Chris for one or another extra lesson in math and for being who you are.
Danke, Paps, Mutti und Susi. Ihr habt den Grundstein gelegt und an mich geglaubt.

This work has been supported by the European Commission NEST Pathfinder initiative on Complexity through project EDEN (contract 043251) and by Volkswagen Stiftung through the initiative Complex Networks as a Phenomenon across Disciplines (contract I / 82 719).

"Die Welt ist nicht, wie sie ist, sondern wie wir sind."

Glück kommt selten allein
Eckart von Hirschhausen

List of Abbreviations

AB model	Aldous' Branching model
DNA	deoxyribonucleic acid
ERM model	Equal Rate Markov model
HIV	human immunodeficiency virus
PANDIT	Protein and Associated Nucleotide Domains with Inferred Trees
PDA model	Proportional to Distinguishable Arrangements model
RNA	ribonucleic acid
SIV	simian immunodeficiency virus

CHAPTER 1

Introduction

Evolution in the context of biology refers to changes of species over time and in particular order. Stating the nature as evolution and attempting to classify different forms of life goes back to time of ancient Greeks and Romans, but the biologist and philosopher Carl Linnaeus (1707-1708) was the first who proposed a simple notation for different organisms and introduced a classification by similarities of organisms (Merkl and Waack, 2003). In 1859, Charles Darwin (1809-1882) published *On the Origin of Species by means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* in which he explains the theory of biological evolution based on the mechanism of natural selection. The theory assumes that all present and extinct life on earth is related by one common ancestor. This implies that all organisms are related with another which can be represented in the so-called "Tree of Life" (see Figure 1.1). Until now, one major aim of biology is the classification of organisms and the elucidation of driving forces which lead to their variety.

Today, different kinds of data can be used to investigate the evolutionary history of organisms or genes. This includes morphological characters and molecular data, e.g., nucleotides and amino acids (Lemey *et al.*, 2009). The relationship between various kinds of entities can be depicted as branching tree-like diagrams, also known as *dendrograms*. Those entities can describe e.g., a single species, groups of organisms or genes. Representing taxonomic data or evolutionary relationships among species or organisms, they are also called *phylogenetic trees*. While the leaves of such diagrams represent extant entities, inner nodes stand for ancestral entities (Clewley, 1998; Lemey *et al.*, 2009). There are different types of dendrograms, visualized in Figure 1.2. For instance, a *phylogram* depicts the phylogenetic relationship of entities under consideration of branch length. The latter represents the evolutionary distance which

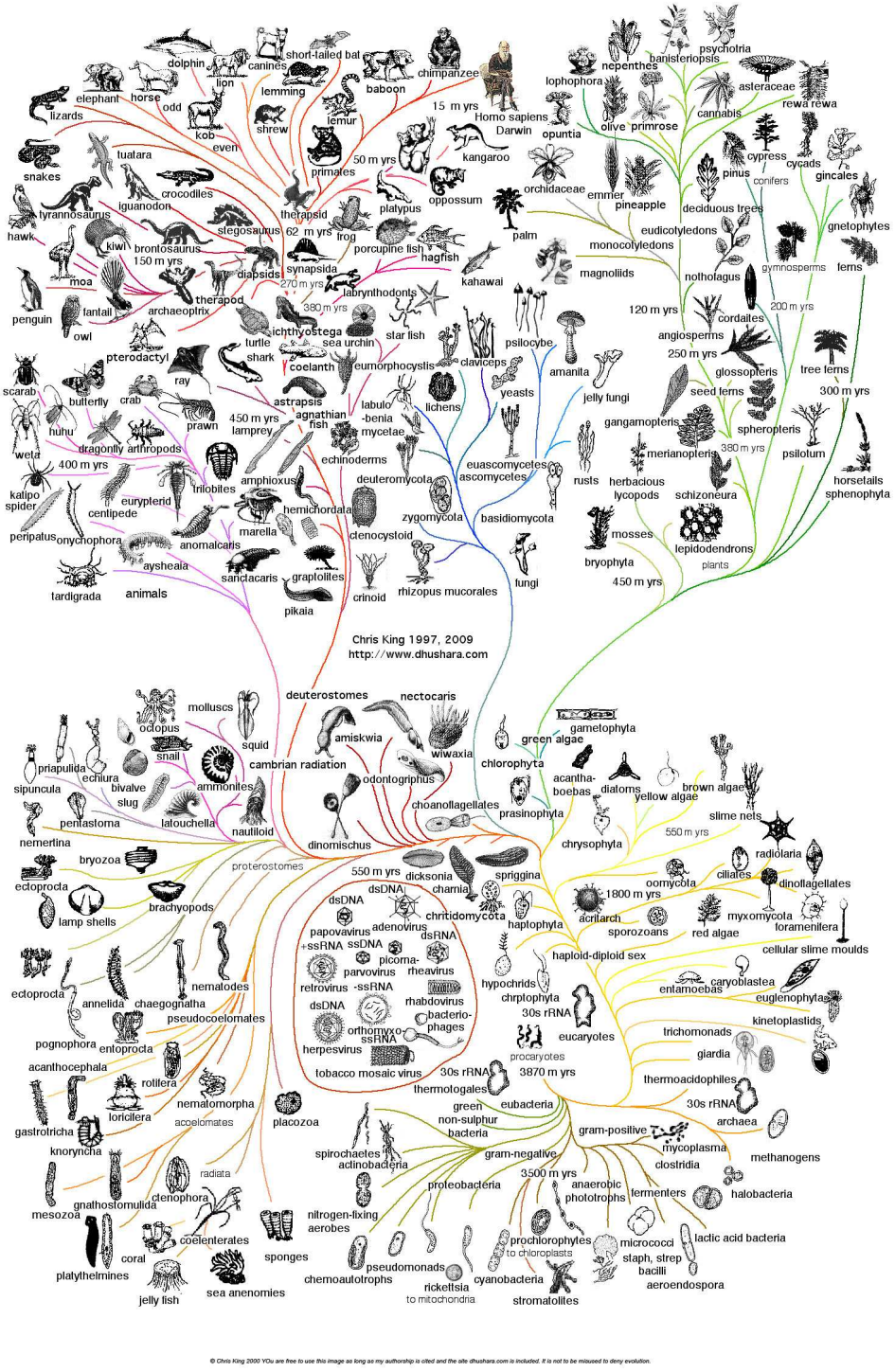


Figure 1.1.: Representation of the "Tree of Life" with five kingdoms. The kingdoms of animals, fungi and plants are depicted in the given order in the upper half. The kingdom of protista and procaryotes, along with archae are represented in the underpart of the tree. Image by Chris King (2009)

can be drawn proportional to the number of evolutionary modifications between two entities (Clewley, 1998). While phylograms consider the branch length, a *cladogram* is assumed to be an estimate of ancestor-descendant relationships where the order of branches is of significance (Pavlopoulos *et al.*, 2010). The third type of a branching diagram is the *phenogram* which depicts the relationship between species or groups of organisms without reconstructing the historical branching process but taking in account the overall similarities (Clewley, 1998).

Explaining the diversity of life is one of the major problems of evolutionary biology. Therefore, one may ask which macroevolutionary processes have given rise to the observed tree shapes, or patterns of species distribution. Within the scope of phylogenetic trees, tree shape refers to the appearance of branching orders and time periods and is also known as, i.e., balance, topology, symmetry, skew, stemminess (Salisbury, 1999; Fiala and Sokal, 1985; Shao, 1990; Kirkpatrick and Slatkin, 1993). Thus, with an increasing number of known species in the context of phylogenetic studies, testing hypotheses about evolution by analyzing the tree shape of the resulting phylogenetic trees became a matter of particular interest. Furthermore the attention of using those reconstructed phylogenies for studying evolutionary processes increased during the last decades (Barraclough and Nee, 2001). An example for patterns of tree shape is the observation of an asymmetry in a tree which is an result of speciations of most of the descendant species by small number of lineages due to an essential adaptation (Felsenstein, 2004). Many paleontologists (Raup *et al.*, 1973; Gould *et al.*, 1977; Gilinsky and Good, 1989; Nee, 2004) tried to describe such patterns of macroevolution by using stochastic models for tree growth. Those models describe a stochastic process to generate phylogenetic trees. The first model was described by Yule (1925) as a continuous time, uneven branching process. In the last decades, further dynamical models were proposed (Yule, 1925; Aldous, 1996; Nee, 2006; Rosen, 1978; Ford, 2005; Hernández-García *et al.*, 2010) to address the investigation of tree shapes and hence, capture the rules of macroevolutionary forces. This is done by analyzing the structures of estimated trees and comparing those to simulated trees of models. Estimated trees can be taken from databases such `TreeBASE` and `PANDIT`. But until now, most of the models can not be explained in a biological sense. Therefore, two new models using an evolutionary dynamics are proposed.

But the evolution of species is not a closed system since species are able to interact. Thus, they may mutually affect their evolution. This can be described by the more complex problem of coevolution or cophylogenetics. Symbiotic relationships between insects and plants respectively between birds and plants or the relationship between predators and prey are just some examples for coevolutionary systems. Here, the focus is on host parasite relationships which are interesting to evolutionary biologists due to the close association between two or more distantly related organisms. The parallel evolution leads to mutual adaptions

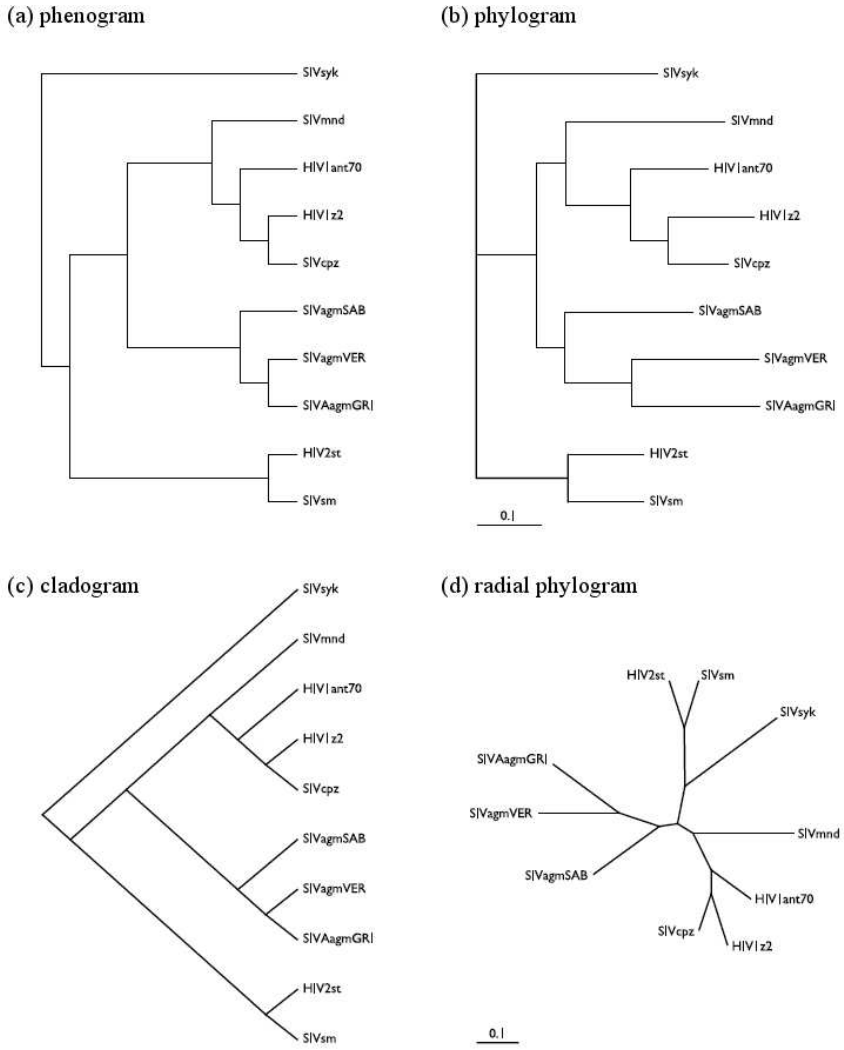


Figure 1.2.: Four types of dendrograms representing the phylogenetic relationships of human and simian immunodeficiency virus (HIV and SIV). The branch length is significant in the cases of both phylograms shown in (b) and (d). the bar ont he bottom of each branching diagram indicated the branch length. For the cladograms depicted in (a) and (c), the relative grouping of the shown entities is of significance. (d) is a radial representations of a phylogram and used for unrooted tree. Illustration modified after Clewley (1998)

in host and parasites and to cospeciation of both groups. Cospeciation is the simultaneous speciation in both lineages which thus gives rates of evolution in two groups of organisms. The additional information can be used in comparative studies (Page and Holmes, 1998). As mentioned earlier, phylogenetic trees for both groups of species can be constructed from sequence data or morphological data. The interactions between the extant species of host and parasite are known empirically. One fundamental problem of coevolution is the inference of the cophylogenetic history when phylogenies of both groups are given. In the recent years, different approaches have been proposed to solve the problem. These algorithms describe a set of events that happened during the coevolution. These events leave their trace on both phylogenies. As far as it is known, no extensively comparative study of those reconciliation methods has been performed. But since not many host parasite systems are available, a benchmark of reasonable test data sets of cophylogenies is necessary. Biologically motivated branching models can be applied to this problem by generating cophylogenies.

1.1. Structure of the Thesis

The central aim of this thesis is the development and study of new biologically motivated approaches which might help to better understand or even discover biological forces which lead to the huge diversity of organisms. This implies the question how speciation emerges and which evolutionary patterns are observable. Therefore, simulated trees generated by models are analyzed in the context of tree shape statistics and compared to real trees. Within the scope of coevolution, one of the porposed models is applied to the generation of cophylogenies to evaluate host parasite reconciliation methods.

At first, a brief introduction to the terminology of graphs and trees, as well as the introduction of appropriate indices used for tree shape statistics is given in Chapter 2. These indices, namely Sackin index, Colless index, and cherry distribution, are quantities for measuring the imbalance of phylogenetic trees.

An overview of methods proposed in the last decade is given in Chapter 3. The chapter starts with the definition of how a tree is generated when using a model. The presented models are divided into different groups. The class of beta-splitting models includes the ERM (Equal Rate Markov) model (Yule, 1925), the birth-death process (Hey, 1992; Nee *et al.*, 1994), the PDA (Proportional to Distinguishable Arrangements) model (Rosen, 1978) and the AB (Aldous' branching) model (Aldous, 1996). The explanation of the ERM model as pure-birth process includes two different approaches of generating a tree. Other models, including Ford's alpha model and the activity model as well as the idea of age-dependent models, are presented in a further section. For each of the models the imbalance of obtained

trees is given and discussed at the end of this chapter.

Chapter 4 presents two novel models of macroevolution. The *age model* is based on the idea that the older a species is the less likely it will speciate. The second model, *innovation model*, assumes the diversification may also be caused by adaptive radiation as a rapid multiplication of species in one lineage after a triggering event. A tree shape statistics is performed for both models in comparison to the common ERM model and AB model. A comparison is also accomplished using three dataset of real trees (`TreeBASE`, `PANDIT`, `McPeek`). It is shown that both models perform at least as well as the AB model and are in good agreement with trees of `TreeBASE` and `PANDIT`.

Until now, the proposed evolutionary models are studied in the context of a tree shape analysis focusing on the tree imbalance. Chapter 5 deals with the question about the quantity of a model. To rank models, a likelihood analysis is employed by asking for the probability of a model of obtaining a given phylogenetic tree. But calculating the likelihood implies a huge computational effort for large trees since each possible order of branching events leading to the tree needs to be considered. A method of resolution is achieved by the development of an efficient sampling method. The exact likelihood is computed for the age model and AB model for small and medium-sized trees of real trees (`TreeBASE`, `PANDIT` and `McPeek`). Results show that the age model performs similar as the AB model for `TreeBASE` data and better for `PANDIT` data. The AB model outperforms the age model for `McPeek` data, but the small data set may lead to ambiguous conclusion.

But not only the understanding of species' evolution but also the problem of coevolution achieves more interest in the last decades. Latter one can be described by the ability of species to interact among one another. Hence, they may mutually affect their evolution. Different host parasite reconciliation methods have been proposed recently. But only a few host parasite systems have been analyzed in sufficient detail. Thus in Chapter 6 one of the proposed models is employed to generate meaningful test data sets to tackle the lack of benchmarks. Those data sets are used for an evaluation of host parasite reconciliation methods. As far as known, this chapter describes an initial contribution to extensively compare methods for cophylogeny reconciliation.

1.2. Manuscripts

A part of the results presented in this thesis have been included in the following manuscripts.

Chapter 4, introducing the new growth models, is based on the following articles:

- Keller-Schmidt S, Tuğrul M, Eguíluz VM, Hernández-García E, Klemm K (2011). **An Age Dependent Branching Model for Macroevolution.** Submitted, *http://arxiv.org/abs/1012.3298*.
- Keller-Schmidt S, Klemm K (2011) **A model of macroevolution as a branching process based on innovations.** Accepted for publication in *Advances in Complex Systems*, *http://arxiv.org/abs/1111.2608*

Chapter 5, discussing the likelihood analysis and presenting a new method for likelihood estimation, is partly based on the following article:

- Keller-Schmidt S, Tuğrul M, Eguíluz VM, Hernández-García E, Klemm K (2011). **An Age Dependent Branching Model for Macroevolution.** Submitted, *http://arxiv.org/abs/1012.3298*.

The application of the age model to Coevolution is presented in Chapter 6. It is based on the following article:

- Keller-Schmidt S, Wieseke N, Klemm K, Middendorf M (2011). **Evaluation of Host Parasite Reconciliation Methods using a new Approach for Cophylogeny Generation.** Submitted.

CHAPTER 2

From Graphs to Trees and Tree Shape Statistics: Definitions and Examples

Every extant and extinct species has one common ancestor. Hence, all species are somehow related. This relation between species can be depicted in the "Tree of Life" whose reconstruction plays a significant role in the research field of evolutionary biology. Such a tree is formed by various types of evolutionary forces which need to be explored. This chapter deals with the fundamentals of phylogenetic trees in a mathematical and biological view. This includes the definition and explanations of terminologies as well as methods for measuring the tree shape.

2.1. Graphs and Trees

Trees are a special kind of graphs. Graphs are a widely used tool in the field of bioinformatic, e.g., for modeling metabolic and regulatory networks, as well as for pattern matching or for the generation and depiction of phylogenetic trees. One can differ between directed and undirected graphs.

Definition 2.1 (directed graph)**.** A *directed graph* $G = (V, E)$ consists of a set of nodes $V = \{v_1, \ldots, v_m\}$ and a set of edges $E \subseteq V \times V$, which connects nodes (Diestel, 2006).

- An edge $e = (v_i, v_j)$ from node v_i to node v_j with $i, j \in [1; m]$ is called *directed edge*, also $v_i \xrightarrow{e} v_j$. Therefore, v_j is a direct successor of v_i and v_i is direct predecessor. The edge is called *loop*, if $v_i = v_j$.
- One differ between an indegree and an outdegree of a node. An *indegree* of $I(v_i)$ of a node v_i is the number of its direct predecessors, $I(v_i) = |\{v_j | (v_j, v_i) \in E\}|$. An *outdegree*

$O(v_i)$ of a node v_i is the number of its direct successors, $O(v_i) = |\{v_j|(v_i, v_j) \in E\}|$. Each node v_i is called *isolated*, if $O(v_i) = I(v_i) = 0$.

- A *path* of a graph is a sequence of nodes $v_1, v_2, \ldots, v_n$, where v_i and v_{i+1} are connected by an edge for each $i = 1, \ldots, n-1$. The length of a path is the count of edges along the path. A *simple* path has no repeating nodes.
- A *cycle* is a path where start and end node are identical. In a *simple cycle* every edge (v_i, v_j) is used once. A graph without cycles is called *acyclic*.
- A graph $G = (V, E)$ is called *connected*, if every pair of nodes can be connected by at least one path.

Definition 2.2 (undirected graph)**.** A *undirected graph* can be understood as a special kind of a directed graph but without relevance of the edge orientation (Diestel, 2006) but:

- An edge is defined by $\{(v_i, v_j), (v_j, v_i)\}$ with $(v_i, v_j) \in E \leftrightarrow (v_j, v_i) \in E$. Since the order of nodes is not considered, an edge can also be expressed as a pair of nodes $\{v_i, v_j\}$.
- Two nodes connected by an edge are called *adjacent*.
- The degree of node v_i is defined by the number of its adjacent nodes.

A tree is a special kind of a graph. Thus, it is defined as follows by using previous definitions.

Definition 2.3 (tree)**.** A *tree* is defined as an acyclic connected graph $G = (V, E)$. Each node can have a number of children nodes (descendants) and at most one parental node (ancestor). Following conditions must be statisfied (Semple and Steel, 2003; Diestel, 2006):

- There exists exactly one path between every pair of nodes in G.
- G is minimal connected, i.e., removing an edge $e \in E$ results in a not connected graph.
- G is connected and $|E| = |V| - 1$.
- G is acyclic and $|E| = |V| - 1$.
- G is maximal acyclic i.e., adding an arbitrary edge e to E results in a graph containing a cycle.

Definition 2.4 (rooted, unrooted)**.** A tree can be rooted or unrooted. A *rooted* tree has one node, called *root*, from which all other node descend. An *unrooted* tree has not such a root node. The number of binary, rooted trees with n leaves can be computed by

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} = (2n-3)!! \; . \tag{2.1}$$

As one can imagine, the number of possible trees topologies grows exponentially with increasing n.

Definition 2.5 (tree topology, shape)**.** A *topology* of a tree is defined by the unlabeled topological branching pattern without information about time (Himmelmann and Metzler, 2007). It is also referred as *shape*.

Definition 2.6 (leaf, inner node, root, cherry)**.** A tree contains different types of nodes which are depicted in Figure 2.2 (p.13). A node is called *leaf* if it is a terminal node. Each terminal node holds a degree of one. All other nodes are called *inner nodes* with a degree not less than two (Mount, 2004; Diestel, 2006). For a rooted tree, one node is signed as *root* and contains no predecessor. Two leaves which are adjacent to a common node are called a *cherry* (McKenzie and Steel, 2000).

2.2. The Link between Phylogenetics and Macroevolution

The evolutionary theory proposed by Darwin says that all existing organisms descend from one common ancestor. The development of new species caused by branching processes out of existing populations makes it possible to depict the evolution of all organisms in an ordered tree (Merkl and Waack, 2003). Figure 2.1 shows a sketch of Darwin's idea.

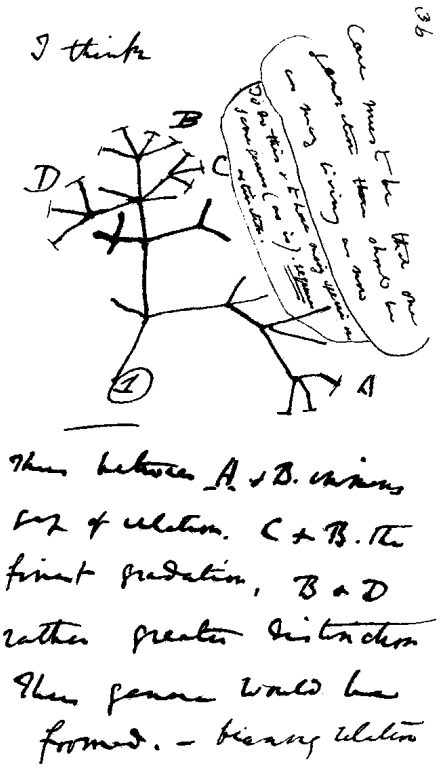


Figure 2.1.: Sketch of an evolutionary tree by Charles Darwin from his first notebook on "Transmutation of Species" (1837).

2.2.1. Phylogenetic Trees

Nowadays a huge amount of organism is identified. The *Tree of Life* is an attempt to arrange all such organism in a phylogenetic tree and captures the relationship between organisms (Mount, 2004; Lemey *et al.*, 2009). These are trees where leaves represent extant species, alive today, and inner nodes stand for ancestral species from which the extant species have descended. The latter one can be hypothetical. The root allegorises the ancestor of every organism. But for extinct and hypothetical ancestors no biological data is available. Thus phylogenetic trees are built from extant species (Page and Holmes, 1998; Semple and Steel, 2003). Therefore, various types of data can be used such as morphological data or data gained from sequence alignment of DNA, RNA or proteins (Merkl and Waack, 2003).

In the following only rooted, binary phylogenetic trees are considered.

Definition 2.7 (phylogenetic tree)**.** A rooted phylogenetic tree T for a set $S = s_1, \ldots, s_n$ of n species is a strict binary tree with n leaves and following conditions.

- It is a tree with exactly one node, called *root*, with degree two or zero.
- Each of the *inner nodes* has exactly two children and thus a degree of three.
- Each *leaf* has degree of one and is labeled with exactly one species $s \in S$.
- Each species exists exactly one time in the set of leaves S.

Definition 2.8 (subtree)**.** Given a phylogenetic tree T with a root w, a subtree T' is obtained as the component not containing w after cutting an edge $\{i, j\}$ of T. T' is again a rooted strict binary tree.

Since the considered trees are rooted, the direction in the tree from the root to all other descending nodes can correspond to evolutionary time (Page and Holmes, 1998). Given the evolutionary time and two nodes i, j connected by a path which starts at the tree root, one can draw conclusion about their relationship of ancestor and descendant. If the node i is closer to the root, i is the ancestor of j and vice versa. An *unrooted* tree has no root and hence, one can not infer any relationship in the sense of descendant and ancestor (Page and Holmes, 1998). With the focus on the tree shape in the following chapters, various characteristics such as the distance, height, and depth of trees can be evaluated.

Definition 2.9 (distance, height, depth)**.** The *distance* between two nodes i and j on a tree T is the number of edges contained in the unique path between i and j. The *height of a tree* is the maximal distance from the root w to a leaf. The *height of a node* i is defined by the height of the subtree rooted at i. The *depth* of a node i is the length of the path to its root (i.e., its root path).

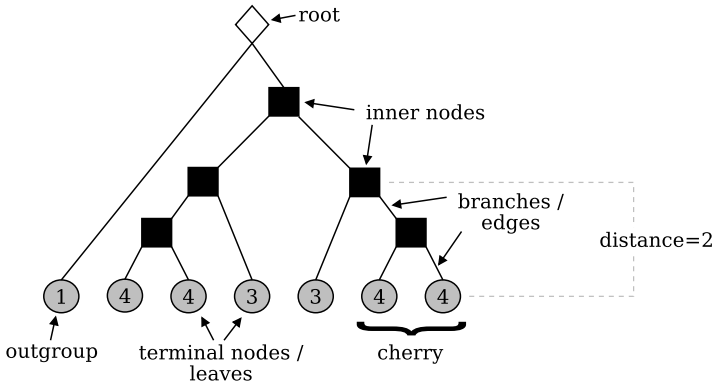


Figure 2.2.: Terminology of phylogenetic trees with seven leaves. The height of the tree is four. The numbers in the circles is the depth for each leave.

Usually, the edge length of phylogenetic trees can be interpreted as time estimations. Thus periods of evolution or counts of morphological and molecular differences between two nodes can be demonstrated. A tree where all leaves have an equal distance to the root, displays only the relative relations between species, but no evolutionary changes. This tree is called *cladogram* whereas *phylograms* feature branch length which can represent evolutionary time or changes (Page and Holmes, 1998). In the following, it is assumed that all edges have a unit length. Thus, only cladograms are considered.

A summary on the terminology of a phylogenetic tree is visualized in Figure 2.2.

2.2.2. Evolutionary History in the Sense of Macroevolution

The idea of biological evolution is based on the assumption that all species descend from one common ancestor which is the origin of the large diversity. It comes along with changes in living organisms over time and occurs at every level of biological classification. This includes biomolecules such as DNA and proteins, an individual species or a population of species (Erwin, 2000; Hall *et al.*, 2008). The research field of biological evolution can be divided into two subfields: microevolution and macroevolution. Both terms were introduced in 1927 by the Russian Entomologist Filipchenko (1927) in German and 1937 translated into English by Dobzhansky (1951).

Microevolution encompasses the small-scale history of life which happens in a short period of time. It refers to changes at the molecular level within a population of species or an individual species. For phylogenetic trees one can say that the focus lies on one branch only. This changes affect the allele frequencies and thus microevolutionary patterns can be observed in the phenotype of organisms. Microevolutionary changes can be caused by, e.g., mutation, gene flow, genetic drift or natural selection (Reznick and Ricklefs, 2009; Kimura, 1983; Page and Holmes, 1998).

In contrast to microevolution, *macroevolution* defines changes at the level of species or

above, i.e., phyla and genera (Ayala and Fitch, 1997). It considers the large-scale history of life and gives rise of the diversity of an entire clade and its stability instead of a single species. Macroevolutionary patterns occurring in a long-term development are caused by dynamic processes such as character changes in lineages, speciation and extinction. Also evolutionary mechanisms defined by microevolution (mutation, gene flow, genetic drift and natural selection) can help biologists to find an explanation of macroevolutionary patterns in the Tree of Life, under the condition that a long-term observation is given. Though micro- and macroevolution can merge seamlessly if microevolution continues and a population educes a new species which is not able to reproduce organisms of that population. This is the reason why some scientists understand macroevolution as a large amount of microevolutionary processes over a long time scale since the mechanisms are identical (Erwin, 2000). This controversy is not further discussed here.

Recent studies (Barraclough and Nee, 2001; Reznick and Ricklefs, 2009; Ricklefs, 2007) discuss the meaning of the phylogenetic tree structure when explaining macroevolutionary processes. In a biological point of view the tree shape might give an answer to the question how the diversity of life has been arisen. Methods for measuring the tree shape are considered in Section 2.3.

2.3. Tree Shape and Appropriate Methods of Measurement

In some contexts, shapes of trees are a result of optimization: prominent examples include minimum spanning trees (Bang and Kun-Mao, 2004) of weighted graphs, self-balancing search trees (Pfaff, 2004) like AVL trees (Bouge *et al.*, 1995; Nievergelt, 1974), the red-black-trees (Guibas and Sedgewick, 1978) and trees of branching blood vessels optimized for a large flux (West *et al.*, 1997). For other types of trees, however, shapes may not be selected by optimization or at least the underlying optimization principle is not known. Then one may ask what *dynamical* branching rules (Harris, 1963) govern the observed tree shapes. Phylogenetic trees are such cases with large datasets available and little knowledge about the mechanisms shaping these structures. Thus the shape of phylogenetic trees is used to test hypotheses about the evolution and corresponding macroevolutionary processes and may give hints on how the biological diversity has arisen. Different methods to study the tree balance have been proposed and applied to simulated and empirical trees in the last decades (Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997; Agapow and Purvis, 2002; Savage, 1983; Matsen, 2007).

2.3.1. Tree Imbalance

Analyzing the tree shape one can focus on the tree balance which refers to a topological structure. It can be described as the degree to which daughter subtrees of internal nodes are of

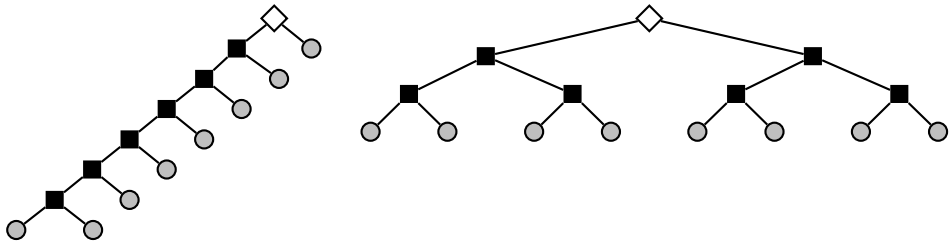


Figure 2.3.: Comparison of tree shapes concerning tree balance. Imbalance refers to an uneven distribution of the number of leave between left and right branches of a tree or subtree. In this case each tree of size eight consists of a root (white diamond), a set of inner nodes (black squares) and a set of leaves (gray circles). The left tree is totally imbalanced, also called comb tree, with Sackin index $d = 35/8 = 4.375$ and Colless index $c = 21/21 = 1$. The right tree is a complete binary tree with Sackin index $d = 24/8 = 3$ and Colless index $c = 0/21 = 0$.

similar or different size (Matsen, 2007). The *size* of a tree refers to the number of leaves. Note that tree balance disregards branch length and does not regard labels (Mooers and Heard, 1997). The balance of a tree is influenced by variation of the speciation or extinction rate (Kirkpatrick and Slatkin, 1993). Systematic deviations between shapes of small phylogenetic trees and trees from uncorrelated stochastic processes of speciation and extinction have been known for decades (Gould *et al.*, 1977). Recent studies (Blum and François, 2006; Herrada *et al.*, 2008) provide a quantitative and exhaustive analysis of tree shapes in large databases (Sanderson *et al.*, 1994; Whelan *et al.*, 2006). The studies by Blum and François (2005) of the phylogenetic imbalance based on trees of the database TreeBASE show that the tree shape undergoes a rapid change from the smaller to the intermediate-sized and larger trees (Aldous, 1996).

The deviation from an evolutionary null model (see Section 3 (p.19)) can be pinned down to an increased imbalance of the phylogenetic trees (Matsen, 2007; Mooers and Heard, 1997; Agapow and Purvis, 2002), a tendency to unevenly split the set of leaves between the left and right subtrees. Figure 2.3 depicts examples for a completely imbalanced and completely balanced tree.

Several indices for balance measurement have been proposed and compared in the literature (see Mooers and Heard (1997); Matsen (2006); Agapow and Purvis (2002); Kirkpatrick and Slatkin (1993) for detailed discussion). The following subsection explains three of these indices, namely the Sackin index (Sackin, 1972) as depth index, the Colless index (Colless, 1982) as metric for tree balance and the cherry distribution (McKenzie and Steel, 2000) as measurement of tree quality. All three metrics do not consider the branch length. Kirkpatrick and Slatkin (1993) and Matsen (2006) concluded that the Sackin index and the Colless index

were the most powerful statistics. Furthermore the study by Matsen (2006) stated the cherry distribution as an appropriate second statistic. Similar conclusion was achieved by Agapow and Purvis (2002).

2.3.2. The Sackin Index

The *Sackin index* (Sackin, 1972) d is the average distance of leaves from root, in the following also called *depth* of a tree

$$d = \frac{\sum_{i=1}^{n} d_i}{n} \ . \tag{2.2}$$

The depth of a tree is commonly needed in the manipulation of the various self balancing trees, AVL trees (Bouge *et al.*, 1995) in particular. Conventionally, the value -1 corresponds to a subtree with no nodes, whereas zero corresponds to a subtree with one node. Here the depth is considered as a measure of imbalance. In a tree with n leaves, d_i denotes the number of edges to be traversed to reach the root from node $i \in \{1, \ldots, n\}$. This measure may be applied to non-binary trees, including polytomies and monotomies.

For a complete binary tree, $d = \log_2 n$ since all $n = 2^k$ leaves are at level k. As the other extreme, a comb (or pectinate) tree has $d = 1 + 2 + \cdots + (n-2) + 2(n-1)$ resulting in asymptotically linear scaling $d \sim n$.

2.3.3. The Colless Index

The *Colless index* (Colless, 1982) c measures the average imbalance of a tree. The imbalance at an *inner node* j of the tree is the absolute difference $c_j = |l_j - r_j|$ of leaves in the left and right subtree rooted at j, denoted by l_j and r_j. Then the average of imbalance can be computed by

$$c = \frac{2}{(n-1)(n-2)} \sum_{j=1}^{n-1} c_j \tag{2.3}$$

with an appropriate normalization. The index j runs over all $n-1$ inner nodes including the root itself. One can easily conclude that $c = 0$ for a complete binary, totally balanced, tree and $c = 1$ for a comb tree, see Figure 2.3.

2.3.4. The Cherry Distribution

Another statistic for tree shape is the *distribution of cherries*. It is an easy computed statistic where the number of pairs of leaves which are adjacent to a common ancestor is calculated McKenzie and Steel (2000). Studies by McKenzie and Steel (2000) and Matsen (2007) show that the distribution of cherries is asymptotically normal under two common null models for generating phylogenetic trees.

2.4. Data Sets of Empirical Phylogenetic Trees

The analysis and comparison of macroevolutionary models (introduced in Chapter 3 (p.19) and 4 (p.31)) is based on three different data sets. The data was preprocessed since some of the trees contain

- *outgroups*: a single node, which is the most distant related one of the root; essential for rooting a tree (Gregory, 2008).
- *monotomies*: a node in a tree which has only one descending branch.
- *polytomies*: a node in a tree which has more than two descending branches.

The preprocessing was done as in studies by Blum and François (2005) including the removal of outgroups by deleting the leaves or cherries branching off of the root. Since not all empirical trees are binary trees, polytomies (multifurcating nodes) and monotomies were solved in a random manner by splitting them based on the ERM model (Blum and François, 2005; Matsen, 2006). In addition, trees with less than four leaves are without meaning for the analysis and hence, were excluded from the used data sets.

Data set of database TreeBASE: `TreeBASE` (Sanderson *et al.*, 1994) is the main phylogenetic database containing phylogenetic trees of species and populations. The data from `TreeBASE` has been downloaded from `http://www.treebase.org` in June, 2007 containing 5,212 phylogenetic trees. After preprocessing the data set contained 5,087 trees of size 4 to 535 .

Data Set of database PANDIT: The database `PANDIT` (Whelan *et al.*, 2006) contains phylogenetic trees representing the evolution of protein families. `PANDIT` has been downloaded from `http://www.ebi.ac.uk/goldman-srv/pandit` in May 2008 and contains 7,738 protein families respectively 46,428 phylogenetic trees. Preprocessing results in a data set containing 36,136 trees of size 4 to 2,562 .

McPeek Data Set: The `McPeek` data set is assembled by McPeek and Brown (2007) and includes 245 species-level molecular phylogenies of 245 clades of animals and plants, namely chordate, arthropod, mollusk, and magnoliophyte. One part of trees are fossil-based estimations while another part is based on molecular phylogenies. One tree (ID: Glor_et_al_2003) was removed since the tree in newick format is missing. For further details on the selected phylogenies it is referred to the work by McPeek and Brown (2007) and McPeek (2008).

An overview of the empirical data sets is given in table 2.1

	TreeBASE	PANDIT	McPeek
leaves representing	species	proteins	species
number of trees	5,212	46,428	245
amount of leaves	3 . . . 535	2 . . . 5,121	4 . . . 116
number of tree after preprocessing	5,087	36,136	244
amount of leaves after preprocessing	4 . . . 535	4 . . . 2,562	4 . . . 116

Table 2.1.: Overview of empirical data sets.

CHAPTER 3

Stochastic Models of Macroevolution

Classifying organisms by their similarities is of great interest since scientists are aware of their diversity. The first attempt to display the diversity in a tree of life, based on the idea that there is one common ancestor for all organism, was published in "The Origin of Species" by Darwin (1859). With an increasing number of species in phylogenetic studies, the asymmetrical branching of different groups of organisms was noticed (Stich and Manrubia, 2009; Willis and Yule, 1922). One reason of the tree asymmetry might be the branching of rare clades which give rise to a large number of descendant species due to an important adaption (Felsenstein, 2004).

For describing and understanding patterns of biological diversity or macroevolution, stochastical models have been used by many paleontologists (Raup *et al.*, 1973; Gould *et al.*, 1977; Gilinsky and Good, 1989; Nee, 2004). But to infer about macroevolution, phylogenetic information is necessary since one is interested in the history of clades (Nee, 2006). With an expanding number of molecular data (Hey, 1992; Nee *et al.*, 1992), the interest in stochastical models increases (Nee, 2006). The continuous-time, uneven branching process was first described as stochastic process for modeling phylogenies by Yule (1925). Since that time stochastical models have been used to address several distinct questions and purposes concerning

- the estimation and comparison of the diversification rate of clades
- the investigation of clade shapes
- the estimation of speciation and extinction rates from fossil data which is only resolved to a certain level in the taxonomic rank, e.g., genus

- the deduction about past speciation and extinction rates from correct phylogenies of extant species
- the reconstruction of phylogenies from molecular data
- the usage as null model when trying to assign data with a biological significance

(Nee, 2004, 2006; Aldous *et al.*, 2011) to name but a few.

When considering the issue on understanding driving forces of evolution that have led to the diversity of living organisms, the reconstruction of phylogenies plays an important role. Several models treat speciation and extinctions as random process (Aldous *et al.*, 2008). Hypotheses about those dynamical rules governing all evolutionary processes may come under scrutiny with large collections of phylogenetic trees available nowadays (Sanderson *et al.*, 1994; Whelan *et al.*, 2006). A suitable starting point and null hypothesis is the ERM (Equal Rate Markov) process suggesting that species undergo further speciation at a constant homogeneous rate, independently of previous events and other species present. The introduction and explanation of the ERM model as well as other common models, such as the PDA (Proportional to Distinguishable Arrangements) model and the AB (Aldous' branching) model, is the main part of this chapter. Beginning with a brief explanation on the basic branching process of generating a phylogentic tree, the imbalance of trees generated with different models, growth models in particular, is discussed as well.

3.1. Tree Generation

From the evolutionary dynamics, an evolving phylogenetic tree $T(t)$ is obtained within the following formal framework. At each time step t, the leaves of $T(t)$ are the species $S(t)$. A species $s \in S(t)$ is choosen according to a probability distribution $\pi(s,t)$ on $S(t)$ and undergoes *speciation*. This is, two new leaves s' and s'' attach to a leaf s such that

$$S(t+1) = S(t) \setminus \{s\} \cup \{s', s''\} \tag{3.1}$$

is the set of species at time $t+1$. After this event, s is an inner node and no longer a leaf of the tree. The initial condition at $t = 1$ is a single species. Therefore, the discrete time t and the number of species n are identical, $n = |S(t)| = t$. In this way, each model of speciation dynamics also defines a model for the growth of a binary tree by iterative splitting of leaves.

Abstracting from the dynamics behind tree generation, one may formulate a model directly in terms of a probability distribution on a set of trees. More precisely, a probability distribution is given separately for each set of all eligible trees of the same given tree size n. Here eligible trees are oriented binary rooted trees. *Oriented* is to say that left and right subtrees are explicitly distinguishable by a left-right labeling. By this choice the isomorphy classes with respect to left-right symmetry can be disregarded.

In a particular class of models, including the ERM model and AB model described in the next sections, the probability $L(T)$ of a tree T is defined by a product over its inner nodes $I = \{1, 2, \ldots, n-1\}$ according to

$$L(T) = \prod_{j \in I} p_{\text{model}}(i_j | n_j) \ . \tag{3.2}$$

The model-specific probability factor p_{model}, to determine the next node for speciation, depends on the total number n_j of leaves in the subtree with root node j and the number i_j of leaves in the left subtree of j. Arguments naturally fulfill $1 < i_j < n_j$. Left-right symmetry is ensured by

$$p_{\text{model}}(i|n) = p_{\text{model}}(n-i|n) \tag{3.3}$$

such that $L(T_1) = L(T_2)$ when T_1 is isomorphic to T_2. The choice of the functional form of p_{model} determines the expected balance of the trees. By concentrating probability mass at values i_j close to two and close to $n-1$, imbalance is enforced.

3.2. Beta-Splitting Models

The so-called beta-splitting models (Aldous, 1996) belong to a one-parametric class of models for stochastic tree generation with expected imbalance tunable by a parameter $\beta \in [-3/2, +\infty[$. Beta-splitting models define a distribution of trees by the probability, depending on the total number n of leaves in a rooted subtree with i numbers of leaves in its left subtree,

$$p_\beta(i|n) = \frac{1}{a_\beta(n)} \frac{\Gamma(\beta + i + 1)\Gamma(\beta + n - i + 1)}{\Gamma(i+1)\Gamma(n-i+1)} \tag{3.4}$$

with appropriate normalization factor $a_\beta(n)$ (see Aldous (1996)) and $\Gamma(x)$ as Gamma function (Abramowitz and Stegun, 1972). Choosing $\beta \to \infty$ produces complete balanced trees whereas $\beta < 0$ corresponds to more unbalanced trees such as trees generated by the ERM model.

The first subsection discusses the well known ERM model as a simple pure-birth process. The birth-death process is considered afterwards. Furthermore, the PDA model as well as the AB model and activity model are considered. Growth models which do not belong to the class of beta-splitting models are pointed out in the last Section 3.3.

3.2.1. Equal Rate Markov Model as Pure-Birth Process

The *Equal Rate Markov (ERM) model* is assigned to the scientific work of Harding (1971) and Cavalli-Sforza and Edwards (1967). Nevertheless it is also called *Yule model* because it is based on models of diversification process which were proposed by Yule (1925). The model is considered as the earliest mathematical model of evolutionary branching. It is well-known

and often used as null hypothesis for phylogenetic tree shape respectively for evolutionary dynamics.

The ERM model is based on the idea that species undergo speciation at a constant homogeneous rate, independently of previous events and other present species. In the process, the probability of choosing a species is uniform at each time step, $\pi(s,t) = 1/t$. It is a pure birth process since the probability for extinction is zero. Additionally, processes that share a similar probability distribution of topologies as the ERM model have been investigated by Moran (1958) and Hey (1992). Simberloff (1987) and Simberloff *et al.* (1981) studied the application of the model to statistical testing of area cladograms (Slowinski, 1990). The model is a particular case of β-splitting with $\beta = 0$ (Blum and François, 2006).

There are two options of growing trees under the ERM model with different probability distributions, both are depicted in Figure 3.1 and explained in the following. In the first case, and as pointed out previously, each species has an uniform probability to split. The algorithm of generating a tree of size n is given in Algorithm 1.

Algorithm 1: Standard ERM model for tree generation.

Input: root, number of nodes n of tree T
Output: tree T
while *n in T is not reached* **do**
    Choose leaf s from all current leaves S in T at random;
    Replace s by a cherry with $s_\prime$ and $s_{\prime,\prime}$ as descendants;

In the second case, called *modificated ERM model* in the following, starts with initializing a root node which is labeld with the target tree size n. Not only the root, but each leaf s is assigned with the number of leaves in the subtree with root s. Since the final tree size n of T must be known in initial conditions, this variant of the ERM model is not a model of open-ended evolution, since no adaptions in tree size while generating the trees are possible. The recursion of generating the tree is given in Algorithm 2.

The ERM model correspond to a particular simple probability distribution on the set of generated trees, as pointed out in the previous part describing the modificated ERM model. For a tree with $n \geq 2$ leaves generated by the ERM model and $i \in \{1, 2 \ldots, n-1\}$, let $p_{\mathrm{ERM}}(i|n)$ be the probability that exactly i leaves are in the left subtree of the root. Then $p_{\mathrm{ERM}}(i|n) = 1/(n-1)$. This is shown inductively as follows.

Proof. Obtaining exactly i leaves at step n, either they were already present at the previous step and the speciation took place in the right subtree, or the number increased from $i-1$ to i by speciation in the left subtree. Addition of these products of probabilities for the two cases yields

$$p_{\mathrm{ERM}}(i|n) = \frac{n-1-i}{n-1} p_{\mathrm{ERM}}(i|n-1) + \frac{i-1}{n-1} p_{\mathrm{ERM}}(i-1|n-1) \ . \tag{3.5}$$

With the induction hypothesis $p_{\mathrm{ERM}}(j|n-1) = 1/(n-2)$ for all j, one can obtain

$$\begin{aligned} p_{\mathrm{ERM}}(i|n) &= \left[\frac{1}{n-1} \cdot \frac{1}{n-2} \cdot (n-1-i)\right] + \left[\frac{1}{n-1} \cdot \frac{1}{n-2} \cdot (i-1)\right] \\ &= \frac{1}{n-1} \cdot \frac{1}{n-2} \cdot [n-1-i+i-1] \\ &= \frac{1}{n-1} \cdot \frac{1}{n-2} \cdot [n-2] \\ &= \frac{1}{n-1} \ . \end{aligned}$$

□

The induction starts with $p_{\mathrm{ERM}}(1|2) = 1$ which holds because a tree with two leaves has one leaf each in the left and in the right subtree. Thus the uniform selection of species turns into a uniform distribution on the number of nodes in the left or right subtree. Note that the same distribution applies to each subtree of an ERM model generated tree. Therefore, p_{ERM} fully describes the statistical ensemble of ERM trees. The probability of obtaining a particular tree is the product of p_{ERM} terms taken over all subtrees. This becomes particularly relevant for modifications of the model taking p non-uniform, as shown in the second case of the ERM model.

Algorithm 2: Modificated ERM model analogous to beta-splitting with predefined tree size n.

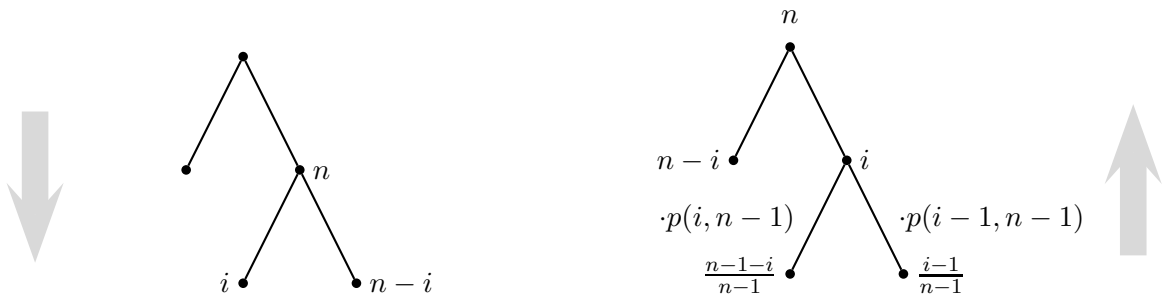
Input: root labeled with target number of leaves n
Output: tree T

while $\exists$ *leaf* s *with label* $l > 1$ **do**
  Choose leaf s of already generated tree randomly;
  Replace s by a cherry with s' and s'' as descendants;
  Assign new leaves with labels i and $l-i$ whereas i is drawn from flat distribution on $\{1, \ldots, l-1\}$

$$p_{\mathrm{ERM}}(i|l) = \frac{1}{l-1}$$

  ;

The topologies generated with the ERM model tend to be compact and nearly *balanced* tree shapes, regardless of the version of the growing tree. When comparing with the shape of observed trees of a certain moderate size, however, the ERM hypothesis can be rejected, as most real phylogenetic trees are significantly less balanced than those generated by the ERM model (Herrada *et al.*, 2008).



$$p(i|n) = \frac{1}{n-1}$$

$$p(i|n) = \frac{n-1-i}{n-1} \cdot p(i, n-1) + \frac{i-1}{n-1} \cdot p(i-1, n-1)$$

(a) Standard ERM model: Calculating the probability distribution in a top-down manner.

(b) Modificated ERM: Calculation of probability of a node being the parent node of two nodes in a bottom-up manner.

Figure 3.1.: Two cases for tree generation under the ERM model in a top-down and bottom-up approach. n is the total number of leaves in the corresponding node and i describes the number of leaves of one of its child nodes. $p(i|n)$ is the probability of node to obtain i given n leaves from its parent node.

3.2.2. The Birth-Death Process

On models based on a *birth-death process* each species has a probability to die with a constant rate r_d, additionally to the constant speciation rate r_b. The death process is called extinction. Assuming that the number of species grows exponentially in time, the probability of a lineage lasting from birth to a later instant of time t is given by

$$p(t) = \frac{1 - \frac{r_d}{r_b}}{1 - \frac{r_d}{r_b} e^{-(r_b - r_d)t}}$$

in which $\frac{r_d}{r_b}$ regulates the tree growth such that it is different to the one from a pure-birth process (Nee, 2006). With the focus on macroevolution, the birth-death model allows to estimate speciation and extinction rates from molecular phylogenies with missing information from extinct species (Nee, 2006). For more details and results on the birth-death process with varying birth and death rates it is referred to Kendall (1948); Hey (1992); Nee *et al.* (1994); Nee (2006).

3.2.3. The Proportional to Distinguishable Arrangements Model

The model of *Proportional to Distinguishable Arrangements* (PDA) was first described by Rosen (1978) in the field of cladograms. As the ERM model the PDA model is a special case of the beta-splitting model but with an observable $\beta = -1.5$. It is no explicit model of growing trees. In fact, each tree topology of n labeled leaves has the same probability.

Within the set of all possible arrangements of n species, the frequency of each topology is proportional to the number of distinguishable trees sharing that topology (Mooers and Heard, 1997). For instance, considering a tree of four leaves, with Equation 2.1 (p.10) one receives 15 possible tree topologies from which three are balanced and 12 are unbalanced trees. Thus, balanced trees have a frequency 0.2, whereas imbalanced trees show a frequency 0.8 (Mooers and Heard, 1997). The random selection from all possible phylogenetic trees is not of interest in the scope of evolutionary dynamics. Hence, it is not considered in the following work.

3.2.4. Aldous' Branching Model

Choosing the parameter value $\beta = -1$ for Equation 3.4 is of particular interest because it has been demonstrated to maximize the agreement of beta-splitting with observed phylogenetic trees (Blum and François, 2006) in terms of imbalance. When β is set to -1, the model is called *Aldous' branching (AB) model* with probabilities

$$p_{-1}(i|n) = \frac{1}{a_{-1}(n)} \frac{n}{i(n-i)} \tag{3.6}$$

whereas $a_{-1}(n)$ is a suitably chosen normalization constant.

Analogous to p_{ERM}, described in previous Section 3.2.1, $p_\beta(i|n)$ is the probability that a tree has i out of its n leaves in the left subtree. While the model can statistically reproduce features of empirical trees in the databases, it does not hint at any biological explanation of these features, as Blum and François (2006) remark.

3.3. Further Models

The focus of the previous section was on beta-splitting models of Aldous including the ERM model, PDA model and AB model as special cases. This section deals with a different family of models including the Alpha model by Ford (2005) and the activity model. Both models are dynamical models which result in growing trees. The last subsection gives a brief introduction to age-dependent models.

3.3.1. Ford's Alpha Model

Ford's alpha model (Ford, 2005), sometimes also called *Uniform model*, belongs to a class of models which are parameterized by tunable parameter value $\alpha \in [0, 1]$. It is a model for recursive tree formation. Assume the tree representation as a set of leaves whereat each leave is connected to an internal node by an edge. Each internal node is again connected to other internal nodes by internal edges. Note that the root is defined as an internal node connected by a single edge to another node, which can be an internal node or a leaf. Thus, a tree of n leaves has $n - 1$ internal edges. For generating a tree, each edge which connects a leave to an

internal node has a weight $1-\alpha$. All other edges between internal nodes are assigned with a weight α each. According to their weights, an edge e is choosen at random. A new leaf connected to an edge e' is then added to the middle of e. When growing a tree by using the Ford's Alpha model, α has a regulating function by controlling the proportion of branching probabilites which is assigned with $1-\alpha$ to each leaf and proportional to α to each internal edge (Ford, 2005; Hernández-García *et al.*, 2010; Jones, 2011). Normalizing the probabilities results in $\frac{1-\alpha}{n-\alpha}$ respectively $\frac{\alpha}{n-\alpha}$. Equal to the ERM model, the branching process at a chosen leaf produces two new leaves. But choosing an internal edge for branching, the new leaf is a result of the insertion of a new internal node into the edge (Hernández-García *et al.*, 2010). For $\alpha = 0$ one gets the ERM model and for $\alpha = \frac{1}{2}$ the PDA model (Hernández-García *et al.*, 2010; Jones, 2011).

3.3.2. Activity Model

Motivated by the fact that Fords alpha model "gives a simple mechanisms for scaling in trees with tunable exponent, the dynamical rule of posterior insertions of inner nodes is hard to justify in the context of evolution" (Hernández-García *et al.*, 2010), Hernández-García *et al.* (2010) proposed the *activity model*. In the model, the set of species $S(t)$ at time t is partitioned into a set of active species $S_A(t)$ and a set of inactive species $S_I(t)$. Starting the branching process with the root, at each time step t a species s is randomly chosen with equal probability from $S_A(t)$ if $S_A(t) \neq \emptyset$, otherwise $S_I(t)$ is drawn uniformly. The emerging new species s' and s'' are added independently of each other to the active set $S_A(t+1)$ with an activation probability p. For $p = 0.5$ a critical branching process is obtained. Otherwise the model is similar to the ERM model. A variation of the activity model has been introduced by Herrada *et al.* (2011) in the context of protein family trees.

3.3.3. Bellman-Harris Model

A further development of birth-death models considering a constant time are the *Bellman-Harris models*. Those models are based on age-dependent processes (Athreya and Ney, 1971) and were first analyzed by Bellman and Harris (1952). The process of growing trees of these models is influenced by the age of each species which is the passed time since its birth. Thus the speciation and extinction is dependent on the age of species. Each species is independent from others and has no information about its parent (Jones, 2011). There has been not much attention on age-dependent processes as models for phylogenetic trees. Lately they were considered in studies by Gernhard *et al.* (2008). A new age-dependent model is presented in Section 4.1 which describes a pure-birth branching process.

3.4. Imbalance Obtained for Empirical Trees

The validity of models can be assessed by comparing the shape of phylogenetic trees (Sackin, 1972; Herrada *et al.*, 2008; Campos *et al.*, 2004; Stich and Manrubia, 2009). In particular comparing their degree of imbalance (Colless, 1982; McKenzie and Steel, 2000), with trees generated by different evolutionary mechanisms (Aldous, 2001; Blum and François, 2006; Hernández-García *et al.*, 2010), a selection of realistic models is possible. For many trees produced by models it is observed that the mean depth scales logarithmically with the number of leaves n. More precisely, models predict more imbalance than observed in trees inferred from real data which has been shown in different studies (Aldous, 1996; Steel and McKenzie, 2002; Pinelis, 2003; Mooers and Heard, 1997; Guyer and Slowinski, 1991).

Blum and François (2006) studied the phylogenetic imbalance based on trees of the database `TreeBASE`. Their analysis shows that the tree shape undergoes a rapid change from the smaller to the intermediate-sized and larger trees (Aldous, 1996). Studies by Guyer and Slowinski (1991) observed more imbalanced trees than the predicted ones by the ERM model when using small samples of trees. This excess of imbalance may be explained by errors in molecular data, incompleteness of trees and bias due to approximate reconstruction methods (Blum and François, 2006; Mooers and Heard, 1997). Overall one can say that tree shapes based on empirical data deviate significantly from those predicted by completely uncorrelated speciation processes. The depth scaling and biological motivation of previously presented models is discussed in the following. An overview is given in Table 3.1 (p.29).

With the focus on beta-splitting models, the parameter $\beta \in [-2; +\infty[$ in Equation 3.4 tunes the expected imbalance. The probability distribution of trees from the ERM model is recovered by taking $\beta = 0$. As the opposite extreme, the *Proportional to Distinguishable arrangements* model is obtained at $\beta = -3/2$ (Pinelis, 2003; Steel and McKenzie, 2001). While the depth of the ERM model scales logarithmically with the number of leaves n (Hernández-García *et al.*, 2010):

$$\langle d \rangle(n) \sim \log n \ , \tag{3.7}$$

the depth grows algebraically with the number of leaves n as

$$\langle d \rangle(n) \sim \sqrt{n} \tag{3.8}$$

for the PDA model (Mooers and Heard, 1997).

The PDA model tends to generate more unbalanced trees (Heard, 1996; Aldous, 1996; Pinelis, 2003). But when tuning the parameter value α the balance decreases. A complete unbalanced tree, also comb tree, is observable for $\alpha = 1$. But unlike the AB model, the PDA model can not generate trees which are more balanced than the ones produced by the ERM model. For the PDA model this is the consequence of choosing $\alpha = 0$ which is unique to the ERM model Hernández-García *et al.* (2010).

The tree shapes produced by the activity model differ from the ones generated by ERM model as a result of the memory in terms of internal states of the nodes (Hernández-García *et al.*, 2010). Studies by Hernández-García *et al.* (2010) have shown that for an activity probability $p = \frac{1}{2}$ the model generates trees with a mean depth growing as the square root of tree size. For $p \neq \frac{1}{2}$ and $p \in (0, 1)$ the depth seems to increase logarithmically with n (Hernández-García *et al.*, 2010).

The trees in `TreeBASE` have been found to match best with a case of the beta-splitting model when choosing the intermediate parameter value $\beta = -1.0$ (Blum and François, 2006) which is the AB model. For this model, the expected mean depth increases as

$$\langle d \rangle(n) \sim (\log n)^2 \ . \tag{3.9}$$

But the AB model and others introduced to account for tree imbalance assign probabilities to tree shapes in a way which is not based on any evolutionary principles. The same holds for the modificated ERM model since both models are not a case of open-ended evolution. Furthermore, the dynamic rule of the PDA model defined by the posterior insertion of inner nodes can also hardly described by evolutionary processes as well (Hernández-García *et al.*, 2010). With the activity model and the age model respectively the innovation model (latter two are discussed in Chapter 4), three approaches of biologically motivated models were introduced.

	depth scaling	evolutionary dynamics
β-**splitting** (Aldous, 1996)	$\begin{cases} \log n & \text{if } \beta > -1, \text{ includes } \textbf{ERM } (\beta = 0) \\ (\log n)^2 & \text{if } \beta = -1, \textbf{ AB model} \\ n^{-\beta-1} & \text{if } \beta < -1, \text{ includes } \textbf{PDA } (\beta = -1.5) \end{cases}$	$\begin{cases} \text{yes (standard)/no (modificated)} \\ \text{no} \\ \text{no} \end{cases}$
Fords alpha model (Ford, 2005)	n^α	no
activity model (Hernández-García *et al.*, 2010)	$\begin{cases} n^{0.5} & \text{if } p = 0.5, \\ \log n & \text{otherwise.} \end{cases}$	yes
age model (Keller-Schmidt *et al.*, 2010)	$(\log n)^2$	yes
innovation model (Keller-Schmidt and Klemm, 2011)	$(\log n)^2$	yes
complete tree	$\log n$	–
comb tree	n	–

Table 3.1.: Overview of the average distance of leaves from root, the depth scaling behavior, of different models. The last column indicates whether the model is biologically motivated and thus affected by an evolutionary dynamics or not. For details of the latter two models, age and innovation, see chapter 4 (p.31).

CHAPTER 4

Two New Approaches For Understanding Macroevolution

Various kinds of models generating phylogenetic trees of different attributes concerning, for instance, tree balance were presented in Chapter 3. Most of these models are simple probability distributions which are not intended to represent any evolutionary process. In addition the produced trees by some of the models show less conformity with real trees. This chapter deals with two new approaches of tree growth. The stochastic analysis and validation of generated trees using both new models show that the generated trees of both models are in good aggreement with the observed balance of empirical trees. Furthermore a biological motivation is given for both new models.

The first section introduces an age dependent growth model, called *age model*. It is based on the fact, that the tree imbalance in terms of a speciation rate is decreasing with the age of a species. In the speciation process of the age model, the branching probability of a species is inversely proportional to the time since the species was last involved in a speciation. Thus the hypothesis is that the speciation rate is a decreasing function of the waiting time since the last speciation. In Section 4.3.1 it is shown that the imbalance in terms of the mean distance of leaves from the root, also Sackin index or depth of a tree, grows as $(\log n)^2$ in leading order with tree size n. Also the shape of trees generated by the age model are in agreement with the scaling observed by exhaustive analysis of the databases `TreeBASE` and `PANDIT`. Compared to the AB model (Blum and François, 2006), the age model yields larger likelihood values on the trees in databases with up to 19 leaves.

In the second section the explanation of the so-called *innovation model* is addressed. In this case, the evolution of species is triggered by the generation of novel features and exhaustive combination with other available traits. Under the assumption that innovations are rare, a

bursty branching process of speciations is obtained. The analysis (see Section 4.3) of trees representing the branching history reveals structures qualitatively different from those of random processes. For a tree with n leaves generated by the introduced model, the average distance of leaves from root scales as $(\log n)^2$ to be compared to $\log n$ for random branching. The mean values and standard deviations for the tree shape indices depth (Sackin index) and imbalance (Colless index) of the model are compatible with those of real phylogenetic trees from databases. Again, earlier models, such as the AB model, show a larger deviation from data with respect to the shape indices.

4.1. The Age Model – an Age-Dependent Method

The *age model* can be defined as a stochastic procedure which describes the growth of binary trees by an iterative stochastic attachment of leaves, similar to the ERM model. In contrast, the branching rate at each clade is no longer constant, but is decreasing in time, i.e., with the age. The age of a species is defined by the time that passed from the birth of that leave to the present time. Put differently, species involved in recent speciation events have a tendency to speciate again. This amounts to bursting behavior in evolutionary activity. That is to say that the probability of speciation is inversely proportional to the age of a species. At each time, a species s is drawn from the set of species $S(t)$ with probability

$$\pi_s(t) \propto \tau_s(t)^{-1} \tag{4.1}$$

normalized properly. Each leaf s is assigned an age $\tau_s(t)$ being the time that passed from the birth of the leaf, t_s, to present time t, i.e. $\tau_s(t) = t - t_s$. The growth proceeds by iterating through the following three steps:

(i) A species s is chosen with probability $p_s(t)$ inversely proportional to its age

$$p_s(t) = \frac{\tau_s(t)^{-1}}{c(t)}, \tag{4.2}$$

where $c(t)$ is chosen such that probabilities of all leaves sum up to 1.

(ii) Two new leaves k and l with creation times $t_k = t_l = t$ are attached to node s.

(iii) Time t is increased by Δt and the process resumes at (i). It is considered that a constant time increment $\Delta t = 1$ unless indicated otherwise. With this choice, time t is equivalent to number of branching events, and $t = n - 1$.

A visualisation of the stepwise process of speciation is depicted in Figure 4.1. A pseudocode is given in Algorithm 3.

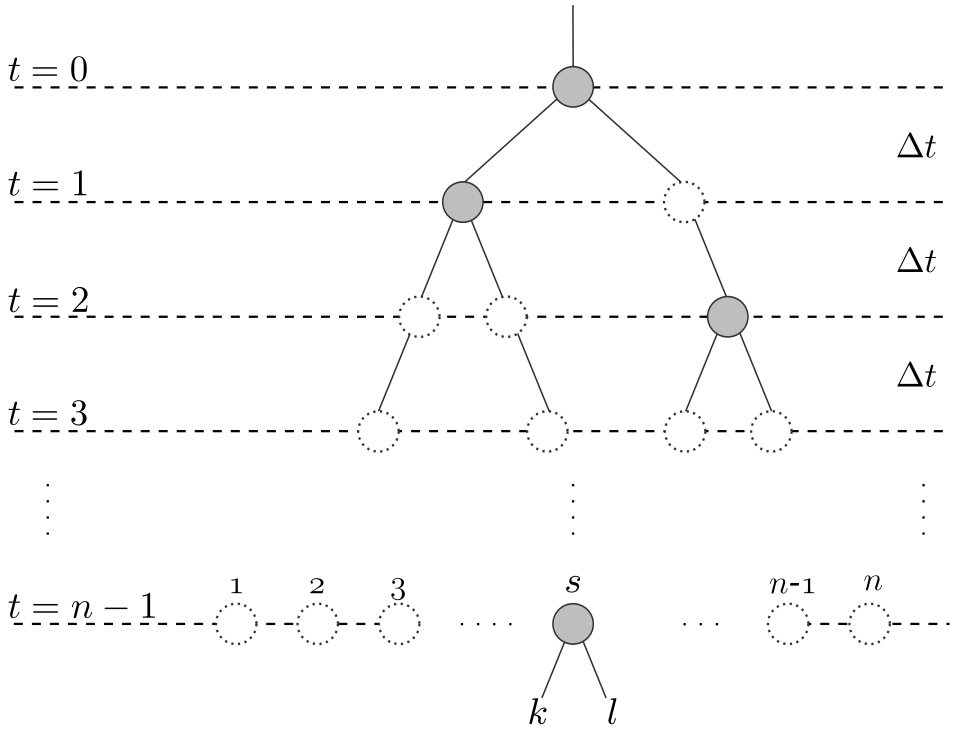


Figure 4.1.: The process of speciation using the age model. At time step $t = 0$ a root of the tree is generated and grows with a discrete time $\Delta t = 1$. t is equivalent to the number of branching events and thus $t = n - 1$. Each level represents a time step containing the set of nodes which are able to speciate. Gray circles define the speciating node in each level and white circles the nodes which were not chosen to speciate.

Algorithm 3: Pseudocode for the age model. Based on the hypothesis that speciation rate is a decreasing function of waiting time since last speciation of a node.

Data: N ... amount of nodes / species which simulated tree T should have;
Δt ... time constant;
Result: T of size $N = |T|$

```
1 set t = 0, create root;
2 while |T| < N do
3     choose leave s ∈ T with p_s(t) = τ_s(t)^{-1} / c(t);
4     attach two new leaves k, l with t_k = t_l = t to i;
5     t = t + Δt;
```

4.2. The Innovation Model

Since the seminal work by Darwin (Darwin, 1859), the evolution of biological species has been recognized as a complex dynamics involving broad distributions of temporal and spatial scales as well as stochastic effects, giving rise to so-called frozen accidents. These are incidents with extensive and manifold consequences to the future. They are reproducable to one chance event which could have turned out differently (Gell-Mann, 1995). There is vast exchange and overlap of concepts and methods between the theory of evolution and the foundations of complex systems such as fitness landscapes (Wright, 1932; Gavrilets, 2004; Klemm and Stadler, 2012) and neutral networks (Kimura, 1983), the evolution of cooperation (Axelrod, 1984) and self-organized criticality (Per, 1996) to name but a few.

A striking feature of biological macroevolution is its burstiness. The temporal distribution of speciation and extinction events is highly inhomogeneous in time (Sepkoski, 1993). As described by the theory of punctuated equilibrium (Gould and Eldredge, 1993), a connection between punctuated equilibrium in evolution and the theory of self-organized criticality (Per, 1996) is established through the model by Bak and Sneppen (Bak and Sneppen, 1993; Sneppen *et al.*, 1995). Ecology, i.e., the system of trophic interactions and other dependencies between species' fitnesses, is driven to a critical state. Then minimal perturbations cause relaxation cascades of broadly distributed sizes.

Rather than through ecological interaction across possibly all species, bursty diversification may also be due to *adaptive radiation* as a rapid multiplication of species in one lineage after a triggering event. About 200 million years ago, a novel chewing system with dedicated molar teeth evolved in the lineage of mammals, allowing it to rapidly diversify into species using vastly distinct types of nutrition (Ungar, 2010). There are many more examples where a single *innovation* triggers adaptive radiation such as the tetrapod limb morphology caused by a binary shift in bone arrangement (Thomson, 1992) and the homeothermy as a key innovation by the group of mammals (Heard and Hauser, 1995; Liem and Nitecki, 1990). Environmental conditions a species has not encountered previously, e.g., when entering a geographical area with unoccupied ecological niches, may also be the source of adaptive radiation. The diversity of finch species on Galapagos islands is the famous example first studied by Darwin. Spontaneous phenotypic or genetic innovations and those caused by the pressure to adapt to a change in environment are treated on the same footing for the modeling purposes in this contribution. Though being a central concept in the theory of evolution, the term innovation has not been ascribed a unique definition so far (Pigliucci, 2008).

The model, introduced and studied in this chapter, is a branching process to mimic the evolution of species driven by innovations. The process involves a separation of time scales. Rare innovation events trigger rapid cascades of diversification where a feature combines with previously existing features. The newly defined branching process is called *innovation model*.

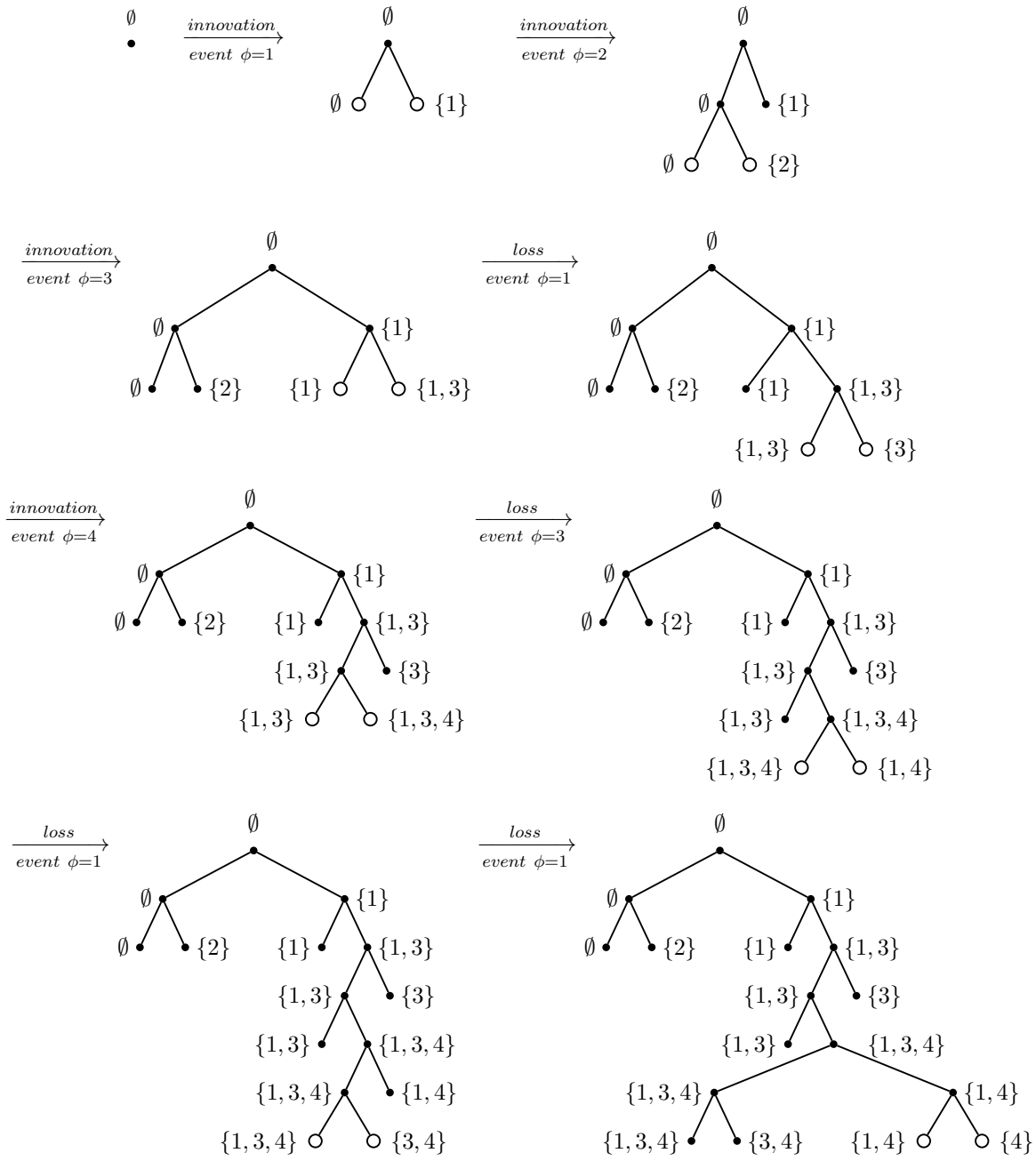


Figure 4.2.: Example for generating a tree of nine leaves applying the innovation model. The root node labeled with the feature set $\{\emptyset\}$ speciates through the application of an innovation event by adding the feature $\{1\}$ to the feature set. This results in two nodes labeled with $\{\emptyset\}$ respectively $\{1\}$. The innovation events are performed until a loss step is possible. The loss event is performed by removing the feature $\{1\}$ and resulting in the new feature set $\{2\}$ which does not occur in the tree yet. The process is repeated until the tree has nine leaves.

The Featured-Based Method In the *innovation model*, each species s is defined as a finite set of features $s \subseteq \mathbb{N}$. Features are taken as integer numbers in order to have an infinite supply of symbols. Let $F(t)$ be denoted by the set of all features existing at time t, that is $F(t) = \bigcup_{s \in S(t)} s$. Each speciation occurs as one of two possible events.

Definition 4.1 (innovation event)**.** An *innovation event* is the addition of a new feature $\phi \in \mathbb{N} \setminus F(t)$ not yet contained in any species at the given time t. One of the resulting species carries the new feature, $s' = s \cup \{\phi\}$. The other species has the same features as the ancestral one, $s'' = s$.

Definition 4.2 (loss event)**.** A *loss event* generates a new species by the disappearance of a feature. A feature ϕ is drawn from $F(t)$ uniformly. The loss event is performed only if $s \setminus \{\phi\} \notin S(t)$ such that elimination of ϕ from s actually generates a new species. In this case, the resulting species are the one having suffered the loss, $s' = s \setminus \{\phi\}$ and the species $s'' = s$ remaining unaltered. Otherwise, ϕ is not present in s or its loss would lead to another already existing species, so nothing happens.

In both cases, a species s is drawn from the set of species $S(t)$ at time t with uniform probability for speciation. For the case of the model, the assumption is made that creation of novel features is significantly less abundant than speciation by losses. This separation of time scales is implemented by the rule that an innovation event is only possible when no more losses can be performed. In order to facilitate further studies in the following, a pseudocode description in Algorithm 4 is provided. Furthermore the process of generating a tree by applying the innovation model is given in Figure 4.2, which shows an example of the dynamics.

4.3. Comparison of simulated and empirical trees

With the introduction of the new growth models, age and innovation model, in Section 4.1 respectively Section 4.2, a validation of those is necessary. A comparison by simple inspection of trees from real data and models may already reveal substantial shape differences. Figure 4.3 shows an example. The trees in panels (a), (b) and (c) are less compact than that of panel (d) of Figure 4.3. Panel (a) represents a tree from `TreeBASE`. Trees generated with the age model and innovation model are shown in panel (b) respectively (c). In panel (d) a tree created as realization of the ERM model is depicted.

For an objective and quantitative comparison of trees generated by models and empirical trees, the three following measures of tree shape are analyzed:

- The Sackin index d (Sackin, 1972) describing the compactness of a tree by the average distance of all leaves from root (see 2.3.2 (p.16)).

Algorithm 4: Pseudocode for the innovation model

Data: N ... final size of simulated tree T;
S ... set of all species s;
$F(t)$ ... set of all features existing at time t;
Result: T of size N

set $t = 1$, $F(0) = \emptyset$, $S(0) = \{\emptyset\}$;
while $|S(t)| < N$ **do**
  if $S(t) \setminus \{s \setminus \{\phi\} : s \in S(t), \phi \in F(t)\} \neq \emptyset$ **then**
    `// loss event`
    draw $\phi \in F(t)$ uniformly;
    draw $s \in S(t)$ uniformly;
    if $s \setminus \{\phi\} \notin S(t)$ **then**
      $S(t+1) = S(t) \cup \{s \setminus \{\phi\}\}$;
      $F(t+1) = F(t)$;
      increment t;
  else
    `// innovation event`
    draw $s \in S(t)$ uniformly;
    set $\phi = 1 + \max(F(t) \cup \{0\})$;
    set $S(t+1) = S(t) \cup \{s \cup \{\phi\}\}$;
    set $F(t+1) = F(t) \cup \{\phi\}$;
    increment t;

- The Colless index c (Colless, 1982) for the evaluation of tree balance (see 2.3.3 (p.16)).
- The cherry distribution (McKenzie and Steel, 2000) as measurement of tree quality by computing the number of cherries of a tree (see 2.3.4 (p.16)).

For the analysis and comparison the simulated data sets were chosen as follows. A data set for each model (AB model, ERM model, age model and innovation model) encompasses 1,000 trees for each tree size from 5 to 535 and ten trees for each tree size from 536 to 2,562. Empirical data sets were taken as explained in Section 2.3.3 (p.16).

4.3.1. Evaluation using the Sackin Index

This section consists of two parts. Each of them deals with the quantification of tree imbalance using the Sackin index, in the following also named depth d of a tree. The first part studies trees generated with the age model, the second deals with trees created by the innovation model. The average depth values d of the empirical trees from `TreeBASE`, `PANDIT`, and `McPeek` are depicted in Figure 4.4

$$d \sim (\log n)^2 \tag{4.3}$$

in good approximation. Alternative analytic expressions for the growth can be fitted, in particular a power law $d \sim n^\gamma$, with $\gamma \approx 0.4$ describes `TreeBASE` data equally well (Herrada *et al.*, 2008). But for the larger tree sizes contained in `PANDIT`, the $(\log n)^2$ form is more accurate (Herrada *et al.*, 2011).

For evaluating the similarity of model generated trees with empirical trees from `TreeBASE`, `PANDIT` and `McPeek`, a p-value analysis is applied. The p-value is calculated as follows. Given a model M and an empirical tree T of depth d and having n_{real} leaves. The fraction f of the model trees, consisting of n_{model} leaves, possessing a depth which is larger as or equal to d can be calculated. For each tree size which is obtained in the empirical data set, 1,000 trees were generated with M. Each model tree was tested against each real tree of size n_{real} if $n_{model} = n_{real}$. Taking that p-value and a significance level $\alpha = 0.05$, a model can be *rejected*, if $f < 0.05$ or $f > 0.95$, i.e., the observed value lies at one of the extremes of the model distribution. Table 4.1 (p.42) shows an overview of the fraction of trees for the age model, ERM model and AB model which were rejected, respectively not rejected compared to empirical trees of `TreeBASE`, `PANDIT` and `McPeek`. The not rejected cases are called *accepted* in the following. The total set of trees for each stochastic data set is divided into four subsets containing about 25% of total trees with an increasing tree size. Each tree size is fully included in one of the subsets. For `TreeBASE`, for instance, 1,347 out of 5,087 (which makes 26%) trees have 18 or fewer leaves. The age model is acceptable for a fraction of 0.8270 (=1,114 trees) out of these 1,347.

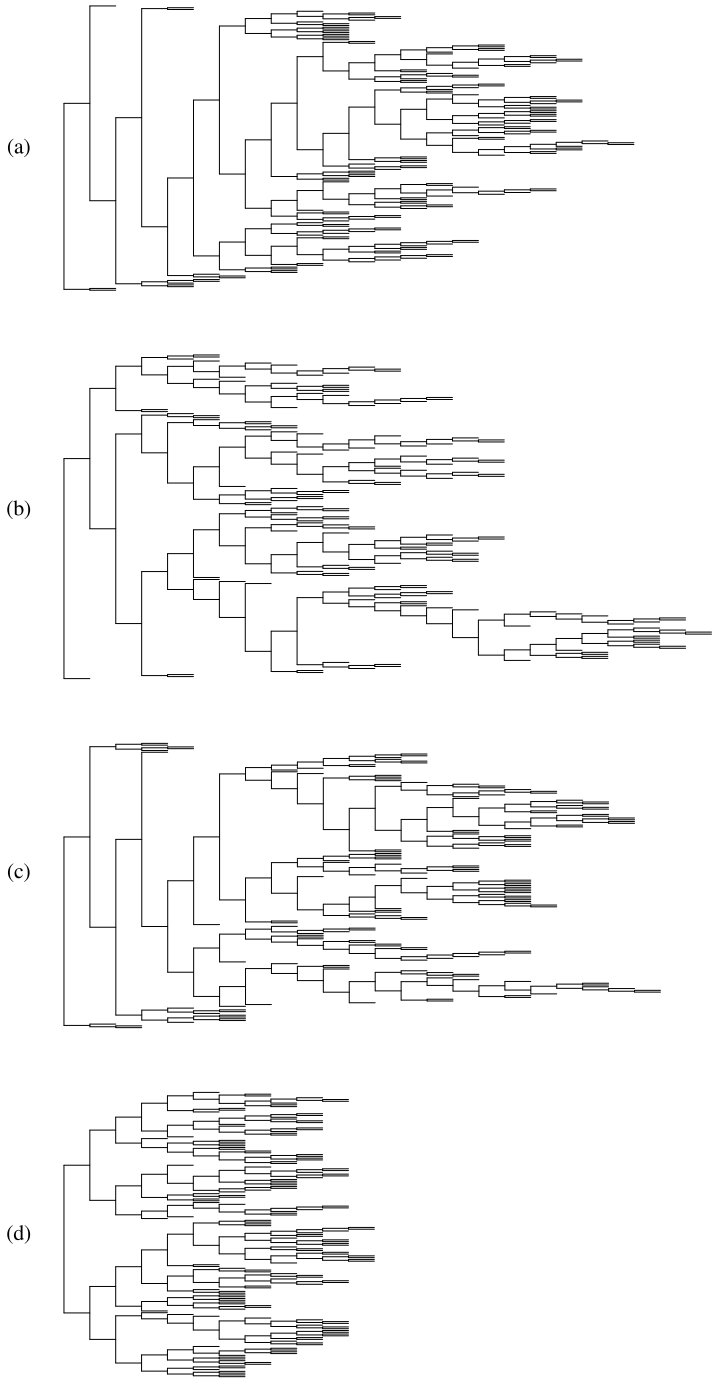


Figure 4.3.: Empirical and simulated trees. The depicted phylogenetic tree in (a) is from the database TreeBASE (Matrix ID M2957, relationships in *rosids* based on mitochondrial matR sequences), (b) is a tree generated with the age model, (c) is a tree created as a realization of the innovation model, and (d) a tree from the ERM (Yule) model. Each of the trees has 161 leaves.

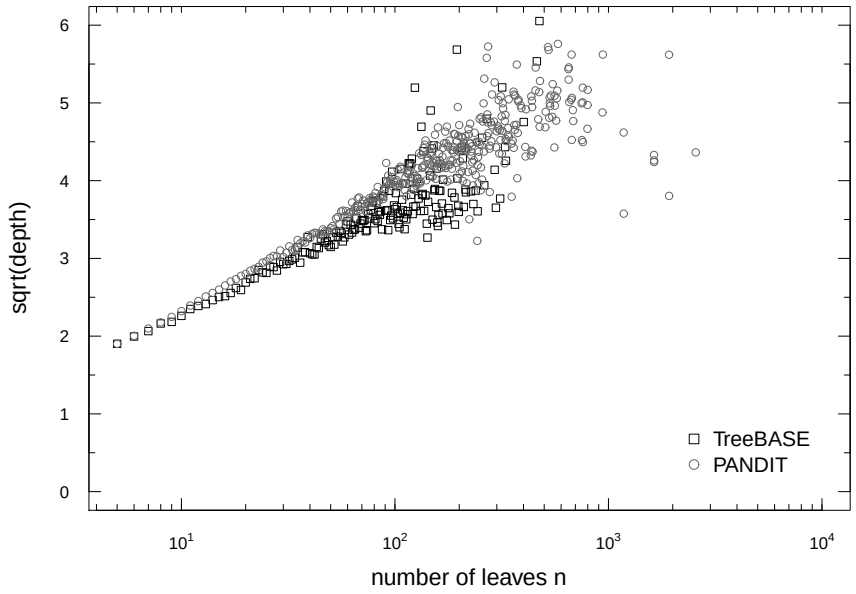


Figure 4.4.: The square-root of the mean depth vs. size of phylogenetic trees contained in databases for species (TreeBASE, □; McPeek, △) and proteins (PANDIT, ○). The depth is averaged for all trees having the same number of leaves. In this scale (log-linear), the behavior $\langle d \rangle \sim (\log n)^2$ is a straight line.

Age model

For the age model, there is evidence that the expected depth $\langle d \rangle$ increases as $(\log n)^2$ with the number of leaves n. An approximation is shown in Section 4.4.3. Hence, the growth law is identical to the AB model. Also, it is in good agreement with the depth values obtained from the databases `TreeBASE` and `PANDIT` (see Figure 4.8 (p.49)). The results of the p-value analysis for supporting the similarity of trees from the databases and the ones generated by the age model with the time increment of $\Delta t = 1$ is depicted in Table 4.1 (p.42). The accepted trees of the ERM model are decreasing with tree size for all three data sets and are the most-often rejected trees. The results show that 80% of the trees with 5 to 47 leaves generated by the age model are accepted when compared the `TreeBASE` data set. Thus, the age model performs slightly better than the AB model. For trees with 48 to 535 leaves 71% of trees generated by the AB model but 70% of trees from the age model are evaluated as accepted. The results for the age model are similar for the `PANDIT` and `McPeek` data set, whereat AB model performs best for the `McPeek` data set.

When comparing the mean values of the Sackin index in Figure 4.5a from the age model, ERM model and AB model to the empirical data sets, the age model shows the smallest discrepancy and fits well with the `PANDIT` and `TreeBASE` dataset. The `McPeek` data set shows a mean depth distribution between the AB model and ERM model generated trees. But compared to the other two empirical data sets, `McPeek` is of a small size and a small number

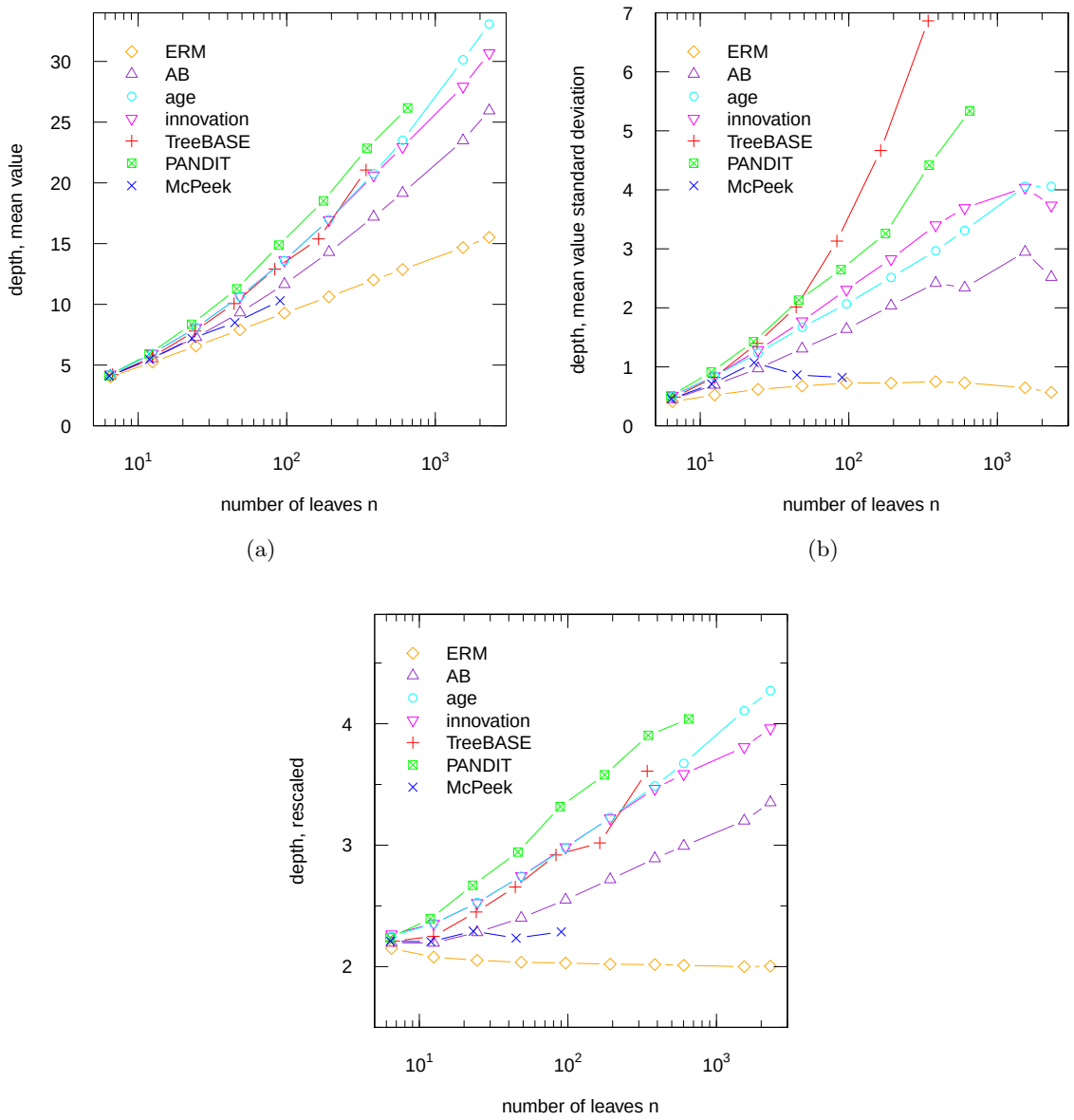


Figure 4.5.: Comparison of size-dependent summary statistics for models and real trees. Symbols distinguish the ERM model, the AB model and the age model respectively the innovation model and the data sets `TreeBASE`, `PANDIT` and `McPeek`. The mean values of depth, panel (a), are binned logarithmically as a function of tree size n. The mean value of n is calculated for each bin. The same procedure is applied to the standard deviation, panel (b). Panel (c) depicts the same values of depth as in panel (a) and (b) with an n-dependent rescaling. This is the average depth divided by $\ln n$. The factor is chosen such that the rescaled values for the ERM model asymptotically approach a constant. See reference Blum *et al.* (2007) for the scaling of the indices of the ERM model.

#leaves	AB m.		ERM m.		age m.		innovation m.		#T
	acc	rej	acc	rej	acc	rej	acc	rej	
(a) `TreeBASE` (5,087 trees in total)									
5 .. 18	0.8478	0.1522	0.6570	0.3430	0.8270	0.1730	0.8315	0.1685	1347
19 .. 31	0.7693	0.2307	0.4335	0.5665	0.7972	0.2028	0.8465	0.1535	1218
32 .. 47	0.7250	0.2750	0.2713	0.7287	0.7976	0.2024	0.8384	0.1616	1349
48 .. 535	0.7187	0.2813	0.1679	0.8321	0.7076	0.2924	0.7349	0.2651	1173
(b) `PANDIT` (36,136 trees in total)									
5 .. 7	0.6472	0.3528	0.6472	0.3528	0.6314	0.3686	0.6314	0.3686	9464
8 .. 12	0.7856	0.2144	0.5665	0.4335	0.8537	0.1463	0.8322	0.1678	9212
13 .. 25	0.6109	0.3891	0.2876	0.7124	0.8146	0.1854	0.8444	0.1556	8806
26 ..2,562	0.4148	0.5852	0.0879	0.9121	0.7148	0.2852	0.7858	0.2142	8654
(c) `McPeek` (238 trees in total)									
5 .. 9	0.8361	0.1639	0.7705	0.2295	0.8033	0.1967	0.7705	0.2295	61
10 .. 14	0.9545	0.0455	0.6970	0.3030	0.8788	0.1212	0.8485	0.1515	66
15 .. 24	0.8654	0.1346	0.6538	0.3462	0.7115	0.2885	0.8077	0.1923	52
25 .. 116	0.8814	0.1186	0.6441	0.3559	0.6271	0.3729	0.6441	0.3559	59

acc = accepted; rej = rejected; m. = model; #T = amount of trees in subset

Table 4.1.: P-values with a significance level $\alpha = 0.05$ for each model regarding real trees from (a) `TreeBASE`, (b) `PANDIT` and (c) `McPeek`. Each model tree of size n_{model} was tested against each real tree of size n_{real} when $n_{model} = n_{real}$. For each n_{real} 1,000 trees were realized by each model. Results where divided into partition of size >= 25% of trees so that all trees of same size are in one subset. The fraction f of the model trees, consisting of n_{model} leaves, possessing a depth which is larger as or equal to depth d for an empirical tree T of size n_{real} is calculated. A model is, if rejected, if $f < 0.05$ or $f > 0.95$. The model showing the highest acceptance rate is displayed in green, the lowest in red.

of trees of large size. This may lead to an improper distribution.

Comparing the standard deviation of the mean depth values (shown in Figure 4.5b), the age model performs slightly worse than the innovation model but still bear a larger resemblance to the ERM model and AB model, besides the innovation model, which is discussed in the next part. The values of `McPeek` are again between those of AB model and ERM.

In Figure 4.5c, the averages of the Sackin index is depicted after rescaling to facilitate the comparison. The age model values are the best matching to those of `PANDIT`, `TreeBASE` and `McPeek`. The curves from the age model and the `TreeBASE` data set possess a high overlap. The innovation model performs similar and is discussed in the following.

Innovation model

The innovation model generated trees show a similar imbalance as real trees from `TreeBASE`, `PANDIT` and `McPeek`. Table 4.1 shows the results of the p-values analysis for each model including the innovation model in the last column. The innovation model shows the highest acceptance rate for the `TreeBASE` data set for the three subset with tree size larger than 18 which is also tree size independent and reaches from 0.7349 to 0.8465. For the `PANDIT` data set, the innovation model shows the highest rate of acceptance for three out of four subsets regarding tree size. Only for small trees of size three to seven the AB model shows a slightly higher acceptance rate with a difference of 0.0158. The p-values are independed of tree size for the AB model as well as for the innovation model. The acceptance rate of trees generated using the ERM do not show such an independency but is decreasing with the tree size. As already pointed out previously, the AB model performs best for the the `McPeek` data set, but for tree of size 10 to 24, the age model and innovation model still show an acceptance rate of 71% to 85%.

An ensemble of mean values and standard deviations of the Sackin index is shown in Figure 4.5. Comparing the results of the three models, ERM model, AB model, and innovation model, to those of the trees from two databases, the least discrepancy is obtained between the innovation model and the trees from `TreeBASE`, representing macroevolution. In Figure 4.5c, the averages of the Sackin index are shown after rescaling to facilitate the comparison. Of all models, the values of the innovation model are also best matching those of `PANDIT`.

According to the p-values and Figure 4.5, out of all models, the results of the innovation model are the best matching those of `PANDIT` and `TreeBASE`. Furthermore one observes the lowest discrepancy to the results of the `McPeek` data set. But the results of the age model are almost as good or identical.

4.3.2. Validation by the Colless Index

For measuring the average imbalance of trees the Colless Index is calculated for each empirical data set (`TreeBASE`, `PANDIT`, `McPeek`) and stochastic data set including ERM model, AB model, age model and innovation model. Figure 4.6a and 4.6b depict the mean value respectively the standard deviation of the Colless index for each data set. Comparing the results of four models to those of trees from the two databases `TreeBASE` and `PANDIT`, the least discrepancy is obtained between the innovation model respectively age model and the trees from `TreeBASE`, representing macroevolution. The `McPeek` data set shows a different scaling for larger trees compared to trees from `TreeBASE` and `PANDIT`. Thus trees of size 40 to 90 the AB model shows the highest conformity. One reason for that could be the small data set of 232 trees compared to 5,087 for `TreeBASE` respectively 36,030 trees for `PANDIT`. Figure 4.6c shows the averages of Colless index after rescaling to facilitate the comparison. Of all models, the values of the innovation model and age model are also best matching those

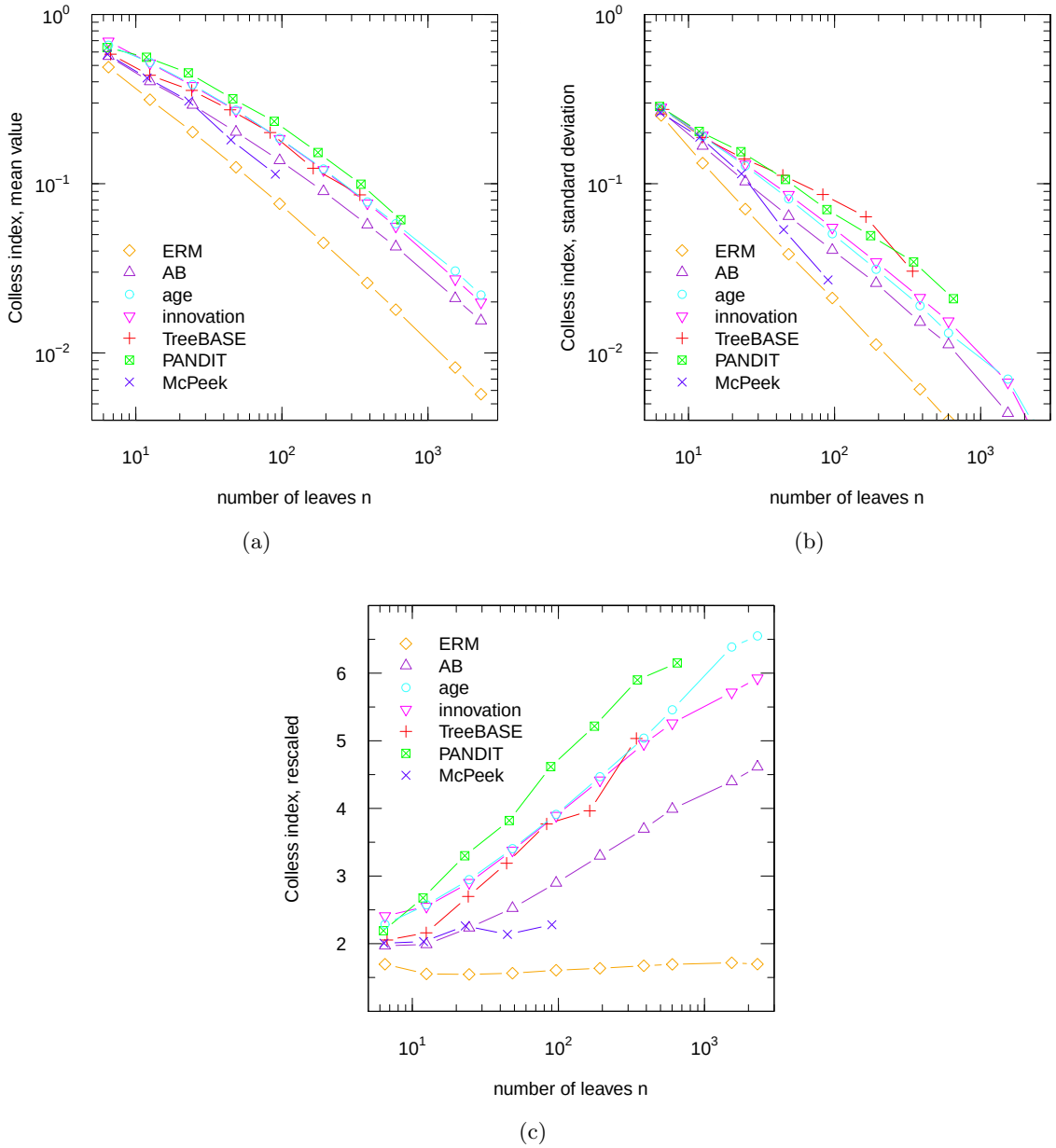


Figure 4.6.: Comparison of size-dependent summary statistics for model generated and real trees. Curbes are shown for ERM model, the AB model, the age model and the innovation model and the data sets TreeBASE, PANDIT and McPeek. The Colless index, panel (a), is binned logarithmically as a function of tree size n. The mean value of n is calculated for each bin. The same procedure is applied to the standard deviations, panel (b). The same values of the Colless index as in panel (a) with an n-dependent rescaling is shown in panel (c). Therefore, the average Colless index is divided by $n^{-1} \ln n$. These factors are chosen such that the rescaled values for the ERM model asymptotically approach a constant. See reference Blum *et al.* (2007) for the scaling of the indices of the ERM model.

of TreeBASE and PANDIT whereat the values of TreeBASE shows a smaller discrepancy for tree of size 250 and up.

4.3.3. Analysis of the Cherry Distribution

For the measurement of the tree quality when testing empirical trees (TreeBASE, PANDIT, McPeek) against simulated trees (from ERM model, AB model, age model, innovation model) the cherry distribution for each data set is studied. For that purpose the amount of cherries for each tree in each data set is calculated.

The mean value of cherries normalized by the tree size as function of leaves n is depicted in Figure 4.7a. For the ERM model the largest number of cherries can be observed. The number of cherries of trees generated by the innovation model seems to increase with n and approaches to the curve of he ERM model. At this point, for an explanation, the process performed during an innovation event is anticipated and described in detail in Figure 4.10 (p.53) and corresponding text. The increase might be caused by the addition of a tree when an innovation event is processed. The added tree is similar to those generated with the ERM model. The curve of TreeBASE is falling at the end due to the appearance of a single tree for that size and thus is not representative. The smallest deviations to TreeBASE for trees of size 30 to 130 approximately is observed for the innovation model. But also the AB model is in agreement with the empiricial data sets when compared to the age model and ERM model.

But without normalizing the mean value of cherries as function of n, shown in Figure 4.7c, all data sets show a similar scaling behavior. Here, the mean values are binned logarithmically. A more detailed plot of the behavior and associated regression lines reflects Figure 4.7d. Here the mean values are binned logarithmically for trees of size 5 to 64 and trees of size < 99 are binned in partitions of size 50. As shown in studies by McKenzie and Steel (2000) and Matsen (2006) the distribution grows asymptotically normal with increasing number of leaves. The coherence between the cherry count and leaf number is also represented by the regression lines whereby the stochastical data sets show a stronger dependency than the empirical data sets. The more narrow both regressions lines for each data set in the intersection the higher the observable stochastic dependencies between the tree size and the number of cherries. For all model generated trees, the regressions lines are one upon the other, also the empirical data sets show only a small difference. Thus, all trees show a high stochastic dependency. The regression line for the age model is the closest to the one from TreeBASE and thus shows a similar dependency of cherries from leave counts. The other simulated data sets are in accordance to the McPeek and PANDIT data sets, which are also similar among one another. The ERM model trees are the most balanced ones which results in more cherries with increasing tree size.

The standard deviation for the amount of cherries is given by Figure 4.7b. The standard deviation is a increasing function with tree size for all data sets. Observing similar values

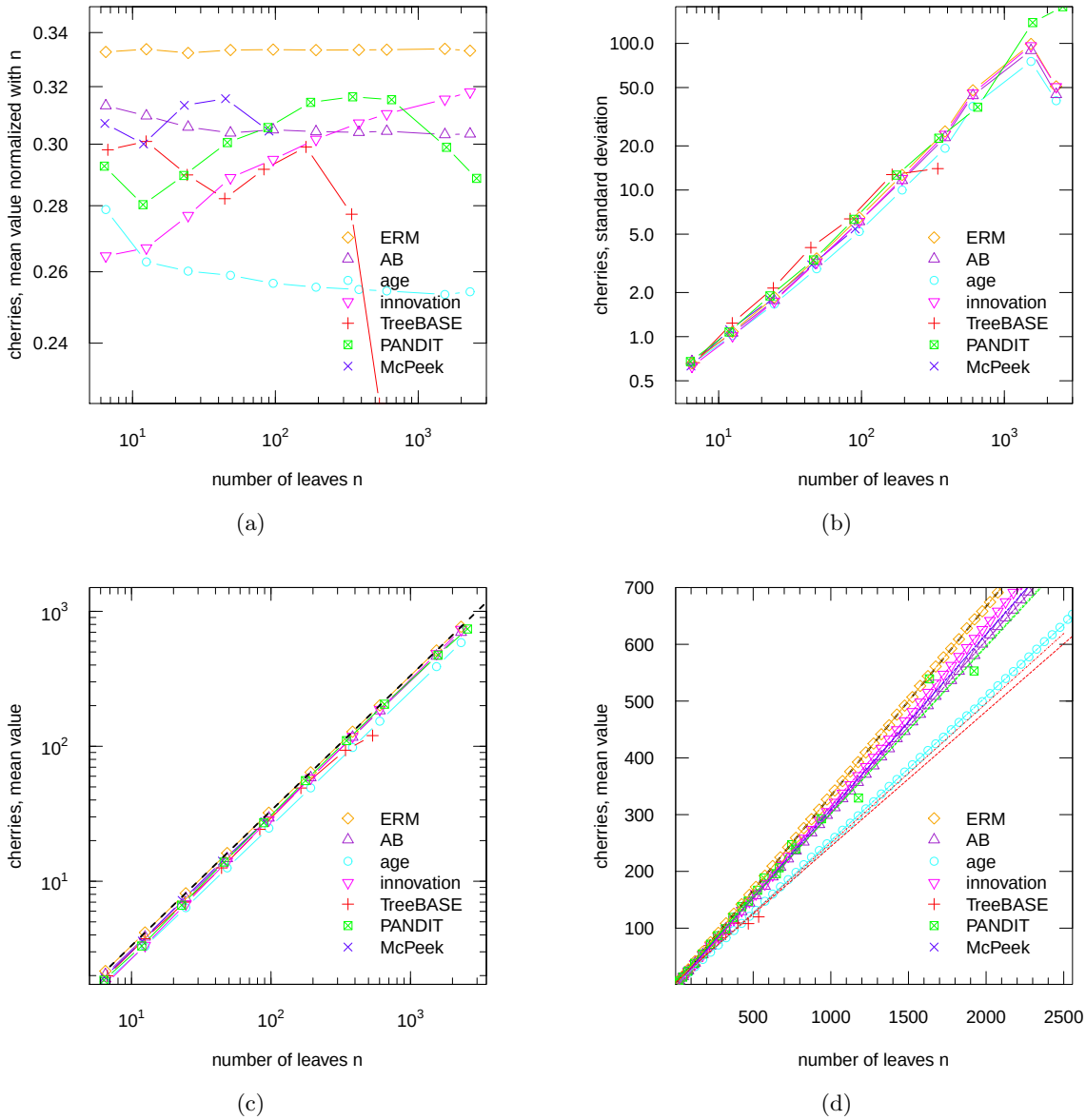


Figure 4.7.: (a) Mean of cherry number normalized with n over trees of logarithmically binned tree size as a function of leaves n. (c) Number of cherries as a function of leaves n. Trees of size 5 to 64 are binned logarithmically, (d) Same as panel (c) with a more detailed binning: trees of size 5 to 64 are binned logarithmically, and trees of size < 99 are binned in partitions of size 50. The mean value of n is calculated for each bin. Additionally, the regressions lines for each data set is depicted. Dashed line presents regression line for number of cherries $\sim n$ and dashed-dotted line corresponds to $n \sim$ number of cherries.

for small trees, the curves of the data sets drift apart with increasing tree size. The curve of the ERM model is the closest to the curves of the empirical data sets with 130 leaves approximately, but the almost overlaying curves of the AB model and innovation model are near as well as the age model data.

Overall, when comparing the mean values of cherries, all models generate trees with a similar behavior as empirical trees. But if taking bins of a small range of tree size, the age model shows the lowest discrepancy to `TreeBASE`. But for the AB model and innovation model a similar behavior to `PANDIT` and `McPeek` is observed. The ERM possess the greatest differences to the empirical data sets.

4.4. Approximation of depth scaling

The mean depth can not only be used for imbalance measurement but also for tree comparison. Thus the depth scaling behavior of model generated trees is used to compare among one another with empirical trees. This is done by answering the question how the mean depth scales with tree size.

4.4.1. Mean Depth Scaling of Most Imbalanced Trees

Considering a rooted, binary but completely imbalanced tree with n leaves, the depth d' for all nodes in the tree is given by

$$n \cdot d'(n) = 1 + 2 + \ldots + (n-2) + (n-1) + (n-1) \tag{4.4}$$

$$= \sum_{k=1}^{n-1} k + (n-1) \ . \tag{4.5}$$

Therefore, the mean depth is obtained by dividing each term by n

$$d(n) = \sum_{k=1}^{n-1} \frac{k}{n} + \frac{n-1}{n} \tag{4.6}$$

$$= \sum_{k=1}^{n} \frac{k}{n} - \frac{1}{n} \tag{4.7}$$

$$= \frac{1}{n} \sum_{k=1}^{n} k - \frac{1}{n} \ . \tag{4.8}$$

Now substitute the sum by a finite sequence which leads to

$$d(n) = \frac{1}{n}\left(n\frac{n+1}{2}\right) - \frac{1}{n} \tag{4.9}$$

$$= \frac{n}{2} + \frac{1}{2} - \frac{1}{n} \tag{4.10}$$

and results in a depth scaling

$$d(n) \sim n \ . \tag{4.11}$$

4.4.2. Mean Depth Scaling of Most Balanced Trees

In a complete balanced tree each leaf has the same distance to the root. The size of such a tree, which additional is rooted and binary, is given by $n = 2^i$. Therefore, i defines the tree levels starting with zero for the root and also stands for the mean depth of each level. Thus $d(n) = i$ and defining $d(n) = \log_2 n$ results in a scaling of

$$d(n) \sim \log n \tag{4.12}$$

(Hernández-García *et al.*, 2010).

4.4.3. Mean Depth Scaling of the Age Model

The n-dependence of the expected depth of trees stochastically generated by the age model is analyzed in the following for $\Delta t = 1$. Numerical and heuristic arguments strongly suggest that $d \sim (\log n)^2$ is the asymptotic growth law for this model (see Figure 4.8), but a fully rigorous demonstration of that is not provided. Instead the upper and lower bounds for the depth in the model are established, and provide numerical evidence for the $(\log n)^2$ scaling of them, from which the same behavior would hold for $d(n)$.

In Keller-Schmidt *et al.* (2010) (version 1), the upper and lower bound recursions were derived assuming that the replacement of the actual age distribution by an extreme case does not yield a decreasing depth. But this assumption does not hold in general, which is shown at the end of this section.

First, a single realization of the stochastic process is considered. For each integer time $t > 0$, let $\delta(t)$ be the distance from root of the two new leaves added at time t. This means that $\delta(t) - 1$ is the distance from root of the leaf chosen to speciate. Let $\tau(t)$ be the age of the leaf chosen at time t. Then $\delta(t)$ obeys the recursion

$$\delta(t) = \delta(t - \tau(t)) + 1 \tag{4.13}$$

for $t > 1$ with $\delta(1) = 1$ as initial condition.

Now, considering the case that the process has generated the sequence $\delta(1), \delta(2), \ldots, \delta(t-1)$,

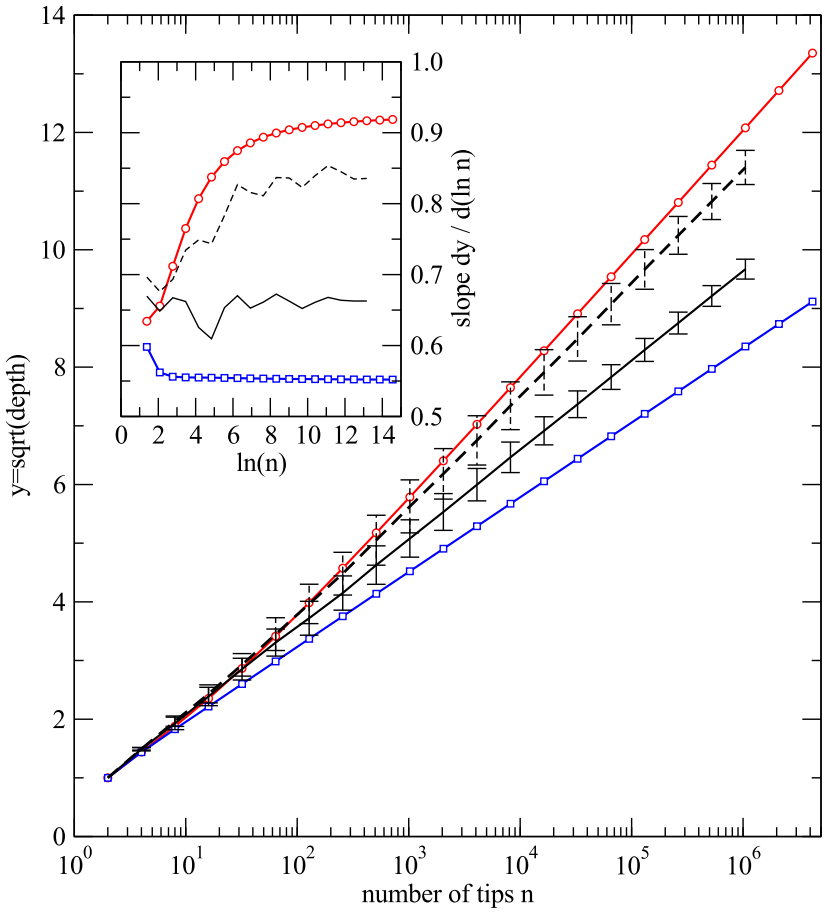


Figure 4.8.: The dependence of depth d on the number of leaves n. Curves with symbols are the lower (□) and upper (○) bounds obtained by the recursion Equations 4.19 and 4.22 inserted into Equation 4.24. Stochastic simulations yield an average depth plotted as the solid line with error bars indicating standard deviation over the 30 independent realizations with $\Delta t = 1$. Analogously, the dashed line is for stochastic simulations but using a time increment $\Delta t = 1/n$. Note that $\sqrt{d}$ is plotted over a logarithmic n-axis, so the dependence $d \sim (\log n)^2$ results in a straight line. The inset shows the slopes of the curves in the main panel, which display better the asymptotic approach to a constant slope, i.e., the approach to a $(\log n)^2$ growth.

one can ask for the expectation value η of $\delta(t)$. In the calculation of η also the distribution $f(\tau, t)$ of ages of the leaves of the tree enters as

$$\eta = 1 + \frac{\sum_{\tau=1}^{t-1} f(\tau, t)\tau^{-1}\delta(t-\tau)}{\sum_{\sigma=1}^{t-1} f(\sigma, t)\sigma^{-1}} \ . \tag{4.14}$$

In the following, an f-independent *lower-bound* on η is established. To this end, a particular age distribution is defined as

$$f_{\leq}(\tau, t) = \begin{cases} 2t^{-1}, & \text{if } \tau = 1 \\ t^{-1}, & \text{if } 2 \leq \tau < t \\ 0, & \text{if } t \leq \tau \end{cases} \tag{4.15}$$

Dynamically, this age distribution is obtained when one of the youngest leaves ($\tau = 1$) is chosen in each step. The assumption is made that the expected level does not increase when replacing the actual age distribution f by $f_{\leq}$. Therefore

$$\eta \geq 1 + \frac{\sum_{\tau=1}^{t-1} f_{\leq}(\tau, t)\tau^{-1}\delta(t-\tau)}{\sum_{\sigma=1}^{t-1} f_{\leq}(\sigma, t)\sigma^{-1}} \ . \tag{4.16}$$

The expectation value $\langle\delta\rangle(t)$ over the whole stochastic process is obtained formally by an average over all histories as follows. Call $\mathcal{D}_t$ the set of all eligible distance sequences of length $t-1$ and $\mathcal{F}_t$ the set of all eligible age distributions at time t. Then one can write

$$\langle\delta\rangle(t) = 1 + \sum_{\delta\in\mathcal{D}_t}\sum_{f\in\mathcal{F}_t} p(\delta, f, t)\frac{\sum_{\tau=1}^{t-1} f(\tau, t)\tau^{-1}\delta(t-\tau)}{\sum_{\sigma=1}^{t-1} f(\sigma, t)\sigma^{-1}} \tag{4.17}$$

with p being the joint distribution of distance sequence and age distribution at a given time. An exact solution for $\langle\delta\rangle(t)$ would thus involve a recursion for p, which is difficult to treat. The lower bound on η in Equation 4.16, however, is valid for each possible realization of the process. Therefore

$$\langle\delta\rangle(t) \geq 1 + \frac{\sum_{\tau=1}^{t-1} f_{\leq}(\tau, t)\tau^{-1}\sum_{\delta\in\mathcal{D}_t}\delta(t-\tau)p'(\delta, t)}{\sum_{\sigma=1}^{t-1} f_{\leq}(\sigma, t)\sigma^{-1}} \tag{4.18}$$

where p' is the marginal of p after summation over $\mathcal{F}_t$. Performing the sum over $\mathcal{D}_t$ yields

$$\langle\delta\rangle(t) \geq 1 + \frac{\sum_{\tau=1}^{t-1} f_{\leq}(\tau, t)\tau^{-1}\langle\delta(t-\tau)\rangle}{\sum_{\sigma=1}^{t-1} f_{\leq}(\sigma, t)\sigma^{-1}} \tag{4.19}$$

Thus a recursion for a lower bound on $\langle\delta\rangle$ is established.

Likewise, the age distribution

$$f_{\geq}(\tau,t) = \begin{cases} 2t^{-1}, & \text{if } \tau \leq \lfloor t/2 \rfloor \\ t^{-1}, & \text{if } \tau = (t+1)/2 \\ 0, & \text{otherwise} \end{cases} \tag{4.20}$$

can be used to establish an *upper-bound* recursion. Dynamically, this age distribution is obtained when an oldest leaf is chosen in each step. Again, it is assumed that the expected level does not increase if the age distribution f is replaced by $f_{\geq}$. Therefore,

$$\eta \geq 1 + \frac{\sum_{\tau=1}^{t-1} f_{\geq}(\tau,t)\tau^{-1}\delta(t-\tau)}{\sum_{\sigma=1}^{t-1} f_{\geq}(\sigma,t)\sigma^{-1}} \ . \tag{4.21}$$

By arguments analogous to the above, the upper-bound recursion can be formulated as

$$\langle\delta\rangle(t) \geq 1 + \frac{\sum_{\tau=1}^{t-1} f_{\geq}(\tau,t)\tau^{-1}\langle\delta(t-\tau)\rangle}{\sum_{\sigma=1}^{t-1} f_{\geq}(\sigma,t)\sigma^{-1}} \ . \tag{4.22}$$

For transforming $\langle\delta\rangle$ into expected depth d, consider the sum of distances of leaves from root, $D(t) = td(t)$. Addition of two leaves at distance x from root increases D by $2x-(x-1) = x+1$. Thus

$$D(t) = \sum_{s=2}^{t}(\delta(s)+1) \tag{4.23}$$

for a realization of the stochastic process with level sequence δ. By linearity of expectation values the expected depth is

$$\langle d(t)\rangle = \sum_{s=2}^{t}[\langle\delta(s)\rangle + 1]/t \ . \tag{4.24}$$

As pointed out previously, assumptions for deriving the lower and upper bound recursions, i.e., the replacement of the actual age distribution by an extreme case does not decrease the depth, does not hold in general. This shown by means of the following example. A comb tree, of a sufficiently tree size $n \geq 6$, is growing by speciation of the "right" leaf in a cherry. But before the last splitting at $t = 6$ the left child node of the root splits which results in $\delta(4) = 2$. At time step $t = 6$ one can obtain the ages one and two, twice each, and three and four. With replacing f by $f_{\leq}$ and applying Equation 4.14 both ages of two are replaced by two and five. This implicates an increase of η in Equation 4.14 and thus, the estimation in Equation 4.16 is not fulfilled.

Figure 4.8 shows upper and lower bounds on the expected depth $\langle d\rangle$ obtained as numerical solutions of the recursion Equations 4.19 and 4.22. The same diagram contains a plot of the results of direct simulations of the model. In one set of simulations the usual time increment $\Delta t = 1$ is used, so that $t \sim n$. Another set of simulations is performed with $\Delta t = 1/n$ to check

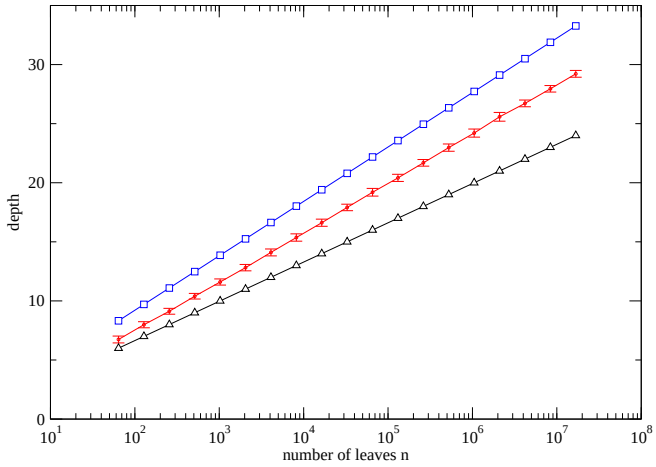


Figure 4.9.: Average depth in dependence of the number of leaves n in trees generated with stochastic loss events (dots with error bars). Each data point is an average over 100 realizations with error bars indicating standard deviations. for comparison, the expected depth for the ERM model (□) and for complete binary trees (△) are shown.

for robustness under different evolution of overall speciation rates. Upper and lower bounds as well as the two sets of simulations strongly suggest that the asymptotic growth behavior for the depth is $(\log n)^2$. An overview of the scaling of the average depth with the number of leaves is given in Table 3.1 (p.29) for different kind of models including the age model.

4.4.4. Mean Depth Scaling of the Innovation Model

The first part considers the deterministic tree growth as an approximation of the innovation model. The given preliminary conditions are essentiel for an approximation of the depth scaling for the innovation model, which is discussed in the second part of this section. There, it is shown that for a tree with n leaves generated by the innovation model, the average distance of leaves from root scales as $(\log n)^2$ to be compared to $\log n$ for random branching.

Preliminary conditions

For calculating the average depth of a tree, one may focus on the subtree generated by an innovation. Supposed that the i-th innovation, generating feature i, affects a species s with f features. Then s is removed from the set S of extant species, turning into an inner node in the tree. Two new species s' and s'' are attached, having feature sets $s' = s$ and $s'' = \{i\} \cup s$. In subsequent loss events, a subtree T_i is built up with 2^f leaves, each of which is a species

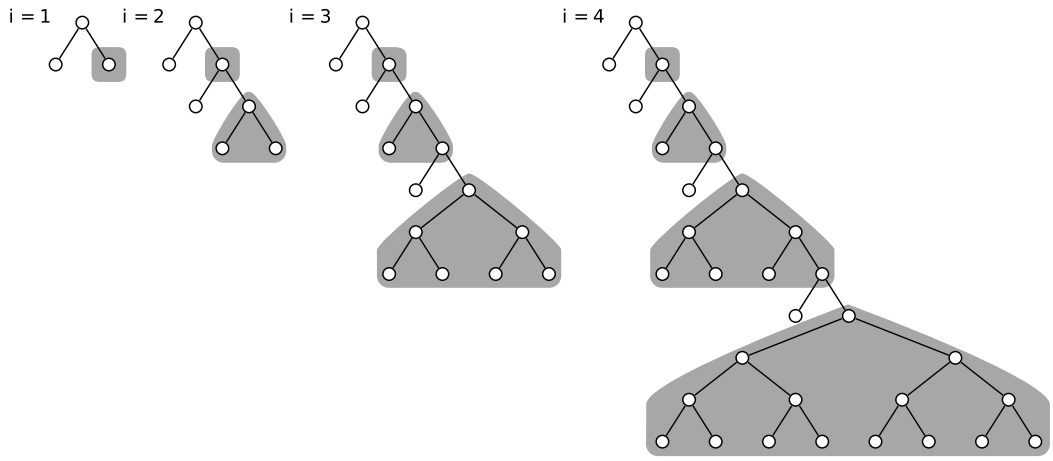


Figure 4.10.: The deterministic growth of a tree considered as an approximation of the innovation model. Each subtree generated by an innovation is indicated as a shaded area.

$\sigma \subseteq s \cup \{i\}$. Call $D(T_i)$ sum of the distances of all the leaves in T_i from the root of T_i.

Now, one can estimate the expectation value $\langle D(T_i)\rangle$, which only depends on the number of features of f. Trivially, $D(T_i)$ is lower bounded by $f2^f$ since the most compact tree is the complete balanced one with all nodes at distance f from root. In particular, we conjecture

$$f2^f < \langle D(T_i)\rangle < D_{\mathrm{ERM}}(2^f) \tag{4.25}$$

with the number of distances for the ERM model D_{ERM}. The second inequality is corroborated by the plots in Figure 4.9. This can be made plausible as follows. Similar to the ERM model, a leaf is chosen in each time step when executing loss events. Here, however, the loss event is performed only if the chosen leaf carries the chosen feature and the reduced feature set is not yet present in the tree. Thus the probability of accepting a proposed loss event at a leaf s is anticorrelated with the number of features $|s|$ at s. The expected number of features carried by a leaf decreases with its distance from root. Therefore, one can argue that the present model adds new nodes preferentially to leaves closer to root than average, resulting in trees with an expected depth increasing more slowly than in the ERM model.

Approximation

The study of tree growth is derived from the innovation model by two simplifying assumptions:

(i) Each innovation is introduced at the leaf with the largest number of features in the tree.

(ii) Introducing an innovation at a leaf with f features triggers the growth of a subtree that is a perfect (complete) binary tree with 2^f leaves at distance f from the root of this subtree.

This leads to the consideration of the following *deterministic growth* starting with a single node and $i = 0$. Choose a leaf s at maximum distance from root; split s obtaining new leaves s' and s''; take s'' as the root of a newly added subtree that is a perfect tree with 2^i leaves; increase i by one and iterate. Figure 4.10 illustrates the first few steps of the growth.

After i steps, the number of leaves added to the tree most recently is 2^{i-1}. Therefore, the total number of leaves after step i is

$$n(i) = 1 + \sum_{j=1}^{i} 2^{j-1} = 2^i \tag{4.26}$$

because the procedure starts with a single leaf at $i = 0$.

The leaves of the subtree added by the j-th innovation have distance

$$\sum_{k=1}^{j} k = \frac{j(j+1)}{2} \tag{4.27}$$

from root because these leaves are j levels deeper than those generated by the previous innovation. Therefore, the sum of all leaves' distances from root is

$$D(i) = i + \sum_{j=1}^{n} 2^{j-1}[j(j+1)/2] \tag{4.28}$$

after the i-th innovation has been performed. The first term i arises because the innovation itself renders one previously existing leaf at a distance increased by one, cf. the leaves outside the shaded areas in Figure 4.10. In performing the sum of Equation 4.28 the following equality is used

$$\sum_{j=0}^{i} x^{j-1}[j(j+1)] = 2^i[i^2 - i + 2] - 2 \tag{4.29}$$

to arrive at

$$D(i) = i + 2^{i-1}[i^2 - i + 2] - 1 \ . \tag{4.30}$$

A substitution of $n(i) = 2^i$, i.e. $i = \log_2 n$, and the division of D by n is performed to arrive at the depth

$$d(n) = \frac{1}{2}[(\log_2 n)^2 - (\log_2 n) + 2] + \frac{(\log_2 n) - 1}{n} \tag{4.31}$$

of the tree with n leaves generated by deterministic growth. For large n, the depth scaling is

$$d(n) \sim (\log n)^2 \ . \tag{4.32}$$

By the comparison in Fig. 4.11, the $(\log n)^2$ scaling is also found for the depth of trees obtained from the innovation model as defined in Section 4.2. Thus one can hypothesize that

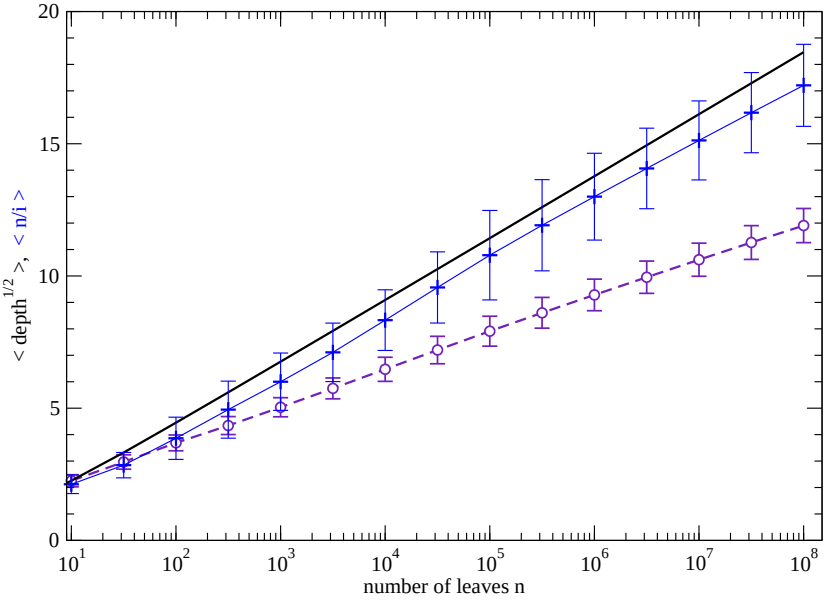


Figure 4.11.: Depth as a function of tree size n for the innovation model ($\circ$) and for the deterministic growth (thick solid curve) according to Equation 4.31. Note that square root of depth is plotted such that a straight line in the plot indicates a depth scaling $d(n) \sim (\log n)^2$. Small symbols ($+$) connected by thin lines give $\langle n/i \rangle$, the average number of leaves per innovation. For each size n, the plotted points ($\circ$, $+$) are averages over $\sqrt{d(n)}$ and i/n for 100 independently generated trees. Error bars give the standard deviation.

the deterministic growth captures the essential mechanism leading to the depth scaling of the innovation model. The prefactor of $(\log n)^2$ is smaller in the innovation model than in the deterministic growth. In the actual model, most innovations hit a leaf with a non-maximal number of features and therefore, trigger the growth of a lower subtree than assumed by deterministic growth. Table 3.1 (p.29) provides an overview of the scaling of average depth with the number of leaves for various tree models .

4.5. Discussion and Concluding Remarks on Age Model and Innovation Model

The proposed *age model* compares with observed phylogenetic trees better than previous models. In addition, it describes the tree generation process in a way which is easy to interpret phylogenetically: it assumes that lineages which have not speciated for a long time would display in the future a still more reduced speciation rate.

Future work should provide a more detailed analysis of the model itself and further compar-

ison to real phylogenetic trees. The depth scaling analysis is a challenging problem. Heuristic arguments suggest an depth scaling $d \sim (\log n)^2$ for the age model. But is is pointed out that the proposed analytic solution for the expected depth of its bounds shows deficiencies. It would be also desirable to obtain expressions or at least numerical results for the second and perhaps higher moments. For the likelihood expressions, a factorization or other kind of decomposition would allow for faster exact computation. Instead of exact computation, estimation by a Monte-Carlo sampling method may circumvent the present size limitation of trees in the likelihood analysis.

An additional interesting point of analysis and comparison of phylogenetic trees is the distribution of branch lengths. Branch length data, however, are not as reliable as the topological structure of phylogenetic trees (Barraclough and Nee, 2001). This argument is supported by Pigolotti *et al.* (2005), summarizing the variety of behaviors of distributions found in the literature. There is evidence to suggest that future studies in the line of Venditti *et al.* (2009) may accumulate sufficiently reliable branch-length data to allow for comparison to models such as the present one.

Finally, timing in the model is worth further clarification. The model describes tree growth as a Markov chain where exactly one speciation event occurs at each time step. A more realistic version would formulate a Markov process that assigns a speciation rate to each species at any moment in continuous time. The choice $\Delta t = 1/n$ in the results of Fig. 4.8 (p.49) is a first step in that direction.

The *innovation model* establishes a connection between the burstiness of macroevolution and the observed imbalance of phylogenetic trees. Bursts of diversification are triggered by generation of new features and combination with the repertoire of existing traits. In order to keep the model simple, the diversification after an innovation is implemented as a sequence of random losses of features. More realistic versions of the model could be studied where combinations of traits are enriched by re-activation of previously silenced traits or horizontal transfer between species. Furthermore, the model as presented here neglects the extinction of species and their influence on the shapes of phylogenetic trees. As the age model the innovation model also produces trees which fit well the oberserved tree structue in estimated trees.

Regarding the robustness of the model, the depth scaling would have to be tested under modifications. In particular, the infinite time scale separation between rare innovations and frequent loss events could be given up by allowing innovations to occur at a finite rate set as a parameter.

In summary, with the innovation model a well-working, biologically motivated model is defined which nevertheless is sufficiently simple to allow for further enhancement regarding biological concepts such as sequence evolution and genotype-phenotype relations.

CHAPTER 5

Likelihood Analysis For Growth Models

A further study on models of tree growth can be performed by the use of a likelihood analysis in order to rank models with respect to their ability to explain observed tree shapes. Thus, when abstracting from the algorithmic formulation of tree generation, a branching model A can be characterized by the probability $L_A(T)$ of obtaining a given tree T. The quantity $L_A(T)$ is also called the *likelihood* of model A under the data (tree) T. When aiming at modeling empirical data, it is supposed that one model A is better than a different model B, if

$$L_A(T) > L_B(T) \tag{5.1}$$

for an observed tree T. Let T be a rooted binary tree, the permutation of inner nodes leading to T is defined as a *branching sequence*. In that process, a node branches at one time step and children cannot branch earlier than their parents. Results of the likelihood analysis for small trees (up to 19 leaves), discussed in Section 5.1.2, show that the age model performs at least as good as the AB model for `TreeBASE` data and better for `PANDIT` data. To support this observation a further analysis using larger trees is necessary. But calculating the exact likelihood for the age model, one needs to sum up the probabilites for each possible branching sequence. Summing up the probabilites of all possible branching sequences implies a huge computational effort because of the exponential increase of branching sequences with tree size. To this effect, a sampling of branching sequences comes in to consideration for estimating the likelihood. Section 5.2 deals with a naive method and an effective method based on importance sampling (Ripley, 1987) for estimating the likelihood. Both methods are compared by means of a tree from `TreeBASE`.

5.1. Exact Likelihood Computation

The way of computing the likelihood needs to be considered individually for each growth model. As it is simple for both beta-splitting models, ERM model and AB model, it is not as easy for the age model. The likelihood computation of the mentioned models is defined in Section 5.1.1 , respectively Section 5.1.2.

5.1.1. Simple Calculation For ERM Model and AB Model

For the ERM model and the AB models (see Chapter 4 (p.31)), the computation of the likelihood is straight-forward:

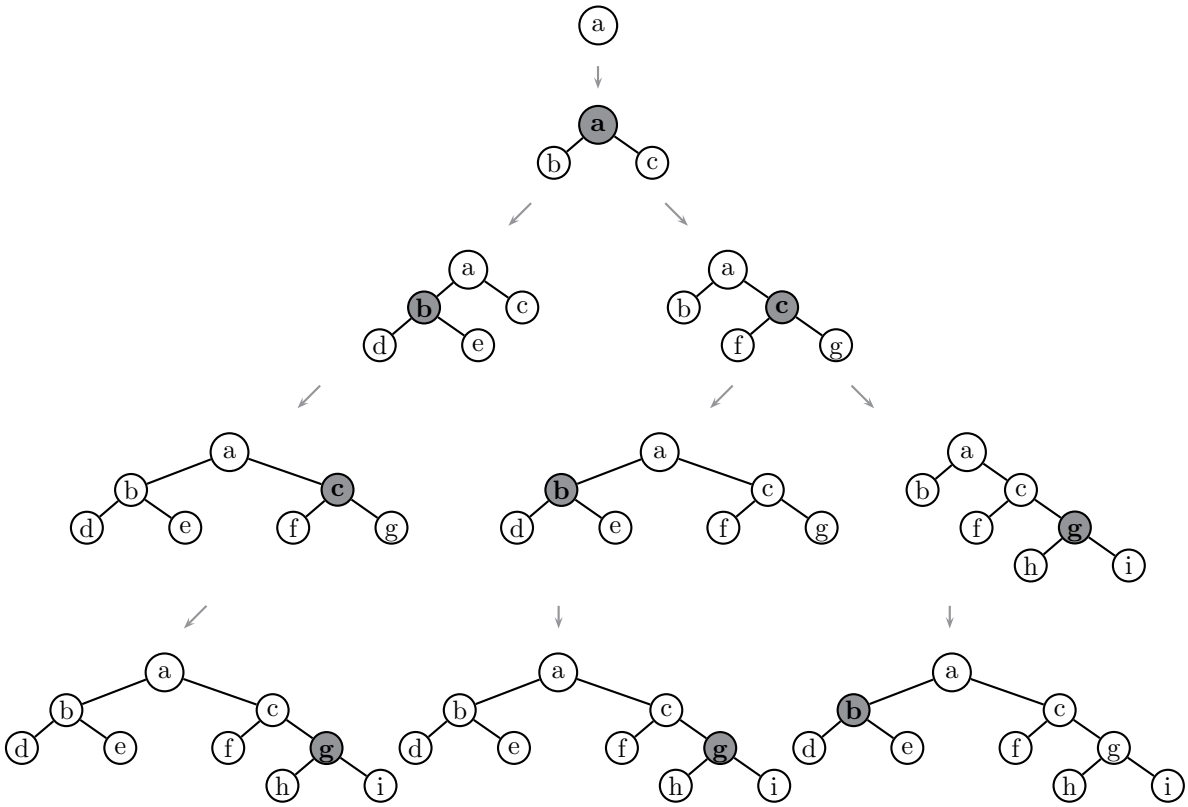
$$L_A(T) = \sum_{x \in I(T)} p_{\mathrm{ERM/AB}}(s(\mathrm{left}(x))|s(x)) \tag{5.2}$$

where A is the model under consideration, $I(T)$ is the set of inner nodes of tree T, $s(x)$ is the number of leaves in the subtree with the root x and $\mathrm{left}(x)$ is the left child node of node x. For the age model it is not clear if a simple method of exact likelihood computation exists. At the moment, the exact value of $L_{\mathrm{age}}(T)$ is calculated by adding up probabilities of *all* branching orders leading to the observed tree T. Details are described in section 5.1.

5.1.2. Likelihood Computation for Age Model

The computation of the likelihood $L_{\mathrm{age}}(T)$ of the age model generating a given rooted binary tree shape T is not as straight-forward as for the ERM model and AB model and can be calculated as follows. The nodes of the tree are assigned unique labels in $A := \{1, \ldots, 2n-1\}$, where the inner nodes have the labels $I := \{1, \ldots, n-1\}$. The root has label 1. For a non-root node $i > 1$, the unique parent node is denoted by $m(i)$. The set of all permutations of I is called S, so each element of S is a bijection $s : I \to I$. Such a permutation is to encode a branching order of a tree: $s(i)$ is the time step at which node i branches. In a valid branching order, children cannot branch earlier than the parent. Thus, one can say that $s \in S$ is *compatible* with T, if $s(i) > s(m(i))$ for each $i \in I \setminus \{1\}$. Let $S_c(T) \subseteq S$ be the set of compatible permutations. When branching according to $s \in S_c(T)$, the set of leaves at time $t > 1$ is

$$\begin{aligned} B(s,t) &= \{j \in I \setminus \{1\} \mid s(m(j)) < t < s(j)\} \\ &\quad \cup \{j \in A \setminus I \mid s(m(j)) < t\} \ . \end{aligned} \tag{5.3}$$



$$L_{\mathrm{AGE}}(T) = \sum_{s \in S_c(t)} p(s,T) = p\left((b,c,g),T\right) + p\left((c,b,g),T\right) + p\left((c,g,b),T\right)$$

Figure 5.1.: Branching orders for tree T with five leaves leading to the same tree topology. Gray nodes represent the speciating nodes. The likelihood of T under the age model is computet by adding up the probability for each of the three branching sequences according to Equation 5.5.

The age of a leaf j at time $t > 1$ is $t - m(j)$. Thus, the age model generates the tree T with the branching order given by $s \in S_c$ with probability

$$p(s,T) = \prod_{i=2}^{n-1} \frac{(s(i) - s(m(i))^{-1}}{\sum_{j \in B(s,s(i))} (s(i) - s(m(j))^{-1}} \ . \tag{5.4}$$

The overall probability of generating T with the age model is obtained by summing over all branching orders generating T,

$$L_{\mathrm{age}}(T) = \sum_{s \in S_c(t)} p(s,T) \ . \tag{5.5}$$

As pointed out, the age model is obtained by summing over all branching orders leading to an observed tree T. But this implies a huge computational effort due to an exponential

increase of branching orders with tree size. A method for likelihood sampling is essential and discussed in the next Section.

5.2. Likelihood Estimation for Certain Growth Models

Since the number of possible ways of generating an observed tree T grows exponentially, the multiplication along each path and summing over each of them for the likelihood estimation of T is time consuming. Hence an exact likelihood calculation is only feasible for small trees and implies a huge computational effort otherwise. Thus, which method allows to calculate the likelihood for large trees? An approach is the estimation of the likelihood for large trees by sampling with an equal probability over all possible branching sequences which lead to T, in the following called naive sampling method. More precisely, the likelihood is computed by the average of the likelihoods along each path of the sample. But one must take into account that branching sequences have different probabilites which results in an improper likelihood. Moreover the necessary amount of samples for a significant likelihood is unknown. Another option for a likelihood estimation would be to sample over the most probable branching sequences using a general efficient Monte-Carlo method instead of factorizing over branches. The idea of considering the most probable branching sequences is based on the method of *importance sampling* (Ripley, 1987). Importance sampling is a technique for reducing the variance of estimates. This is done by drawing attention to the values with an higher impact of a set of random variables in a simulation. Thus, importance sampling attends to approach quantities for a variety of applications where the computation of exact results is restricted or difficult to obtain (Wiuf *et al.*, 2006). Keeping in mind that trees are a special kind of networks, different ways of likelihood computation using importance sampling to ascertain how well a network growth models fits to data were already proposed by Wiuf *et al.* (2006) and Guetz and Holmes (2010). At this point, it is referred to Ripley (1987); Liu (2008) for more details on importance sampling.

Now, starting with a formal introduction of calculating the likelihood leads to an efficient sampling method which considers the most probable branches. Therefore, let $n > 1$ be a natural number and $X_1, \ldots, X_n$ finite sets (of states) with $|X_1| = 1$. Considering a stochastic dynamics that starts at the unique state in X_1 at time $t = 1$ and makes a transition to one state of the next set at each time step. After $n - 1$ steps, the process stops at a state in X_n. The dynamics, henceforth called p-dynamics is given by the transition probabilities

$$p_i(y|x),\ i \in \{1, \ldots n-1\},\ x \in X_i,\ y \in X_{i+1}\ . \tag{5.6}$$

A trajectory is a sequence $\Theta \in X_1 \times X_2 \times \cdots \times X_n =: X$. The probability R with which the system produces trajectory Θ is simply the product over the transition probabilities ("from

one state to another")

$$R(\Theta) = \prod_{i=1}^{n-1} p_i(\Theta_{i+1}|\Theta_i) \ . \tag{5.7}$$

Of interest is the probability L that the system ends up in a given set of target states $Z_n \subseteq X_n$. In terms of trajectories, L is the sum of probabilities over all trajectories that end up in a state $\Theta_n \in Z_n$

$$L = \sum_{\Theta \in X, \Theta_n \in Z_n} R(\Theta) \ . \tag{5.8}$$

In the application to growing binary trees, non-vanishing transition probabilities are sparse and from many intermediate states the target state set is not reachable at all. For $1 \leq i < n$, the set

$$Z_i = \{x \in X_N | \exists y \in Z_{i+1} : p(y|x) > 0\} \tag{5.9}$$

is defined iteratively and contains the states from which Z_n is reachable. Restricting the summation in Equation 5.8 to trajectories with non-zero probabilities, one can write

$$L = \sum_{\Theta \in Z} R(\Theta) \tag{5.10}$$

with $Z = Z_1 \times Z_2 \times \cdots \times Z_n$.

If $|Z|$ is too large to perform the sum explicitly, sampling may be employed. In many cases of interest, however, the distribution of trajectory probabilities is very broad: There are many trajectories with negligible probability while the value of L is determined by a few trajectories with relatively large probability. Then sampling each trajectory with equal probability does not yield good convergence.

By introducing a different sampling procedure, the likelihood calculation may be more efficient in such difficult cases. Starting with the introduction of the q-dynamics restricted to the sets $Z_1, \ldots Z_n$. The q-dynamics is also depicted in Figure 5.2. The transition probabilities are

$$q_i(y|x) = \frac{p_i(y|x)}{s(x)}, \ i \in \{1, \ldots n-1\}, \ x \in Z_i, \ y \in Z_{i+1} \tag{5.11}$$

with the normalization

$$s(x) = \sum_{y \in Z_i + 1} p(y|x) \ . \tag{5.12}$$

In this stochastic process, a trajectory $\Theta \in Z$ has the probability

$$S(\Theta) = \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i) \ . \tag{5.13}$$

To each trajectory $\Theta \in Z$ an *output* is assigned

$$A(\Theta) = \prod_{i=1}^{n-1} s(\Theta_i) \ . \tag{5.14}$$

The expectation value of A over trajectories under q-dynamics is the probability L that the p-dynamics ends up in the target set Z_n, as shown by the following sequence of term replacments.

$$\begin{aligned}
\langle A \rangle &= \sum_{\Theta \in Z} S(\Theta) A(\Theta) && (5.15)\\
&= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i) \prod_{j=1}^{n-1} s(\Theta_j) && (5.16)\\
&= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i) s(\Theta_i) && (5.17)\\
&= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} p_i(\Theta_{i+1}|\Theta_i) && (5.18)\\
&= \sum_{\Theta \in Z} R(\Theta) && (5.19)\\
&= \sum_{\Theta \in X} R(\Theta) && (5.20)\\
&= L && (5.21)
\end{aligned}$$

Thus L can be approximated as an average of A over sufficiently many trajectories generated with q-dynamics. This approximation method is applicable, if for each $i \in \{1, \ldots, n-1\}$ and all states $x \in X_i$

1. it can be decided efficiently (fast) if $x \in Z_i$ or not,
2. the normalization $s(x)$ can be computed efficiently.

5.3. Results and Discussion on Likelihood Analysis

For the likelihood analysis based on the exact likelihood calculation, the age model and AB model were compared under the tree shapes on small and medium-sized trees with up to 19 leaves in the databases `TreeBASE`, `PANDIT` and `McPeek`. Figure 5.3 shows that the likelihoods of the age model and AB model are clearly correlated under the trees in the databases. The variation of likelihoods across trees of the same size n is smaller in the age model compared to that in the AB model. Notably, the age model has larger likelihood than the AB model under more than half of the trees under consideration for `PANDIT`, so that it can be considered a

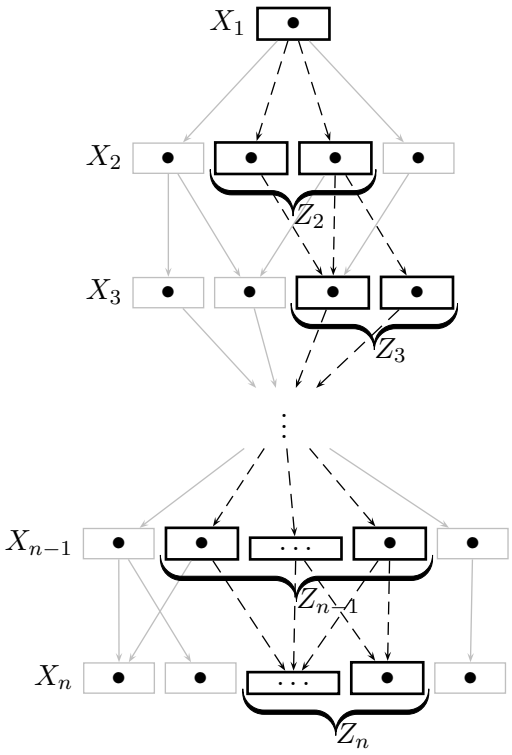


Figure 5.2.: Visualization of dynamics of the systems. Starting at unique state X_1 of size one, the number of trajectories leading to Z_n (black boxes), and thus leading to T, is of main interest. For each level i, Z_i contains the states from which the target states Z_n is reachable. But from a large number of intermediate states (gray boxes), $Z_n \subseteq X_n$ can not be accessed. For an efficient estimation of the likelihood for the systems reaching the set of target states Z_n, one starts with states in Z_n. Going backwards in the system, only transition probabilities leading from Z_i to Z_{i-1} will be considered for likelihood computation.

better description of the evolutionary process. But one can observe larger likelihoods for the AB model than for the age model for more than half of trees of `TreeBASE`. In case of `McPeek` the AB model outperforms the age model as well but one must take the small amount of trees for each tree size into consideration which may lead to biased results. Thus for example, the number of trees of size 15 to 19 reaches from three to nine.

The results of the likelihood analysis for trees up to 19 leaves have shown that the age model performs slightly worse as the AB model for `TreeBASE` data and worse for `McPeek` data but better for `PANDIT` data. To support this observation a further analysis using larger trees is necessary. But the overall probability of generating T with the age model is obtained by summing over all branching orders leading to the observed tree T. This implies a huge computational effort due to an exponential increase of branching orders with tree size. Two ways for sampling were proposed in Section 5.2. One is easy to apply to models since it assumes that each branching order has the same probability. For the efficient likelihood estimation

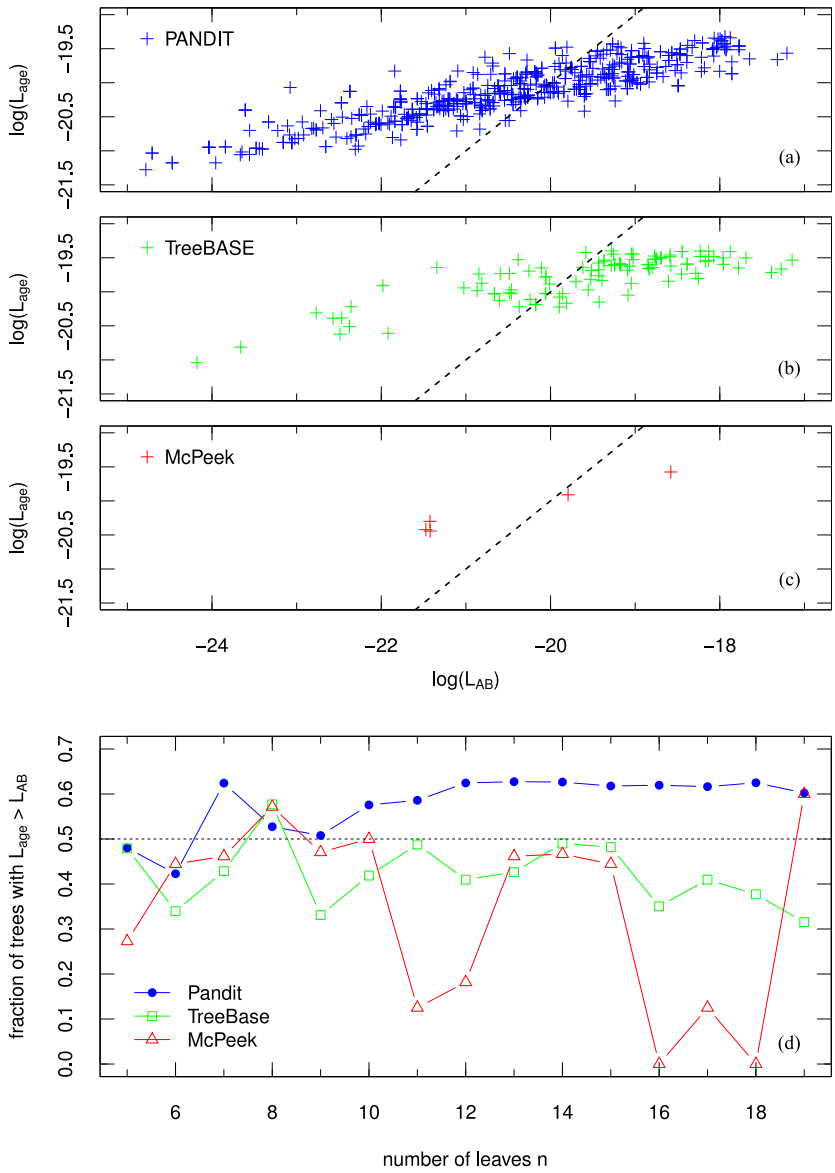
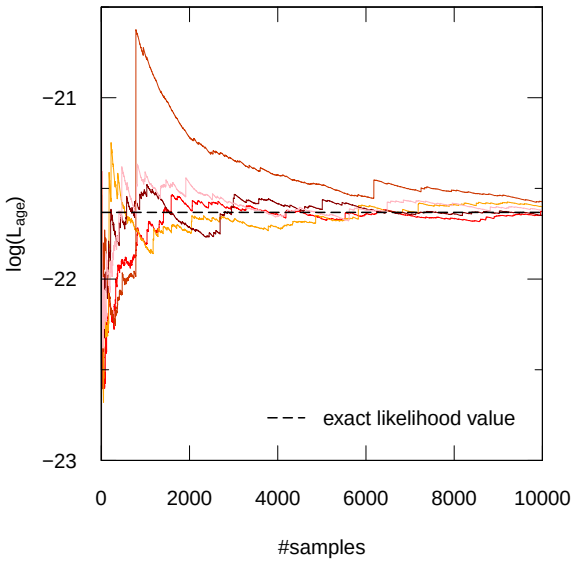


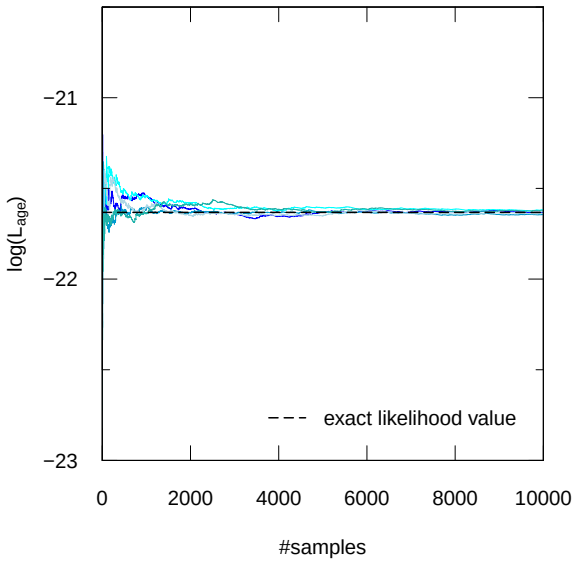
Figure 5.3.: Comparison between age and AB model by likelihoods under tree shapes from databases. (a) $\log(L_{\text{age}}(T))$ versus $\log(L_{\text{AB}}(T))$ for each of the 538 tree shapes T with $n = 19$ leaves in the database PANDIT. The dashed line is the identity. (b) Same as (a) for the 111 tree shapes with $n = 19$ leaves in the database `TreeBASE`. (c) Same as (a) for five tree shapes with $n = 19$ leaves in the database `McPeek`. (d) Fraction of trees T with $L_{\text{age}}(T) > L_{\text{AB}}(T)$, separately for each $n \in \{5, \ldots, 19\}$. The overall fraction is $0.552 = 13,674/24,754$ for `PANDIT` and $0.415 = 605/1,458$ for `TreeBASE`respectively $0.382 = 58/152$ for `McPeek`. The number of available tree instances is one order of magnitude smaller in `TreeBASE` than in `PANDIT` leading to larger fluctuations in the `TreeBASE` results. The results for `McPeek` are based on a small data set which may lead to ambiguous conclusion.

method, the normalization in each step i is calculated by summing over the probabilites of all possible states x if $x \in Z_i$.

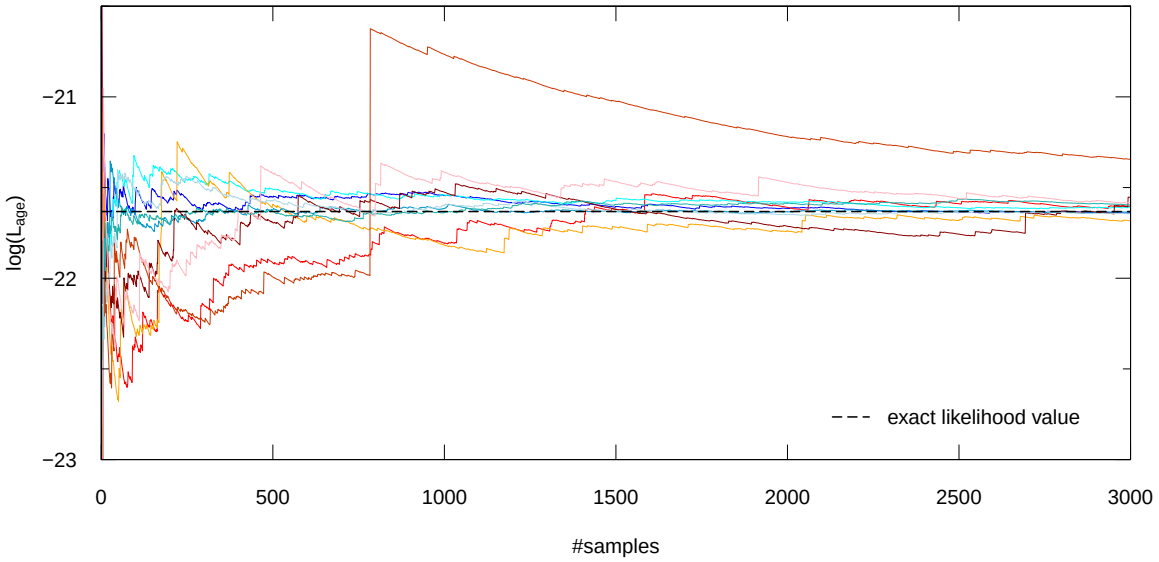
The likelihood was estimated using both methods for trees of `TreeBASE`. For each of the methods five runs with 10,000 samples each were computed. The likelihood for an example tree of size 19 chosen from `TreeBASE` (Matrix ID: M954) is given in Figure 5.4. For each run, Figure 5.4a visualizes the likelihood for each sample for the naive method respectively in Figure 5.4b for the effective method. One can obtain that the naive sampling results in large increasing and decreasing leaps until it converges to the value of the exact calculated likelihood. The curve with the highest peak in Figure 5.4a compared to the other samples shows clearly that the likelihood after a certain amount of samples does not need to converge to the exact likelihood. This is a problem since the number of needed samples for an estimated likelihood close to the exact likelihood is unknown. For the efficient likelihood estimation in Figure 5.4b all samples converge to the exact likelihood value with a smaller number of samples compared to the naive sampling method. Also the leaps can not be observed as strong as for the naive way of likelihood estimation. The difference in converging to the exact likelihood value between both method for each run up to 3,000 samples is again visualized in Figure 5.4.



(a) Five runs of naive likelihood sampling.



(b) Five runs of effective likelihood sampling.



(c) Five runs of effective (blueish) and naive (reddish) likelihood sampling with 3,000 samplings each.

Figure 5.4.: Comparison of likelihood estimation for a tree of TreeBASE (Matrix ID: M954; phylogenetic tree representing the relationship of symbiotic cyanobacteria and related species of the lichen fungus Peltigera (O'Brien *et al.*, 2005)) using the naive way of sampling (reddish curves) and the the effective method (blueish curves) by calculation the most probable branching sequences. The dashed line represents the exact value of the logarithmic likelihood of the tree using the age model. Five runs with 10,000 samples each were performed.

CHAPTER 6

Evaluating Host Parasite Reconciliation Methods Using The Age Model For Cophylogeny Generation

Coevolution between species is a common phenomenon in biology: species interact across groups such that the evolution of a species from one group can be triggered by a species from another group. Most prominent examples are systems of host species and their associated parasites. Typically in this field, phylogenetic trees for both groups of species can be constructed from sequence data or/and morphological data. In addition, the host parasite interactions between the extant taxa are known empirically. The problem is then to reconcile the common history of both groups of species and to predict the associations between ancestral hosts and their parasites. Some algorithmic methods have been developed in recent years to solve this reconciliation problem. Only few host parasite systems, however, have been analyzed in sufficient detail to serve as benchmarks for the evaluation of the reconstruction methods. In this chapter a dedicated approach for generating cophylogenies is introduced to tackle the lack of benchmarks by generating meaningful test data sets. The method builds on biologically motivated branching models to generate cophylogenies under the assumption of the widely used coevolutionary model. It pictures coevolution as a stochastic process with cospeciation, duplication, lineage sorting and (host) switching as discrete events. The probability of an independent parasite speciation as well as the ratio between cospeciations and sortings and between duplications and host switches are user defined parameters. Results on the evaluation of reasonable parameter settings under the aspect of producing realistic coevolutionary scenarios, giving rise to a large set of test scenarios, are discussed in the end of the chapter. Furthermore a detailed analysis and comparison of the common tools TreeMap 3b, Jane 2.0, and `CoRe-Pa` with a focus on the significance of the computed reconstructions is

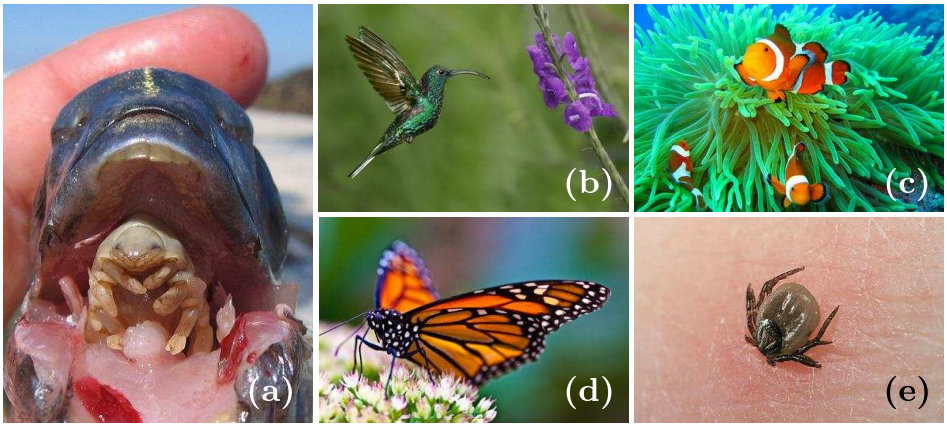


Figure 6.1.: Examples of coevolutionary systems on Earth. (a) Host parasite relationship between Ceratothoa imbricata and a fish[1]. The parasitic crustacean attaches itself on the fish tongue. There, it draws blood out of the tongue which then degenerates. The parasite stays as kind of tongue in the fish's mouth (Brusca and Gilligan, 1983). (b) Symbiotic relationship between Humming bird and ornithophilous flowers[2]. Birds serve as pollinators while flowers allocate nectar according to the birds diet. See (Stiles, 1981) for more details regarding the coevolution. (c) Symbiotic mutualism between Anemonefish and Sea anemone[3]. The coevolution of some clownfish species with certain sea anemones may have lead to an immunity of the fish to the toxins of the anemone (Mebs, 1994). (d) An insect-plant relationship as symbiosis between Monarch butterflies and Milkweed plant[4]. The Monarch only lay eggs on the milkweed plant from which its larvae feed become poisonous to other animals (Malcolm and Brower, 1989). (e) Parasite host relationship of Ixodida and host, here Human[5].

provided. All three tools are based on the maximum parsimony principle but using different heuristics and cost models.

A brief introduction to the field of cophylogenies and the necessity is given in Section 6.1 (p.69). In Section 6.2 (p.70) definitions such as phylogenetic trees in the sense of host and parasite trees are introduced as well as the usage of growth models for trees to accommodate a coevolutionary event model in order to generate cophylogenies. The principle of coevolution and the considered coevolutionary event model is explained as well. The chapter ends with a discussion on the properties of the resulting cophylogenies to assess their biological plausibility.

[1]Copyright by Nico Smit (Nico.Smit@nwu.ac.za)
[2]Copyright by frogger - Fotolia.com
[3]Copyright by TommySchultz - Fotolia.com
[4]Copyright by Jearu - Fotolia.com
[5]Copyright by Ste2.0 - Fotolia.com

6.1. Introduction

In the research field of phylogenetics, the recent advent of large genetic data sets offers increased insight into the evolutionary histories of species. Representations of such histories are phylogenies, which typically are binary trees with leaves corresponding to extant taxa and inner nodes representing ancestral species. In order to understand the driving forces of evolution leading to a high diversity of species, the reconstruction of phylogenies is inevitable. Statistical macroevolutionary growth models (see Chapter 4 (p.31)) are used to understand the dynamical rules of evolutionary processes such as the speciation and extinction. For example the Yule model (Yule, 1925) (Section 3.2.1 (p.21), the simplest model and generally referred to as the null hypothesis, describes a continuous-time branching process where each speciation is equally likely (Blum and François, 2006; Aldous, 2001). But the evolution of species cannot be understood as a closed system. Species are able to interact and may mutually affect their evolution. This can be described by the more complex problem of coevolution or cophylogenetics. Examples for coevolutionary systems are relationships between hosts and their associated parasites, between predators and prey, or between groups of species with symbiotic interactions. Some of those are depicted in Figure 6.1 . With a focus on the coevolution of parasites with their hosts, several methods have been proposed (Charleston and Perkins, 2006; Doyon *et al.*, 2011; Merkle and Middendorf, 2005; Merkle *et al.*, 2010; Ronquist, 1998; Conow *et al.*, 2010) to infer plausible cophylogenetic histories from given phylogenies for the host and the parasite species and an assignment of the extant parasite species to their hosts. Assessing the accuracy of these methods requires benchmarks, preferably based on empirically confirmed data of coevolutionary histories. However, such data are scarce. The main reason is that it is very difficult to get clear evidence about the former relations between the predecessors of the extant host and parasite species. Data from simulated coevolution might be able to fill the gap. A first step in this direction is taken with the proposal of a method for the generation of cophylogenies in Section 6.3 (p.73). Based on sets of cophylogenies that have been generated by this method, the accuracy of several cophylogeny reconstruction methods that have been proposed in the literature have been studied. For evaluation purpose Doyon *et al.* (2011) presented a simulation approach for coevolutionary scenarios. Therefore, an ultrametric tree (i.e., the host tree) was generated with a standard birth death process. Additionally the dependent tree (i.e., the parasite tree) was created by generating coevolutionary events according to a Poisson process with respect to the rates of the respective events. Unfortunately this approach requires a dating scheme of the independently generated host tree and biologically motivated estimations of the coevolutionary event rates. To avoid timing issues and evolutionary rates a new method of generating cophylogenetic scenarios can be used. Utilizing stochastic branching models like the ERM (Yule, 1925) or the age model (Keller-Schmidt *et al.*, 2010) presented in 3.2.1 (p.21) respectively 4.1 (p.32) the intention was to extend these models to produce evolutionary dependencies

between two simultaneously generated phylogenies. Such type of dependencies have been described in the well-known coevolutionary event-model (see, e.g. Charleston and Perkins (2006)). The branching models are used to generate binary trees iteratively by speciating a leaf chosen with a probability distribution given by the model. This process is combined with the four types of events that are typically used to describe host parasite coevolution, namely cospeciation, duplication, host switch, and sorting. A comparison of cophylogenies that have been generated by the proposed method using different growth models is performed with a focus on the proper choice for the parameter values of the generation model. Furthermore, generated pairs of phylogenetic trees consisting of a host tree and a parasite tree need to be compared in the context of a cophylogenetic analysis such that biologists are able to explore the relative rate of evolution with the knowledge about the coevolution of hosts and their parasites (Charleston and Perkins, 2006). Common cophylogenetic reconstruction methods are TreeMap 3b (Jackson and Charleston, 2004), Jane 2.0 (Conow *et al.*, 2010), and `CoRe-Pa` (Merkle *et al.*, 2010). These methods are evaluated with a focus on the significance of the reconstructions that they deliver for the test sets of cophylogenies that have been generated with the different dynamical branching models.

6.2. Basic Definitions on Cophylogenies

This chapter introduces basics which are necessary for the following chapter. The first part deals with the principle of maximum parsimony since it is a basic concept of common reconciliation methods. As introduced in Section 2.1 (p.9), phylogenetic trees describe the phylogenetic history between different organisms and are considered as rooted binary trees with inner nodes representing ancestral species and leaf nodes representing extant species. But species are able to interact and affect their evolution, i.e., the relationship between host and parasite. This is called coevolutionary system and the evolution can be represented as cophylogeny which will be discussed in the second part of this section. Finally, a brief introduction of all three tools is given.

6.2.1. The Principle of Maximum Parsimony

The principle of maximum parsimony in this context means reconstructing an optimal solution by minimizing the total number of evolutionary changes, i.e., amount of events, respectively the total costs which mapped to the events. The resulting solution is called *most parsimonious*. Originally, the method was establihed for molecular data (Hennig, 1966) and based on the assumption of minimal evolution. This criteria of optimality deduces from the so called *occam's razor* which states with the words of Einstein (about 1900) "Everything should be kept as simple as possible, but no simpler" . It is known in many research fields and for the reconstruction of phylogenetic trees it simply means that the easiest explanation for consistent

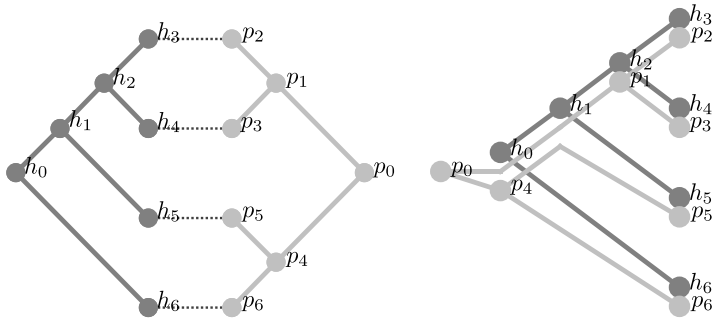


Figure 6.2.: Example for a coevolutionary system and a corresponding reconstruction. Left: Example for a small coevolutionary system with four extant host species (leaf nodes in dark grey tree) and four extant parasite species (leaf nodes in light grey tree). Right: Example of a cophylogenetic reconstruction for the coevolutionary system. The three associations (p_5, h_5), (p_6, h_6) and (p_4, h_0) induce one cospeciation and one sorting event. The three associations (p_1, h_2), (p_4, h_0), and (p_0, h_0) induce one duplication and two sorting events. The reconstruction needs two cospeciations, one duplication, and three sortings.

characters between species is a common ancestor.

6.2.2. Coevolution

Definition 6.1 (Cophylogeny, Coevolutionary System)**.** A *coevolutionary system*, also called *cophylogeny*, consists of two (coevolved) phylogenetic trees, a host tree T_h and a parasite tree T_p. It describes the interaction of species across groups such that the evolution of a species from one group, i.e., the parasite, developed in dependence from a species of another group, i.e., the host.

The coevolution of two groups of species is studied in order to explore the combined phylogenetic history. Therefore, the two phylogenetic trees T_h and T_p of both species are inferred. To this end, the observed host parasite associations in the extant species have to be known. Such associations are defined as follows:

Definition 6.2 (association)**.** The association between host h and parasite p can be seen as a relation ϕ between the different leaf sets, i.e., $\phi \subset L(T_p) \times L(T_h)$. Thereby it is assumed that one parasite species can be associated to at most one host species.

The latter assumption is widely used in the literature on algorithms for the analysis of coevolution. Note, however, that there are several empirical examples where this assumption does not hold. An example of an artificial coevolutionary system is given in Figure 6.2 (left). Figure 6.3 depicts an instance taken from the real life on earth. Shown is the relationship between parasitic primate lice and their vertebrate hosts studied by Reed *et al.* (2007).

A common approach for the reconstruction of cophylogenetic histories establishes a mapping from the parasite tree onto the host tree. In this way, ancestral dependencies between parasites

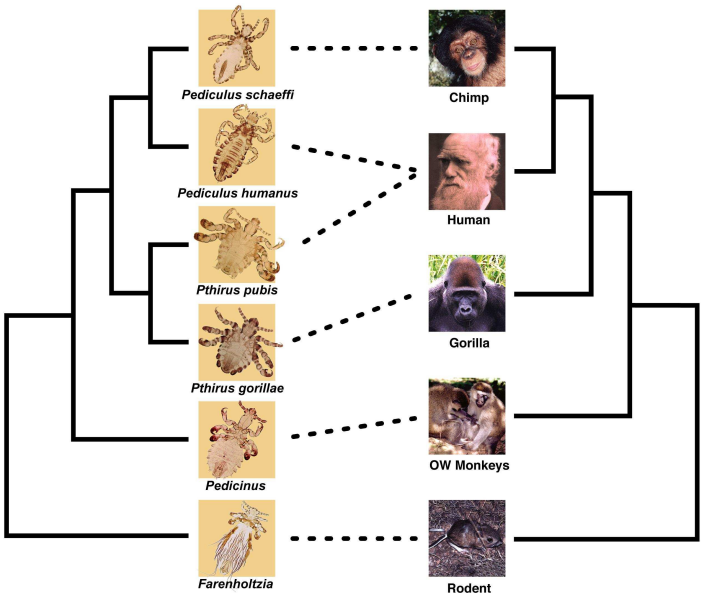


Figure 6.3.: Example for a real coevolutionary system of a host parasite relationship. The figure by Reed *et al.* (2007) represents the phylogenetic trees for parasitic primate lice and their vertebrate hosts. The tree is assembled by the use of morphological and genetic characters.

and their hosts are predicted. Coevolution is captured in terms of *events*. In this case four different types of events are employed and visualized in Figure 6.4.

Definition 6.3 (host dependent events: copseciation, sorting)**.** The events *cospeciation* (co) and *sorting* are host dependent and describe the reaction of a parasite if its associated host performs a speciation. In case of the cospeciation event, host and parasite speciate simultaneously. A sorting event describes the lineage sorting of a parasite across the speciation of its associated host. In this case, the parasite species remains on only one of the newly emerged host species.

Definition 6.4 (host independent events: duplication, host switching)**.** The remaining two events are host independent, namely, *duplication* (du) and *host switching* (sw) where the speciation of a parasite occurs without a speciation of an associated host. The duplication event describes the speciation of a parasite alone. The resulting two child species are associated to the same host as the parent species. A host switching event refers to a host shift of one of the parasite child species immediately after a speciation (Charleston, 1998).

To each of the four event types, a cost value is assigned taking into account the likelihood of the event. Less likely events incur larger cost. Using maximum parsimony a reconstruction is sought such that the total costs of all events that occur is minimal. Depending on the chosen event costs, i.e., the cost model, different reconstructions can be optimal. A reconstruction

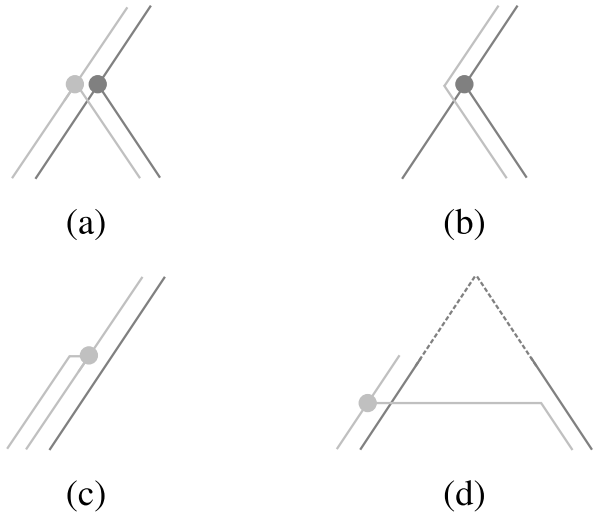


Figure 6.4.: Coevolutionary Events. Host tree T_h (dark gray), parasite tree T_p (light gray); (a) *cospeciation*: node of T_h and T_p associated; (b) *sorting*; (c) *duplication*: both child nodes of T_p are associated with a node in the same subtree of T_h; (d) *host switch*: only one child node of T_p is associated with a node in the respective subtree of T_h.

being optimal under a certain cost model is called a Pareto optimal solution of the coevolutionary system. *Pareto optimal solutions* are optimal compromises of two criteria where the enhencement of one criteria affects the other by decreasing its importance. Thus, if S is a Pareto optimal solution, there exists no other solutions $\overline{S}$ which performs better as S with respect to the same cost model (Beier *et al.*, 2007). In the scope of cophylogenies, S defines the reconstructed coevolution. An example of a reconstruction for the coevolutionary system depicted in Figure 6.2 (left) is given in Figure 6.2 (right).

6.3. The Generation of Cophylogenies

The common and new growth models introduced in Chapter 3 (p.19) respectively 4 (p.31) for phylogenetic tree generation can not directly be used for the generation of cophylogenies. The reason is that it is essential that the two phylogenetic trees are generated simultaneously with respect to the intended dependencies between the corresponding groups of species. Therefore, the aim here is to adopt common growth models to meet these demands.

6.3.1. The Model

The starting point is a host tree and a parasite tree, both consisting of a single node. Furthermore, the parasite node is associated with the host node. Then, one node is chosen for an upcoming speciation. If the selected node is from the parasite tree, this results in a host independent coevolutionary event (i.e., a duplication or host switch). Otherwise the event is

host dependent (i.e., a cospeciation or sorting). To decide which node is the next to speciate, a parameter p_{hc} is introduced, giving the probability that the node belongs to the host tree. The respective probability for selecting a parasite node is defined by

$$p_{pc} = 1 - p_{hc} \tag{6.1}$$

With this probability, only the type of the node (i.e., host or parasite) is chosen. The decision of which leaf in the host tree, respectively parasite tree, has to be done according to the considered branching model. Thus it is ensured that both created trees satisfy the particular branching model. Furthermore, in each step it is clear which are the current extant species. This information is needed later for producing time consistent host switching events, as a parasite can only switch to a host which existed at the same time.

To achieve the intended dependencies between host and parasite species, additional parameters have to be considered. These parameters p_{co}, p_{so}, p_{sw}, p_{du} define the probabilities for the respective coevolutionary events cospeciation, sorting, host switch, and duplication. Thereby it holds

$$p_{co} = 1 - p_{so} \tag{6.2}$$

$$p_{sw} = 1 - p_{du} \tag{6.3}$$

It can be seen that the probability for p_{so} respectively p_{du} can be inferred from p_{co} respectively p_{sw} using Equation 6.2 and 6.3. Therefore, these parameters can be obtained from the ratio between the event frequencies of the two host dependant (respectively the two host independent) event types. Compared to the approach presented in (Doyon *et al.*, 2011) it is easier to estimate these ratios than the true evolutionary rate for each of the events.

In case of a host dependent event occurring after a host node is chosen for speciation, for each associated parasite it has to be decided with probability p_{co} if the parasite speciates too, resulting in a cospeciation event. Otherwise a sorting event occurs and the parasite remains on only one of the newly emerged host children. The respective host child is selected randomly with an equal probability. In case of a host independent event after a parasite node is chosen for speciation, at least one of the child species remains on the same host species. The other child species can switch to a randomly selected host leaf with probability p_{sw} or otherwise remains on the same host species too.

In that way the generation of both trees T_h and T_p and their respective associations is done iteratively until there exists a given total number s of extant species, i.e., $s = |L(T_h)|+|L(T_p)|$. The pseudocode describing this method is shown in Alg. 5.

Algorithm 5: Pseudocode for the generation of a cophylogentic history.

Input: trees T_h, T_p each with only a single node, size s, probabilities p_{hc}, p_{co}, p_{du}, p_{sw}, p_{so}

Output: cophylogeny composed of a parasite tree T_p associated with a host tree T_h and $s = |L(T_h)| + |L(T_p)|$

```
1  while s < |L(T_h)| + |L(T_p)| do
2      with uniform probability chose r ∈ [0, 1];
3      if r ≤ p_hc then
4          choose leave l ∈ L(T_h) w.r.t. a branching model;
5          foreach parasite associated with host l do
6              with uniform probability chose r ∈ [0, 1];
7              if r ≤ p_co then
8                  do cospeciation;
9              else
10                 do sorting;
11     else
12         choose leave l ∈ L(T_p) w.r.t. a branching model;
13         with uniform probability chose r ∈ [0, 1];
14         if r ≤ p_sw then
15             do switch to a randomly selected host from L(T_h);
16         else
17             do duplication;
18     update T_h, T_p;
```

6.3.2. Properties of Generated Cophylogenies

It is obvious that not all combinations of parameter values for the proposed cophylogeny generation method lead to "relevant" cophylogenies. For example choosing $p_{hc} = 0$ will result in a single host node, as no host will ever be chosen for a speciation. Thus all parasite nodes will be associated with this single host. On the other hand if the probabilities p_{hc} and p_{so} are both 1 then only host nodes are chosen and the one associated parasite does always a sorting. This results in a parasite tree with a single node associated to one of the host leaves.

To decide whether a generated host parasite system is a "relevant" data set or not properties have to be found which describe if a certain cophylogeny is similar to real biological data. The number of empirically confirmed cophylogenies does not allow a meaningful statistical analysis on that. But the host parasite systems seems to have several things in common. Studies by (Charleston, 1998; Hafner and Nadler, 1988; Kikuchi *et al.*, 2009; Refrégier *et al.*, 2008; Reed *et al.*, 2007; Hughes *et al.*, 2007; Banks *et al.*, 2006; Ramsden *et al.*, 2009) have shown that the sizes of the two trees T_h and T_p differ only slightly in the way that T_p is somewhat larger. Additionally, there is usually no host taxa included which is not associated with at least one parasite. Also every host harbors approximately the same number of parasite species.

Thus, in order to evaluate the generated host parasite systems the following two characteristics are considered: The ratio between parasite tree size and host tree size and the variance of the number of associated parasites per host leaf. Generated cophylogenies with a size ratio close to 1 and variance close to 0 are considered to be more likely similar to biological cophylogenies.

Formally the ratio between the sizes of parasite and host tree (*scale*) is defined as

$$scale = \frac{|L(T_p)|}{|L(T_h)|} \tag{6.4}$$

The variance of the number of associated parasites (*var*) is defined as

$$var = \frac{\sum_{h_i \in L(T_h)} (x_{hi} - \mu)^2}{|L(T_h)|} \tag{6.5}$$

with x_{hi} being the number of parasites associated with host leaf h_i, i.e., $x_{hi} = |\{(p, h_i) \in \phi\}|$ and μ being the average number of associations per host leaf. Note, that $\mu = scale$ since assuming that each parasite leaf is associated with exactly one host leaf.

To compare cophylogenies of different sizes *scale* and *var* are normalized to range from -1 to 1, respectively 0 to 1. For this purpose cut off values of 1/10 and 10 were defined for *scale* such that a value of *scale* that is 10 or larger is rated 1 and a value of *scale* which is 1/10 or below is rated -1. Furthermore, a *scale* value of 1, i.e., equal size host and parasite trees,

should result in a normalized value of 0. The formal definition is given in Figure 6.6.

$$scale^* = \begin{cases} 1 & \text{if } scale > 10 \\ \frac{scale-1}{9} & \text{if } scale \geq 1 \wedge scale \leq 10 \\ \frac{-\frac{1}{scale}+1}{9} & \text{if } scale \geq \frac{1}{10} \wedge scale < 1 \\ -1 & \text{otherwise} \end{cases} \tag{6.6}$$

Accordingly a threshold of 10 is defined for *var* such that a variance of 10 or above results in a normalized value of 1. Equation 6.7 describes this normalization.

$$var^* = \begin{cases} 1 & \text{if } var > 10 \\ \frac{var}{10} & \text{otherwise} \end{cases} \tag{6.7}$$

A threshold of 10, respectively 1/10, was chosen, as this is the maximal, respectively minimal, value when considering cophylogeny systems of size 10, which are the smallest systems being analyzed in this study. Both normalizations result in a value of 0 in the best case, i.e., equal sized host and parasite trees, respectively equally distributed number of parasite associations. Conversely values of ± 1, respectively $+1$ indicates that a host parasite system is likely to be unrealistic.

To combine both measures $scale^*$ and var^* they are multiplicatively linked to obtain a *quality* value which is used as a measure of how likely a cophylogeny can be considered to be realistic. Formally, Equation 6.8 is defined by

$$quality = (1 - |scale^*|) * (1 - var^*) \tag{6.8}$$

6.4. Results

In the following, the space of parameter values for the cophylogeny generation method is analyzed in order to identify "good" sets of parameter values that lead to realistic cophylogenies. Then, an evaluation data set of cophylogenies is generated. This data set is used to evaluate the cophylogeny reconstructions that are delivered by the tools TreeMap 3b, Jane 2.0, and `CoRe-Pa`. The result of this evaluation is given at the end of this section.

6.4.1. Parameter Values

For the parameter evaluation, the modified ERM and the age model were used with the generation method to generate 100 cophylogenies for each combination of parameter values $s = \{10, 15, \ldots, 50\}$, $p_{hc} = \{0.0, 0.05, \ldots 0.95, 1.0\}$, $p_{co} = \{0.0, 0.05, \ldots 0.95, 1.0\}$, and $p_{sw} = \{0.0, 0.05, \ldots 0.3\}$. Only values up to 0.3 have been considered for p_{sw} because in typical biological host parasite systems it is much more likely for a parasite to remain on an associated host than to switch to another host. Moreover, a very high switching probability would mean

that there is only a very loose relation between hosts and their parasites. Those systems are not of high interest and thus, it is not essential to analyzed them by reconciliation methods. The cophylogenies generated with the different sets of parameter values have been evaluated with respect to $scale^*$, var^*, and *quality*. To analyze in more detail the influence of the system size cophylogenies have also been generated for size $s = 100$.

The *quality* of cophylogenies that have been generated with different combinations of parameters values p_{hc} and p_{co} and for different values of s are shown in Figure 6.5. Recall that a set of cophylogenies may be considered to be more realistic if i) both trees are of similar size ($scale^* \approx 0$), and ii) every host is associated with approximately the same number of parasites ($var^* \approx 0$). Cophylogenies where T_p equals T_h and each of the parasites is associated with the corresponding host can be generated using parameter values $p_{co} = 1$ and $p_{hc} = 1$. In this case no host independent events occur and there is always a cospeciation of the parasite whenever a host speciates. These perfect scenarios belong to the upper right corner of each of the *quality* plots given in Figure 6.5.

It comes as no surprise that independently of all other parameter values a value for p_{hc} of at least 0.4 is needed to obtain equal size trees. Otherwise, there will be too few host speciations resulting in very small host trees. By increasing the probability of cospeciations p_{co} the parasite tree becomes larger. Hence p_{hc} must be increased simultaneously in order to obtain the same results. Surprisingly there is nearly no influence of the switching probability p_{sw} and the system size s on the ratio of both tree sizes. On the other hand, the variance of the associations varies strongly depending on the system size. In general it holds that the larger the system size s is, the higher the host choosing probability p_{hc} has to be in order to obtain a small variance. Additionally, if a higher probability of cospeciations p_{co} is chosen then smaller values of p_{hc} are possible for producing quite reasonable variances. If a higher switching probability p_{sw} is used, p_{hc} can be decreased further while retaining a small variance.

Figure 6.5 shows that the range of "good" parameter values strongly depends on the system size. With an increasing system size, the range of parameters leading to realistic cophylogenies shrinks to the upper right quarter of the plot. Thus for systems with 50 or more leaves, the probability p_{co} should be at least 0.4. Choosing smaller values for p_{co} is not recommended, when considering highly dependent host parasite systems. Additionally, p_{hc} should be greater than 0.7. Otherwise, the variance becomes large. Surprisingly, there is only a small influence of the switching probability p_{sw} such that the ranges of "good" values for p_{hc} and p_{co} can be larger. This means that any of the considered p_{sw} values can be chosen to produce realistic cophylogenies.

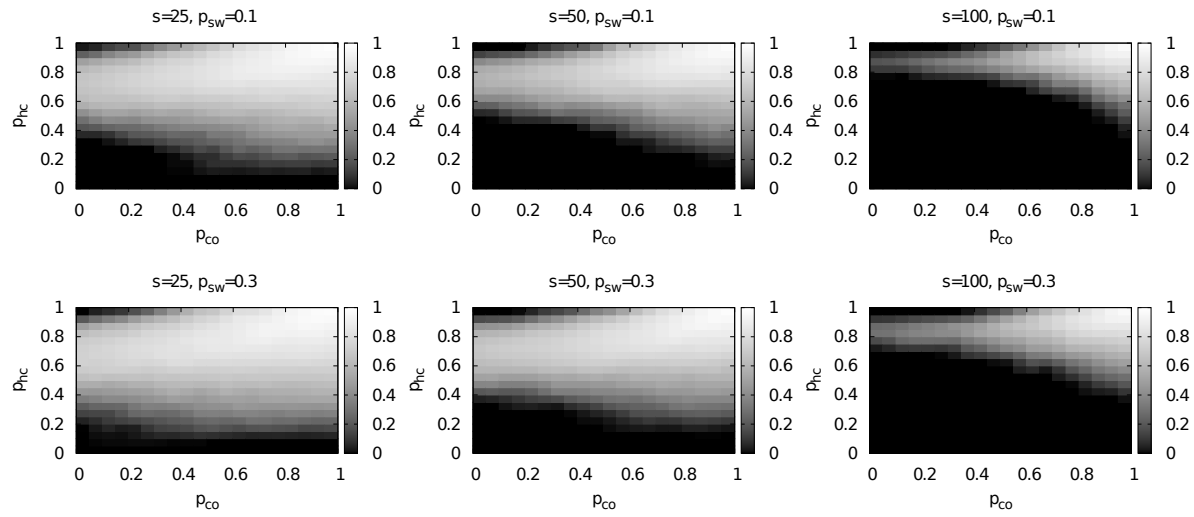


Figure 6.5.: Mean *quality* of 100 generated cophylogenies per parameter combination of p_{hc} and p_{co} for a tree sizes $s = (25, 50, 100)$ (left to right) and $p_{sw} = (0.1, 0.3)$ (top to bottom) for the age model. (See Appendix C.1,C.2 and C.3 for more results of further parameter combinations considering the age model as well as the ERM model.)

6.4.2. Evaluation of Reconciliation Methods

This section starts with an overview of common reconciliation methods which is followd by a definition of the test data set used for the analysis. The last section deals with the reconstruction of host parasite system using the different tools, Treemap, Jane 2.0 and `CoRe-Pa`.

Overview of Common Reconciliation Methods

In this part, the three software tools TreeMap 3b (Charleston, 2011), Jane 2.0 (Libeskind-Hadas, 2010) and `CoRe-Pa` (Wieseke *et al.*, 2010) are compared in terms of accuracy when reconstructing cophylogenetic histories for the generated test data. TreeMap is probably the most common tool for computing reconciliations of host parasite systems. It is now in its 3rd major release and was rewritten completely in Java. Jane 2.0 and `CoRe-Pa` are quite novel tools which offer several additional features. For instance, Jane 2.0 includes an advanced reconstruction viewer where the user can browse easily through all possible reconstructions. `CoRe-Pa` provides a graphical user interface for designing host parasite systems and is able to deal with non-binary species trees.

Although all three methods are based on the same coevolutionary event model, they differ in how the costs for each of the events are counted. This is due to the fact that in one approach the costs are counted per event while in the other they are counted per emerged sibling in the parasite tree. An overview on the different cost methods is given in Table 6.1.

[6]The cost method used by Jane 2.0 is the same as the one that was used in former versions of TreeMap; Jane

	Cospec.	Dupl.	Sorting	Switch
TreeMap 3b	$2c$	$2d$	s	$d + w$
Jane 2.0[6]	$2c$	$2d$	s	$2d + w$
`CoRe-Pa`	c	d	s	w

Table 6.1.: Different methods of costs assignments per event for the reconciliation methods considering specified cost values c, d, s and w

All three approaches are based on the maximum parsimony principle. Given a certain cost vector (i.e., a cost value for each type of event) the tools search for the reconstruction which results in the minimum total cost. For that reason, the resulting reconstructions depend highly on the used cost model. Jane 2.0 uses costs $c = 0$, $d = 1$, $s = 2$ and $w = 1$ by default. But as in all three applications the cost model can also be user specified. TreeMap 3b and `CoRe-Pa` offer more sophisticated methods to solve this issue. Since version 3b (build 1234), TreeMap uses a heuristic to find several reconstructions that are potentially optimal under a certain cost model. This set of so called Pareto optimal solutions may be huge and the reconstructions differ very much. So it is hard to decide for one of the reconstructions being the most likely. `CoRe-Pa` also tries to find all these Pareto optimal solutions by applying multiple cost models. In addition every reconstruction is then rated by a value which indicates how good a reconstruction fitted to the appropriate cost model.

The cophylogeny reconstruction problem is NP-hard (Ovadia *et al.*, 2011). Therefore, all tools use heuristics for the optimization. Only TreeMap gives the opportunity to search for an exact solution. Depending on the size of the host parasite system the computation can be time and space intense so that only small instances can be solved in reasonable time. By default TreeMap 3b uses a heuristic, too. While `CoRe-Pa` always finds an optimal solution, the reconstruction may be chronologically invalid, involving sets of inconsistent host switches. TreeMap 3b and Jane 2.0 always produce consistent though not necessarily optimal solutions.

Test Data Generation

To evaluate the reconciliation methods 1,000 test data sets per branching model are computed. The sizes of the generated cophylogenies and the other parameter values are chosen randomly with a distribution proportional to the *quality* gathered from the parameter space evaluation discussed in the previous section. In this way it is ensured that each combination of parameters can be selected, but it is more likely that parameters are chosen that will result in cophylogenies that are similar to cophylogenies that occur in biological systems.

For each model it is distinguished between the complete cophylogenies as they were gen-

2.0 can also be configured to count the costs in the same way as `CoRe-Pa` does.

erated and a pruned version. In this pruned cophylogenies host nodes are removed which have no assigned parasites. This was done due to the fact that most biological studies also disregarded hosts without associated parasites. So one might ask if this lack of information would have a measurable impact on the reconstructions. An overview of removed simulated systems which were not used for the evaluation is given in Appendix C.4.

This results in four test set-ups, one for each combination of ERM or age model with complete or pruned cophylogenies. But not each of the 1,000 generated cophylogenies per set-up could be considered for reconstruction. Due to the wide range of possible parameter values combinations 7% to 24% of the datasets were cophylogenies with one of the trees having less than 3 nodes. These trivial instances were not included in the analysis. Very few cophylogenies could not be processed with TreeMap 3b resulting in an "out of memory" error. These cophylogenies were also removed. Altogether between 771 to 920 cophylogenies were used per test set-up.

Reconstruction Evaluation

The reconstructions computed with TreeMap 3b were done with the default heuristic trying to find different Pareto optimal solutions. Although in some rare cases more than 500 different reconstructions were found for a single data set, the dominant number of computations (around 60%) produces only three or less distinct reconstructions. The command line version of Jane 2.0 was used with its default cost model, producing exactly one reconstruction per data set. `CoRe-Pa` was configured to evaluate 2,500 different cost models and the best rated reconstructions were considered for the analysis. In most cases (more than 90%) `CoRe-Pa` produced a single reconstruction. In the other cases up to seven different solutions were found, all having the same event distribution.

To measure the accuracy of each tool, the amount of correctly predicted host parasite associations were analyzed with respect to the generated cophylogenies. If more than one solution was found by one of the tools (TreeMap 3b or `CoRe-Pa`), the average amount of correct hits was taken. As TreeMap 3b tries to find different Pareto optimal reconstructions the solution with the highest, respectively lowest, number of hits were analyzed additionally. But it should be noted that for a determination the best (or worst) of the solution knowledge about the exact history is necessary (which will not be available for the application to biological data). For normalization purposes, the fraction of the exact hits compared to the total number of associations - including false positives and false negatives - was used.

Figure 6.6 depicts the strip chart of the sorted fraction values with one dot for each data set and method. Only the results of the complete age model data set are shown. For the results of the pruned cophylogenies and the ERM model it is referred to the Appendix C.7, as these are quite similar.

`CoRe-Pa` turns out to be the most precise method in this analysis. Depending on the

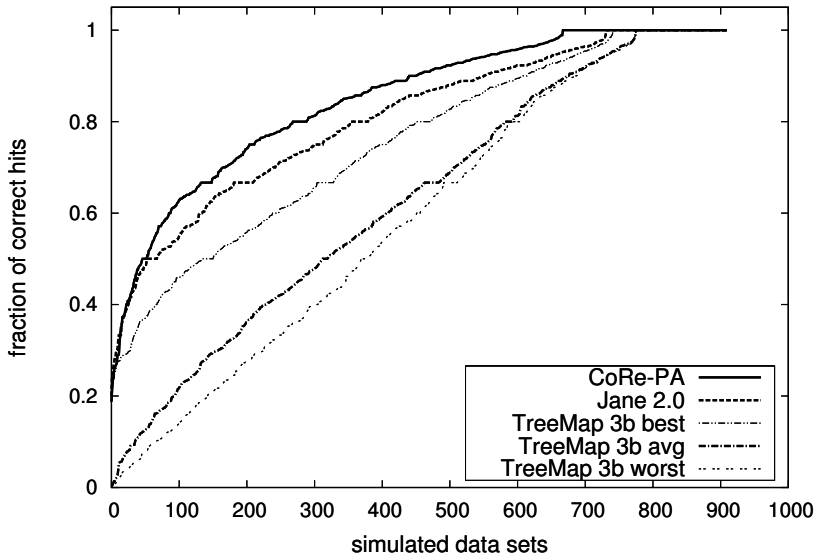


Figure 6.6.: Sorted fraction of exact predicted host parasite associations for each tool for the complete cophylogenies with age model data set.

used branching model it produces significantly more exact hits than Jane 2.0. It comes as no surprise that the average fraction of hits computed from the multiple solutions of TreeMap 3b is much lower. By considering multiple Pareto optimal solutions there are many reconstructions which differ very much from the corresponding generated cophyogeny. This obviously lowers the average fraction of hits. On the other hand one would assume that by considering only the most similar of these reconstructions the fraction of hits would be much better, especially because the solutions of Jane 2.0 and `CoRe-Pa` are Pareto optimal too. This leads to the assumption that the heuristic used within TreeMap 3b misses a significant amount of Pareto optimal solutions, not reaching the results of Jane 2.0 and `CoRe-Pa`.

Additionally the reconstructed events were analyzed. This was done by computing the difference between the number of reconstructed and generated events. The difference was normalized by division with the parasite tree size. It turns out that each method has its advantages and disadvantages. Using the default cost model Jane 2.0 results in a good estimation for the number of cospeciation events. But it underestimates the number of sortings and duplications and slightly overestimates the number of host switches. Both methods, TreeMap 3b and `CoRe-Pa`, overestimate the number of cospeciations. Whereas TreeMap 3b overestimates the number of sortings and underestimates the number of duplications `CoRe-Pa` is quite exact in predicting the total number of both types of events. On the other hand, `CoRe-Pa` seems to produce too few host switches whereas TreeMap 3b tends to produce slightly too many of them. Figure 6.7 shows boxplots of the deviations of the number of events for each type of event and application gathered from the complete set of cophylogenies of the age model data set. For results using the prunded data set and the ERM model it is referred to

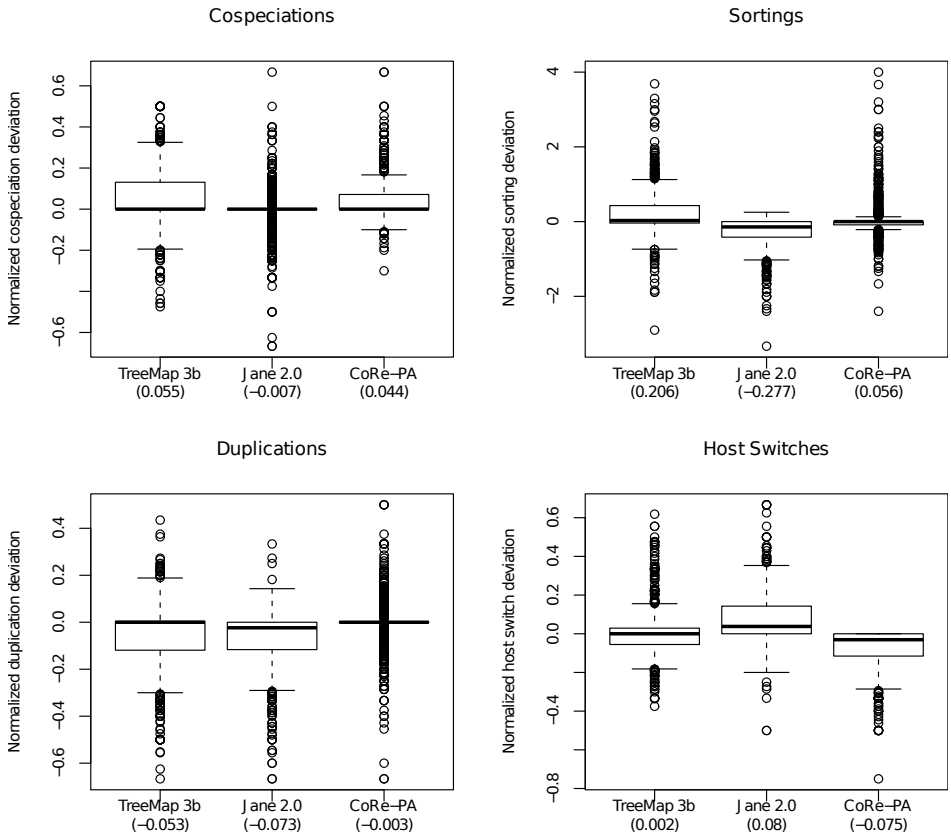


Figure 6.7.: Normalized deviation between number of simulated and reconstructed events for the complete cophylogenies with age model data set. The average deviation per event and tool is depicted in brackets in the x-axis.

Appendix C.6 .

By comparing the runtime of the three tools TreeMap 3b and Jane 2.0 perform quite similar on the test data with the complete phylogenetic trees, but Jane 2.0 is significantly faster on reconstructing the pruned test data set. On average TreeMap 3b needs around 3 to 15 times longer, but this was due to the fact that there were several instances where TreeMap 3b had exceptional long runtimes. `CoRe-Pa` was around 40 to 100 times slower compared to Jane 2.0. But it should be noted that Jane 2.0 considers only a single cost model whereas `CoRe-Pa` analyzes 2,500 different cost models per computation. Figure 6.8 shows boxplots of the runtimes for each application required for the reconstructions of the complete cophylogenies with the age model data set. The results on the runtime for the ERM model and pruned cophylogenies are given in Appendix ap:cophylo:runtime.

It is interesting that the different branching models seem to have only a small impact on the accuracy of the reconstructions. However, when considering pruned cophylogenies the deviation between the reconstructed number and the original number of the host dependent events (cospeciations and sortings) becomes larger. This does not hold for duplications or

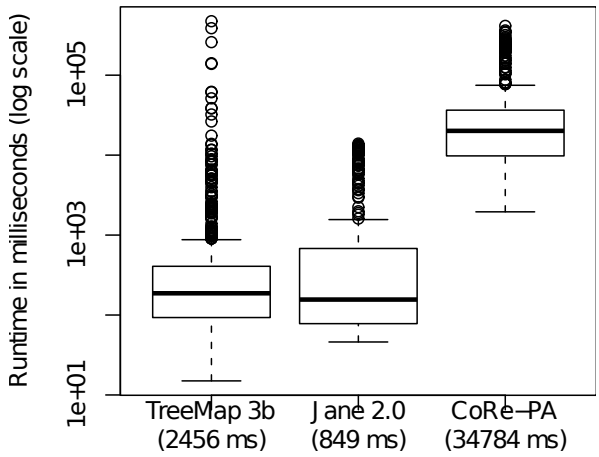


Figure 6.8.: Runtimes of the three tools for the complete cophylogenies with age model data set. The y-axis is in log scale. The average runtime per tool is depicted in brackets in the x-axis.

host switches. Hence, for coevolutionary studies it might be useful to enrich the host parasite systems with data from host species without associated parasites to obtain more precise reconstructions.

6.5. Concluding Comments

In this work, a method for generating cophylogenies that describe the common evolution of two groups of species were presented. In particular, the case of cophylogenies that can describe the coevolution of hosts and their parasites have been considered. Existing branching models for creating phylogenetic trees have been combined with a coevolutionary event model considering cospeciation, duplication, lineage sorting, and host switching events. The influence of different parameters (e.g., the probabilities for different types of coevolutionary events) on the characteristics of the generated cophylogenies have been analyzed. It was shown which parameter values are relevant for generating cophylogenies that have similar properties to cophylogenies found in biological systems. Based on this analysis, different sets of cophylogenies have been generated, that can be used as test data for reconciliation methods. These data sets have been used to make the first systematic study to evaluate the common tools TreeMap 3b, Jane 2.0, and `CoRe-Pa` on test data.

The evaluation has shown that on the generated data sets `CoRe-Pa` is the most precise of the three tools in predicting the correct host parasite associations. But `CoRe-Pa` is not best in estimating the correct number of cospeciation and switching events. Furthermore, `CoRe-Pa` is the computational most intense method. Jane 2.0 is best to estimate the correct number of cospeciations and is the fastest of the three tools. A disadvantage is that it always relies on a single user specified cost model. TreeMap 3b is the only tool which can be configured

such that it always finds the optimal reconstruction for a specified cost model. Using the implemented heuristic it is much faster, but the accuracy of the computed reconstructions is not as good as that of the other tools. Additionally TreeMap 3b sometimes computes several hundred solutions making it hard to decide which is the best without any further evaluation.

CHAPTER 7

Conclusion

Phylogenetic trees capture the evolutionary relationship between species or groups of species in tree-like branching diagrams based on morphological and/or molecular data. In the last decades, a variety of dynamical models have been proposed (Yule, 1925; Aldous, 1996; Nee, 2006; Rosen, 1978; Ford, 2005; Hernández-García *et al.*, 2010) to address the investigation of tree shapes and hence, capture the rules of macroevolutionary forces which gave rise to the diversity of organisms.

In this work, two new models for growing trees in the context of macroevolution have been developed and analyzed. Both models, namely the age model and innovation model, are introduced in Chapter 4. The age model is defined as a stochastic procedure which describes the growth of binary trees by an iterative stochastic attachment of leaves. The branching rate at each clade is no longer constant, as in common models, but decreasing in time i.e. with the age. Thus, species involved in recent speciation events have a tendency to speciate again. The second introduced model describes a branching process which mimics the evolution of species driven by innovations. The process involves a separation of time scales. Rare innovation events trigger rapid cascades of diversification where a feature combines with previously existing features.

Both models were compared intensively in the scope of a tree shape statistics to the most-often used nullmodel, the ERM model, and to AB model. Latter one is known to produced trees with an imbalanced similar to the one from estimated trees of `TreeBASE` but is not intended to model evolutionary processes. These alternative models are simply probabilities distributions on trees while the age model and the innovation model describe a more complex mechanism of generating trees. Three data sets, including of estimated trees were considered

for the tree shape statistic, including `TreeBASE` (phylogenetic trees about the evolution of species and populations), `PANDIT` (phylogenetic trees representing the evolution of protein families) and the small data set `McPeek` (molecular phylogenies on species-level). A tree shape statistic was performed under consideration of a variety of imbalance measurements including the Sackin index and Colles index, which are stated as most powerfull indices. Results show that simulated trees of both growth models fit well to the tree shape observed in estimated trees.

Chapter 5 deals with a further study on the age model based on the likelihood computation in order to rank models with respect to their ability to explain observed tree shapes. The likelihood is calculated by summing up the probabilities of all possible branching sequences. Results show that the likelihoods of the age model and AB model are clearly correlated under the trees in the databases when considering small and medium-sized trees with up to 19 leaves. To summarize, when compared with the AB model, the age model yields larger likelihoods for `PANDIT` data set and slightly less likelihoods for `TreeBASE` on small and medium-sized trees in the databases, where likelihood computation is feasible. In case of the small `McPeek` data set the AB model outperforms the age model but one must take the small amount of trees for each tree size into consideration which may lead to biased results. To support this observation a further analysis using larger trees is necessary. But an exact computation of likelihoods for large trees is computationally too intensive since the number of branching sequences leading to the observed trees grows exponentially. Therefore, an efficient method for likelihood estimation was proposed and compared to the estimation using a naive sampling strategy.

For future work one can consider a comparative analysis of phylogenetic trees not by analyzing the tree structure but using the distribution of branch length data. Altough, branch length data is stated as not reliable as the topological structure of pyhlogenetic trees (Barraclough and Nee, 2001; Pigolotti *et al.*, 2005) at this time, it is assumed that future studies will yield to more approved branch length data (Venditti *et al.*, 2009). In that case, a further comparision of the age model to others regarding branch length may be accomplished.

Both models are sufficiently simple to allow for further enhancement regarding biological concepts. This involves the sequence evolution and genotype-phenotype relations in case of the innovation model. Formulating the age model as a Markov process which allows a speciation at any moment in continuous time would result in a more realistic version.

The coevolution between species, describing the interaction of species across groups such that the evolution of a species from one group can be triggered by a species from another group, is discussed in Chapter 6. Considering systems of host species and their associated parasites a major problem is the reconciliation of the common history of both groups of species. Different heuristic approaches have been proposed recently to solve the problem of predicting

the associations between ancestral hosts and their parasites. But only a few host parasite systems have been analyzed in sufficient detail to serve as benchmark of evaluating reconciliation methods. As far as known there is no approach to generate a reliable test data set in the context of cophylogenies. In this work a method based on the age model is presented to generate such a test data set. The method generates cophylogenies that describe the common evolution of two groups of species. The appliance of the age model exhibits several advantages such as the simultaneously generation of host and the parasite tree. Also the chronological information in terms of age can be used to improve the coevolutionary reconstruction. Three reconciliation methods for cophylogenies, Jane 2.0, Treemap 3b and CoRe-Pa were applied to the test data set generated with the age model. All software tools use an event-based methods to find cost minimal reconstructions. Results have shown that CoRe-PA yields the most precise predictions of the associations between hosts and parasites. However, it does not optimally estimate the number of events and is the computationally most expensive method. Jane 2.0, being the fastest of the three tools, is best at estimating the correct number of cospeciations. TreeMap 3b is the only tool with the option to find the optimal reconstruction for a specified cost model.

To conclude, the presented age model as well as the innovation model produce tree shapes which are similar to obtained tree structures of estimated trees. Both models describe an evolutionary dynamics and might provide a further opportunity to infer macroevolutionary processes which lead to the biodiversity which can be obtained today. Furthermore with the application of the age model in the context of coevolution by generating a useful benchmark set of cophylogenies is a first step towards systematic studies on evaluating reconciliation methods.

APPENDIX A

Program for Likelihood Computation

A.1. General Information

The tool takes as input an oriented (rooted) binary tree T and a stochastic model M of tree growth defined in terms of a Markov chain. The output is the likelihood $L(M, T)$ of the model for the given tree.
When the likelihood cannot be expressed in closed form the calculation is performed by an importance sampling over histories of the Markov chain. The tool is able to deal with a large class of tree growth models including dynamics of hidden variables such as sequence information on the nodes of the tree.

A.2. Availability and Installation

An executable file of the program `LiCoMoPhy` for calculating the likelihood of model generated trees is available at:

- `http://www.stephie-it.de/software/LiCoMoPhy.tar.gz`

It is written in Java 6.0 . Just extract the archive and run `LiCoMoPhy.jar`. An installation is not necessary.

For some examples, change to directory *test* and perform a run by typing:

- `java -jar LiCoMoPhy.jar -m all -e 1 -a 30 -i 5 -c s -p 20 -s y -b T -d ./simpletrees/ > example.out`

After the calculation for all files in `simpletrees` for all available models, the results can be found in the file `example1.out`.

A.3. Input Format

An inputfile or a directory containing files must be committed. Each input files contains the tree data in form of a *node list*. It is essential that the tree data file ends on `.tree` . Example files are given in directory *test*.

A.4. Options

Several options which can be used to manage the estimation of the likelihood are given in Table A.1. Usage: LiCoMoPhy.jar

A.5. Output Format

Output of data stored in a simple file is ordered in columns as followed.

- counter
- filename
- amount of inner nodes
- amount of possible branching sequences
- `llh`, which is the loglikelihood

Additionally for the likelihood estimation for each used model:

- `s` standing for standard deviation
- `lb` gives the lower bound of loglikelihoods
- `lq` gives the lower quartil of loglikelihoods
- `m` stands for the median
- `uq` gives the upper quartil of loglikelihoods
- `ub` gives the upper bound of loglikelihoods

When sampling the likelihood and the output of the likelihood for each sample or is requested, an additional output file is stored. The file is named after the inputfile and used parameters divided by an underscore,
e.g., `M954.wop.treeAGE_cn_e12345_p10000_t1_l0-1.stepwiseout`.

The columns are ordered by:

- index of sample
- logarithm of likelihood for a sample
- average of log likelihoods for recent samples
- seed for random number generator

short	long	type	description	default
`-a`	`--maxleave`	int	Options used for defining maximal number of leaves of a tree, for which a likelihood computation is performed.	`30`
`-ae`	`--ageexp`	int	When using the age model and if one like to change the age paramter, this can be done tuning this parameter.	`1`
`-b`	`--beginstr`	string	String with which treefiles should start (to limit files to process).	
`-c`	`--calctype`	e, s, n	Defines the way of likelihood computation or estimation. Therefore, `e` stands for exact, `s` for sampled and `n` for the naive way of sampling.	`e`
`-e`	`--seed`	int	Defines the initial seed value for random processes.	`12345`
`-h`	`--help`		Calling `help` will give an overview of all possible options.	
`-i`	`--minleave`	int	Options used for defining minimal number of leaves of a tree, for which a likelihood computation is performed.	`5`
`-m`	`--model`	model	Defines the model. It can be choosen from the AB model (`ABM`), ERM model (`ERM`), age model (`AGE`) and innovation model (`INNOV`). If the calculation is requested for all models, just use the option `all` .	`all`
`-o`	`--stepoutdir`	string	Defines the directory for stepwise output of likelihoods. If no output directory is defined, the user's default starting directory is ued.	`user's dir`
`-p`	`--samples`	y, n	Defines the amount of samplings if `--calctype` equals `s` or `n` .	`1`
`-s`	`--sortout`	y, n	Define with this option, if the output files should be sorted by the number of leaves of tree.	`y`
`-t`	`--stepwide`	int	The size of bins for the output of the likelihood computation, when `--stepout` is defined by `y`.	`1`
`-w`	`--stepout`	y, n, l	If the output of the likelihood should take place in bins, use `y` here. For logarithmical binning use `l`. Otherwise use `n`	`n`
Just one of the following must be defined (files must end on .tree!):				
`-f`	`--file`	string	Defines the path of the input file containing tree. Keep in mind to use only `-f` or `-d` .	
`-d`	`--dir`	string	Defines the path of the input directory containing tree files for which likelihood computation should be performed. Again, keep in mind only define `-f` or `-d` .	
If data should be calculated for Innovation model:				
`-l`	`--alpha`	double	The parameter alpha defines the probability to perform an innovation step. If alpha equals zero, the probability distribution equals the one from the ERM model.	`0.1`

Table A.1.: Options supported by the `LiCoMoPhy` algorithm.

APPENDIX B

Tree Statistics Program

B.1. General Information

The tool takes as input an oriented (rooted) binary tree T and calculates different tree shape characteristics including:

- the mean depth, known as Sackin index (Sackin, 1972), describing the compactness of trees by the average distance of a leaf from root in a tree.
- the Colless index (Colless, 1982), for the evaluation of tree balance.
- the cherry distribution (McKenzie and Steel, 2000) as measurement of tree quality by computing the number of cherries of a tree.

The results can be used, e.g. in the scope of a tree shape statistic using the tree imbalance as quantification.

B.2. Availability and Installation

An executable file of the program `TreeStatistics` for calculating some characters of trees generated by different models is available at:

- `http://www.stephie-it.de/software/TreeStatistics.tar.gz`

It is written in Java 6.0 . Simply extract the archive and run the `TreeStatistics.jar`. An installation is not necessary.

For some examples, change to directory *examples* and perform a run by typing:

- `java -jar TreeStatistics.jar -d ./examples/ -m n -o ./ -s y -t example`

After processing all files in `examples` ending on .tree, the results can be found in the cmmitted output directory.

B.3. Input Format

An inputfile or a directory containing files must be committed. Each input files contains the tree data in form of a *node list*.

B.4. Options

Several options which can be used to manage the estimation of the likelihood are given in Table B.1. Usage: LiCoMoPhy.jar

B.5. Output Format

There are two different output files. First the `stat4[dataID]_singlefiles.out` containing the results for each of the processed trees in the following order:

- `n` defining the number of leaves of the processed tree
- `n_left` defining the smallest number of leaves in branches of root
- `d` states the depth of the tree
- `<d>` stands for the mean depth, also Sackin index
- `#cherries` gives the number of cherries
- `I_colless` gives the Colless index
- `node_ID` represents the ID of the node whereas zero refers to the root node

The second generated file `stat4[dataID]_lclsum_l[number of leaves].out` contains for all trees of the same size (number of leaves) how often the amount of leaves in the left branch (refering to the smallest number of leaves of both branches of the root) can be obtained. The data is given in columns in the following order:

- `n_l` represents the smallest number of leaves in branches of root
- `how often appearing` gives the number of observed `n_l` for trees of the same size

short	long	type	description	default
`-a`	`--maxleave`	int	Options used for defining maximal number of leaves of a tree, for which a likelihood computation is performed.	`30`
`-b`	`--beginstr`	string	String with which treefiles should start (to limit files to process).	
`-h`	`--help`		Calling `help` will give an overview of all possible options.	
`-i`	`--minleave`	int	Options used for defining minimal number of leaves of a tree, for which a likelihood computation is performed.	`5`
`-m`	`--subtreeDepthCalc`	y, n	Defines if for each root of tree and each subtree the depth should be calculated or if only the root node of the whole tree should be taken into account for tha calculation.	
`-o`	`--outdir`	string	Defines the output directory for the storage of the results.	
`-s`	`--sortout`	y, n	Define with this option, if the output files should be sorted by the number of leaves of tree.	`y`
`-t`	`--dataID`	string	String for a unique labeling of processed data, e.g. outputfiles for processed trees generated with the age model can be named `stat4age*.out` when using `-t age`	
Just one of the following must be defined (files must end on .tree!):				
`-f`	`--file`	string	Defines the path of the input file containing tree. Keep in mind to use only `-f` or `-d` .	
`-d`	`--dir`	string	Defines the path of the input directory containing tree files for which likelihood computation should be performed. Again, keep in mind only define `-f` or `-d` .	

Table B.1.: Options supported by the tool `TreeStatistics`.

APPENDIX C

Supplement for Results on Cophylogenies

The following Sections contain additional material on the evaluating study. The results are depicted for both models, the ERM model and age model as well as for the complete and pruned cophylogenies. The generated benchmark test data set of all simulation set-ups with respective reconstructions and results can also be downloaded from:
`http://pacosy.informatik.uni-leipzig.de/files/19/suppl11.zip` .

C.1. Results for Variance of the Number of Associated Parasites to a Host

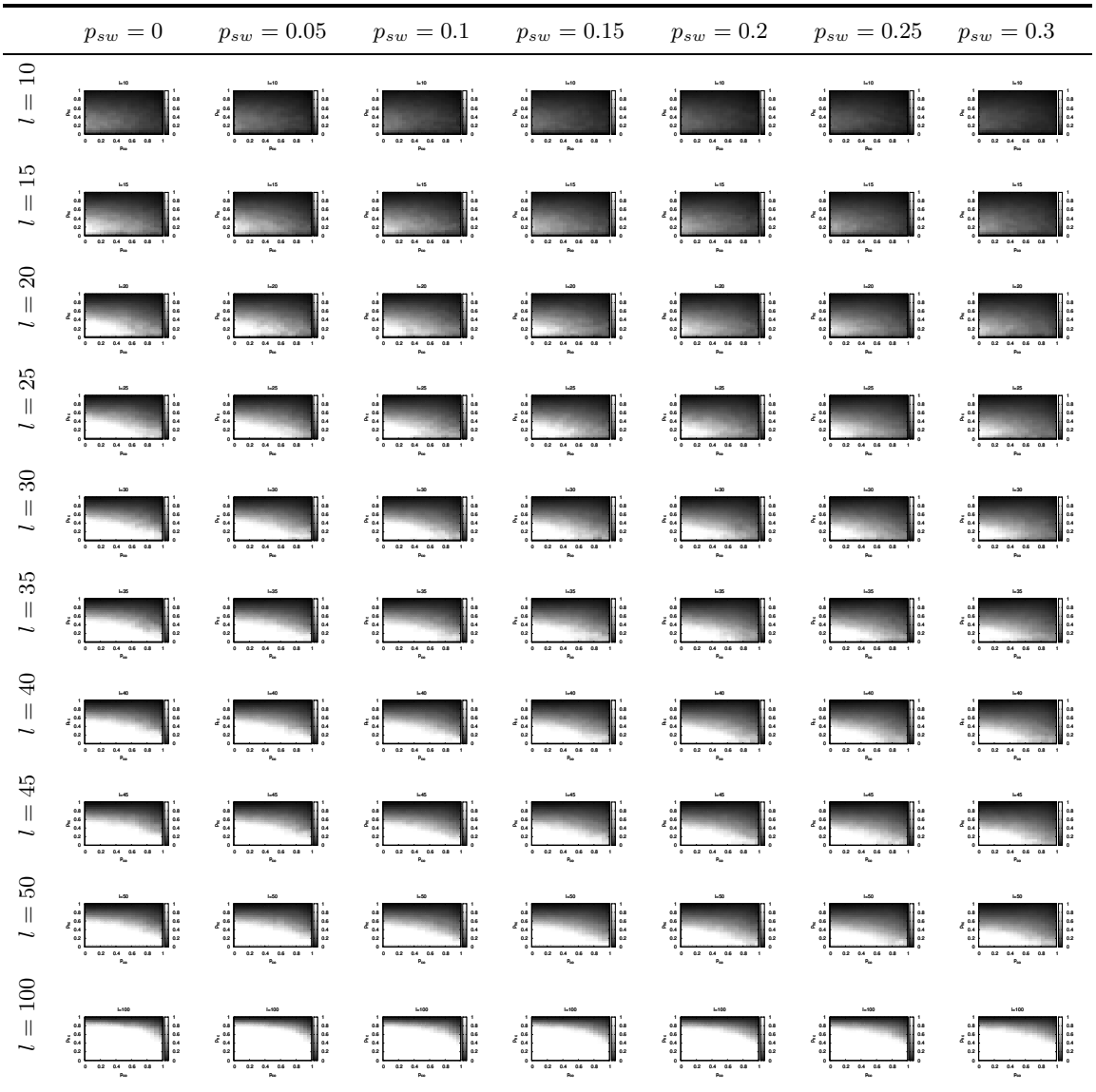


Table C.1.: Results for the normalized variance $var^* \in [0, 1]$ of the number of associated parasites to a host when using the age model for generating cophylogenies. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The variance is depicted as gray scale. The x-axis represents the probability of cospeciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

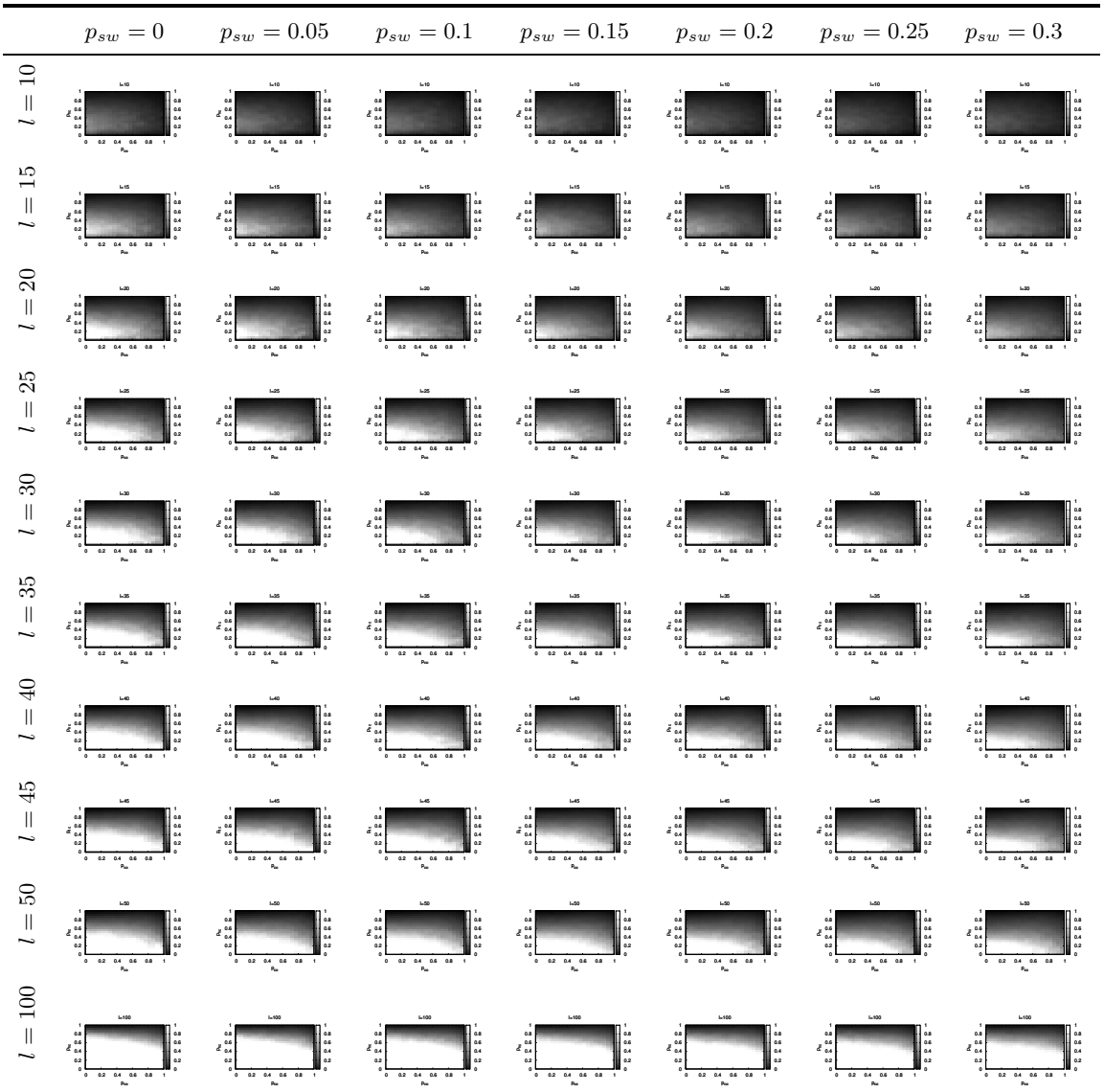


Table C.2.: Results for the normalized variance $var^* \in [0, 1]$ of the number of associated parasites to a host when using the ERM model for generating cophylogenies. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The variance is depicted as gray scale. The x-axis represents the probability of cospeciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

C.2. Results for Ratio Between the Sizes of Parasite and Host Tree

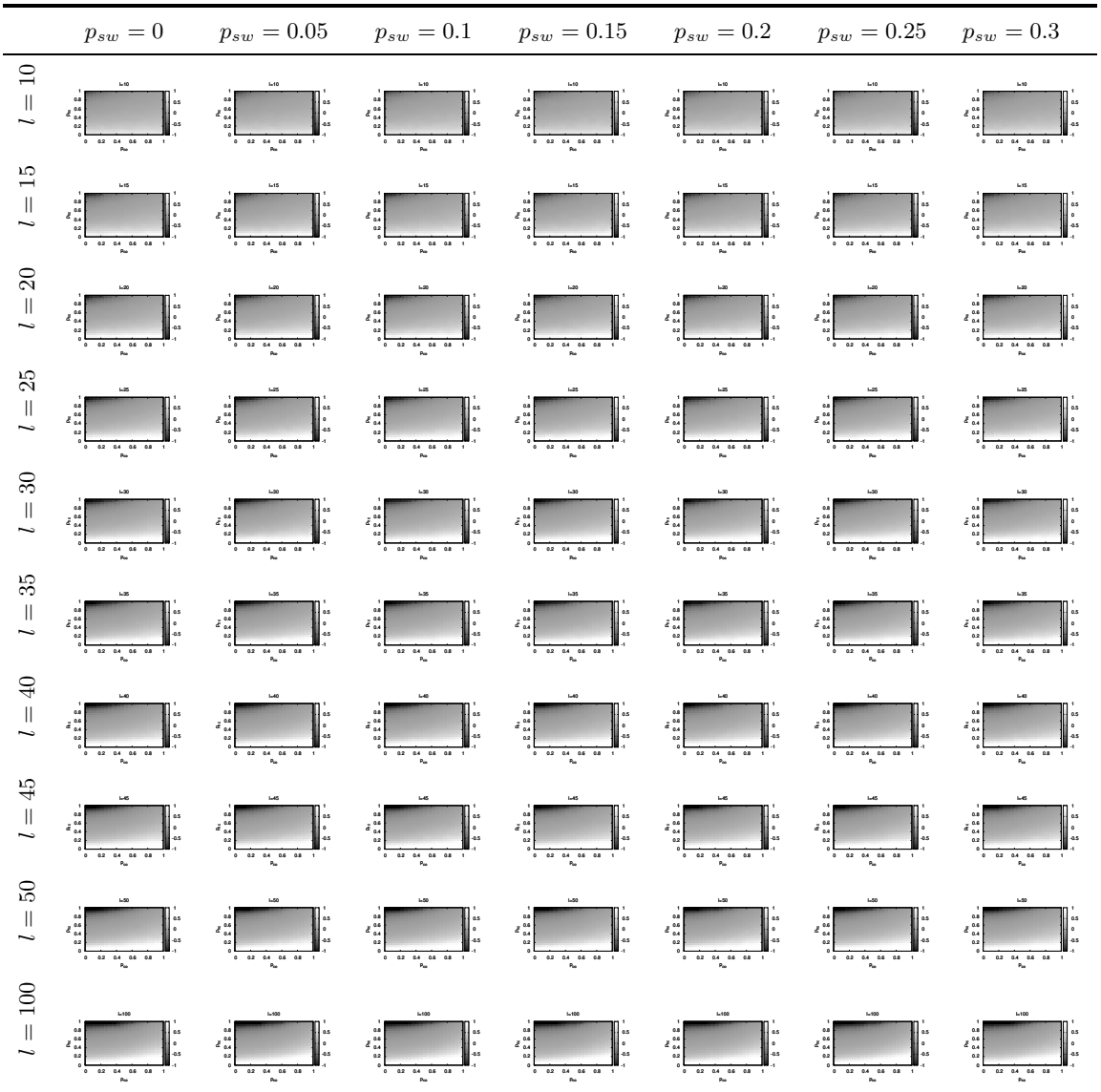


Table C.3.: Results for the normalized ratio $scale^* \in [-1, 1]$ between the sizes of parasite and host tree when using the age model for generating cophylogenies. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The ratio is depicted as gray scale. The x-axis represents the probability of copseciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

Table C.4.: Results for the normalized ratio $scale^* \in [-1, 1]$ between the sizes of parasite and host tree when using the ERM model for generating cophylogenies. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The ratio is depicted as gray scale. The x-axis represents the probability of cospeciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

C.3. Quality Measurement of Cophylogenies

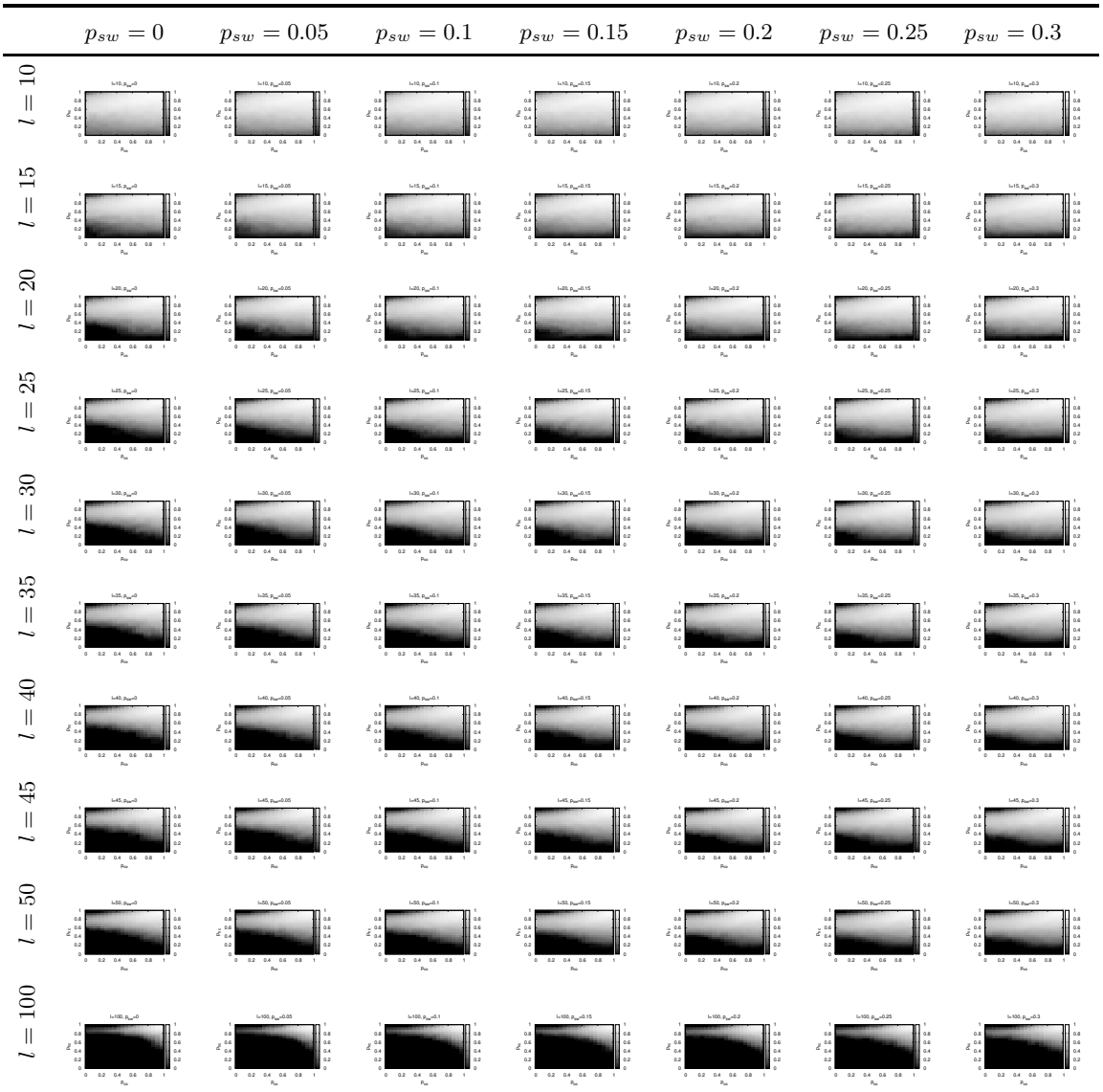


Table C.5.: Result for quality measurement using the measures $scale^*$ and var^* when using the age model for generating cophylogenies. The $quality \in [0, 1]$ measures how likely a cophylogeny can be considered to be realistic. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The quality is depicted as gray scale. The x-axis represents the probability of cospeciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

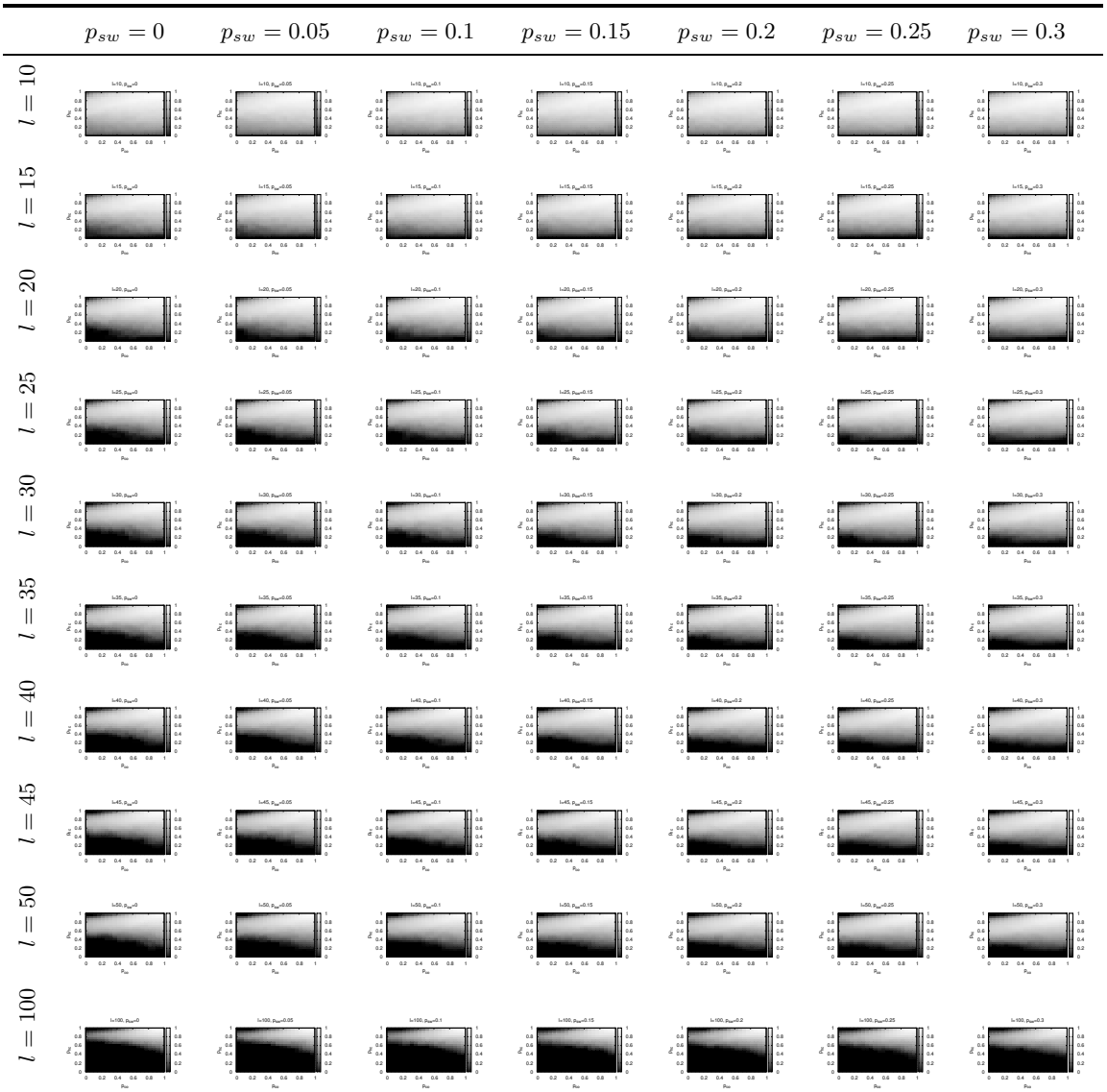


Table C.6.: Result for quality measurement using the measures *scale** and *var** when using the ERM model for generating cophylogenies. The *quality* $\in [0, 1]$ measures how likely a cophylogeny can be considered to be realistic. The use of different parameter combinations is shown for different tree sizes $l \in \{10, 15, 20, 25, 30, 35, 40, 45, 50, 100\}$ and different host switch probabilities $p_{sw} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The quality is depicted as gray scale. The x-axis represents the probability of cospeciation p_{co} while y-axis depicts the probability of performing an event on the host tree.

C.4. Data Sets

	age model		ERM model	
	complete	pruned	complete	pruned
systems with a single node tree	46	112	47	82
systems unfeasible with Jane 2.0	42	117	32	104
systems unfeasible with TreeMap 3b	2	0	1	1
applicable systems	910	771	920	813

Table C.7.: Number of simulated systems which had to be skipped for the analysis, i.e., those containing a single node tree or being unfeasible with one of the reconciliation methods.

C.5. Runtime of Reconciliation Methods

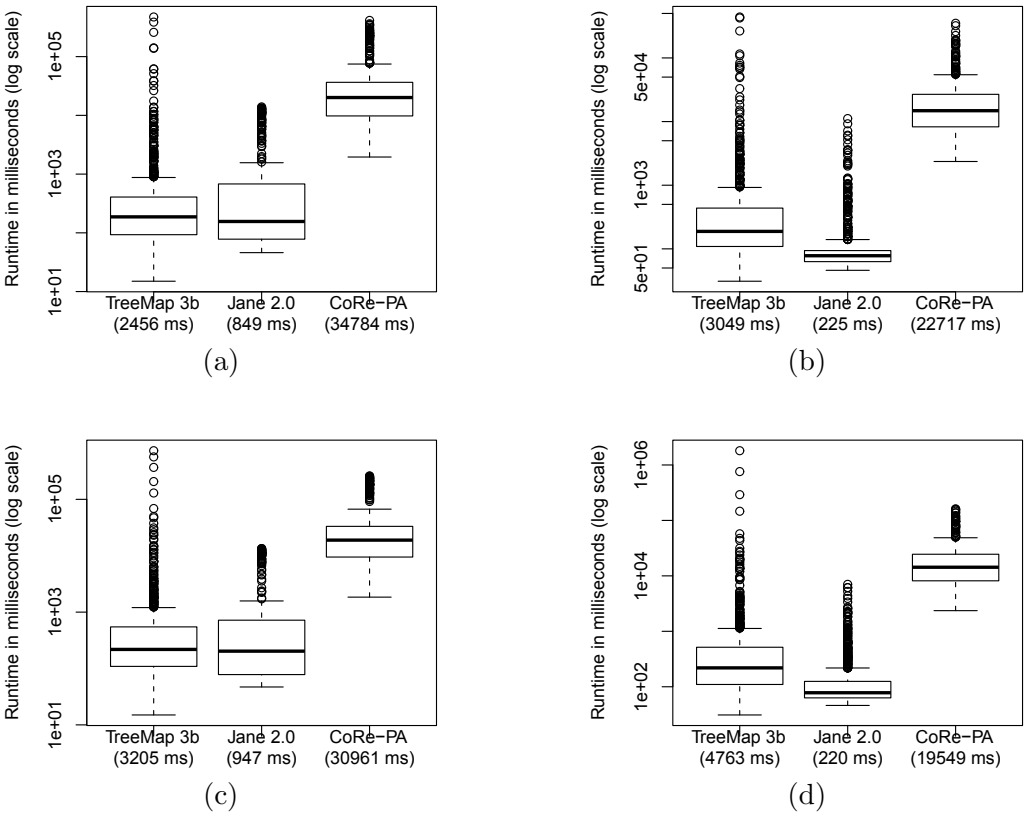
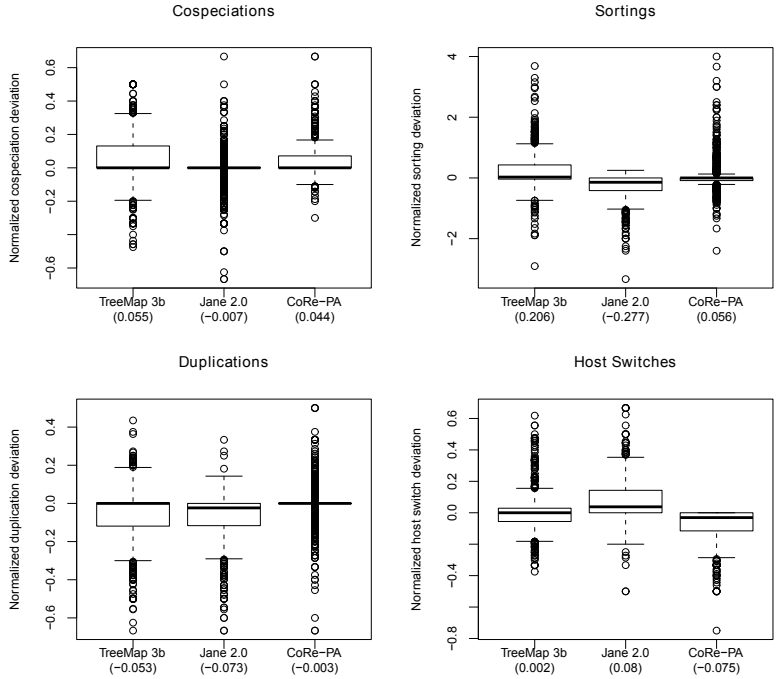
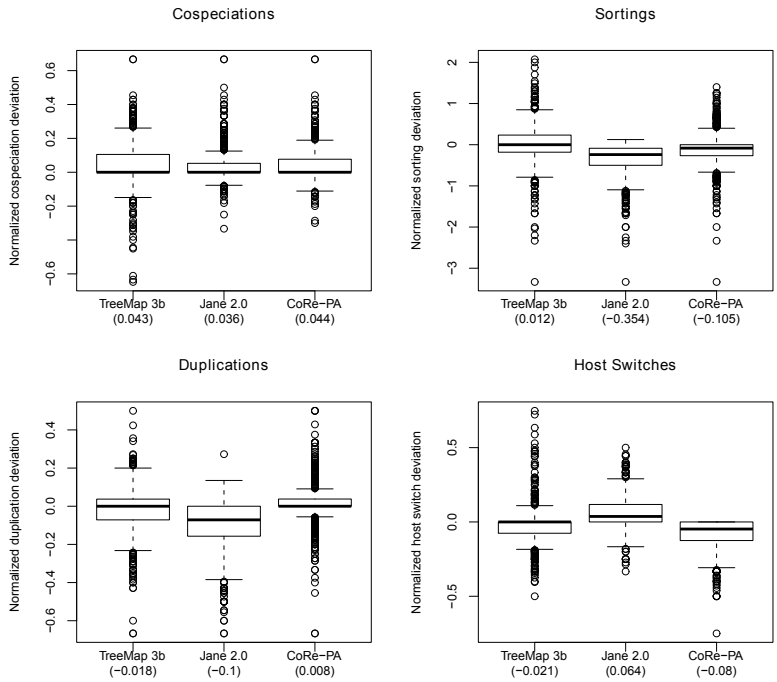


Figure C.1.: Runtimes of the three methods for the complete (a, c) and pruned (b, d) age model (top) respectively ERM model data set (bottom). The y-axis is in log scale. The average runtime per method is depicted in brackets in the x-axis.

C.6. Deviation of Events

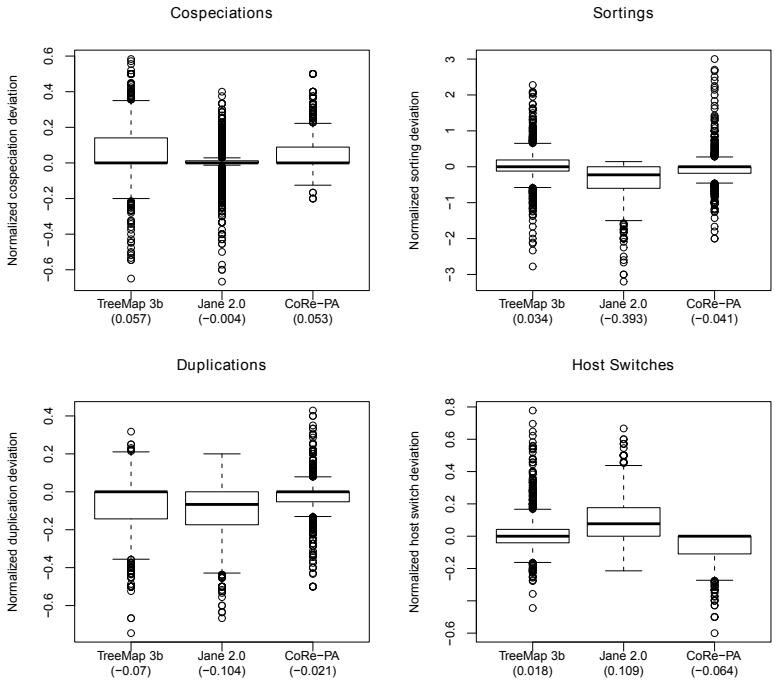


(a)

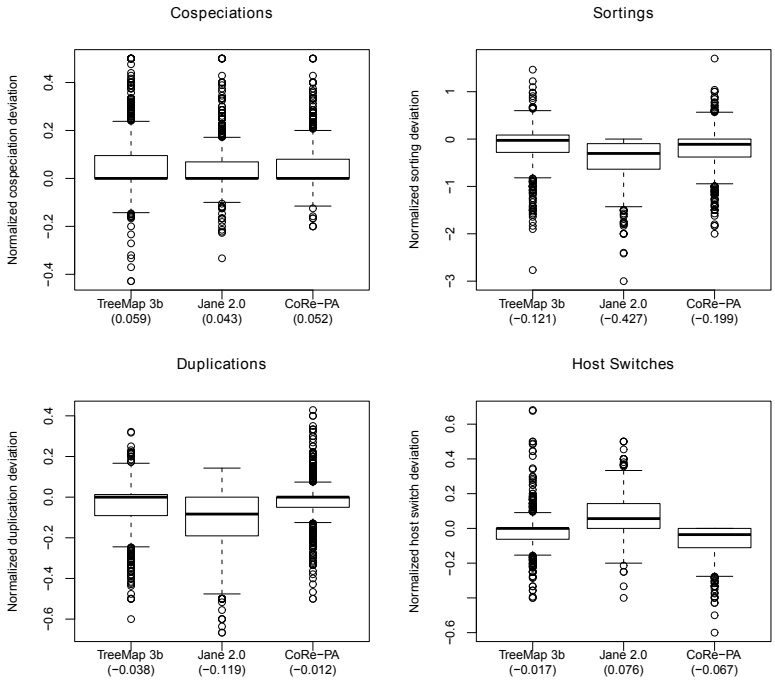


(b)

Figure C.2.: Normalized deviation between simulated and reconstructed events for the complete (a) and pruned (b) age model data set.



(a)



(b)

Figure C.3.: Normalized deviation between simulated and reconstructed events for the complete (a) and pruned (b) ERM model data set.

C.7. Fraction of Exact Predicted Host Parasite Associations

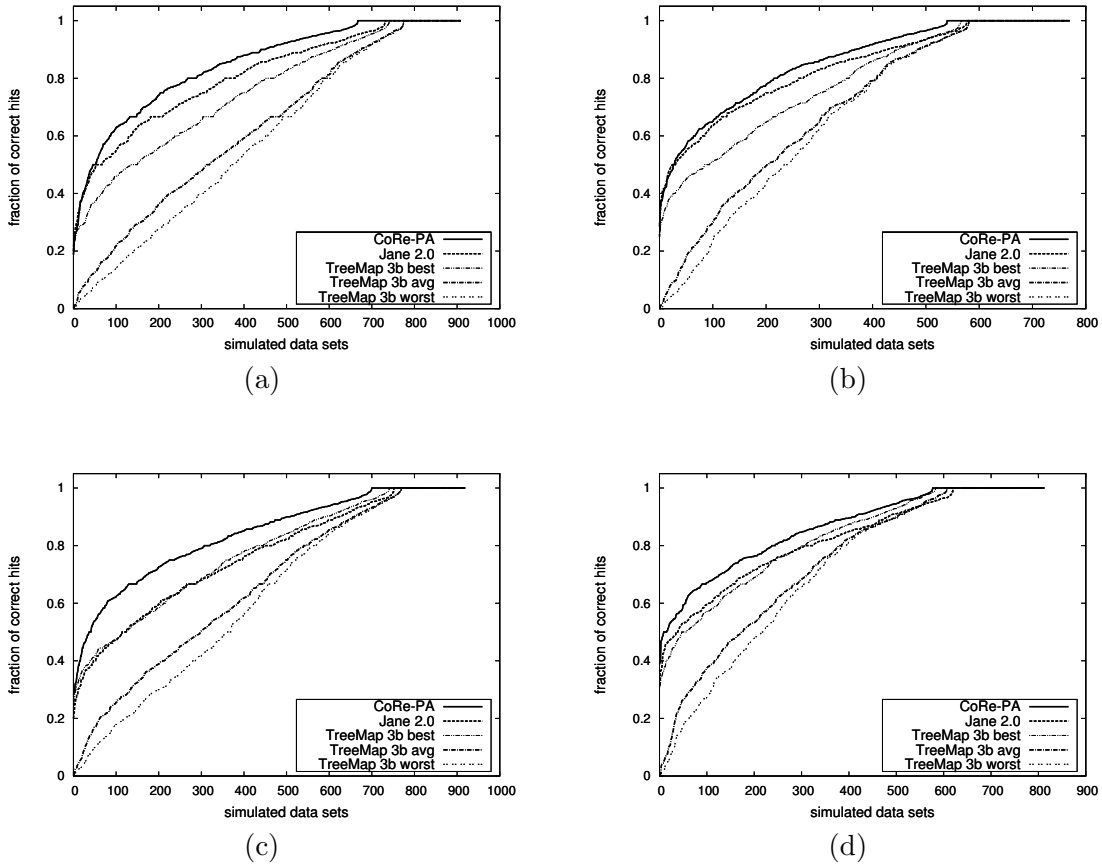


Figure C.4.: Sorted fraction of exact predicted host parasite associations for the complete (a, c) and pruned (b, d) age model (top) respectively ERM model data set (bottom).

List of Algorithms

List of Figures

List of Tables

Bibliography

Abramowitz, M. and Stegun, I. (1972). Handbook of mathematical functions, with formulas, graphs, and mathematical tables. *Dover books on advanced mathematics.*

Agapow, P. and Purvis, A. (2002). Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology*, **51**(6), 866–872.

Aldous, D. (1996). Probability distributions on cladograms. *Random Discrete Structures*, **76**, 1–18.

Aldous, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, **16**, 23–34.

Aldous, D., Krikun, M., and Popovic, L. (2008). Stochastic models for phylogenetic trees on higher-order taxa. *Journal of Mathematical Biology*, **56**(4), 525–557.

Aldous, D., Krikun, M., and Popovic, L. (2011). Five statistical questions about the tree of life. *Systematic Biology*, **60**(3), 318.

Athreya, K. and Ney, P. (1971). Branching process. *American Mathematical Society*, **30**(3).

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Ayala, F. J. and Fitch, W. M. (1997). Genetics and the origin of species: An introduction. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(15), 7691–7697.

Bak, P. and Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, **71**, 4083–4086.

Bang, Y. W. and Kun-Mao, C. (2004). *Spanning Trees and Optimization Problems*. Chapman and Hall/CRC.

Banks, J., Palma, R., and Paterson, A. (2006). Cophylogenetic relationships between penguins and their chewing lice. *Journal of Evolutionary Biology*, **19**(1), 156–166.

Barraclough, T. and Nee, S. (2001). Phylogenetics and speciation. *Trends in Ecology & Evolution*, **16**(7), 391–399.

Beier, R., Röglin, H., and Vöcking, B. (2007). The smoothed number of pareto optimal solutions in bicriteria integer optimization. *Integer Programming and Combinatorial Optimization*, pages 53–67.

Bellman, R. and Harris, T. (1952). On age-dependent binary branching processes. *The Annals of Mathematics*, **55**(2), 280–295.

Blum, M. and François, O. (2005). On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. *Mathematical Biosciences*, **195**(2), 141–153.

Blum, M. and François, O. (2006). Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. *Systematic Biolology*, **55**(4), 685–691.

Blum, M., François, O., and Janson, S. (2007). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, **16**(4), 2195–2214.

Bouge, L., Gabarró, J., and Messeguer, X. (1995). Concurrent avl revisited: self-balancing distributed search trees.

Brusca, R. and Gilligan, M. (1983). Tongue replacement in a marine fish (lutjanus guttatus) by a parasitic isopod (crustacea: Isopoda). *Copeia*, **1983**(3), 813–816.

Campos, P., de Oliveira, V., and Maia, L. (2004). Emergence of allometric scaling in genealogical trees. *ADVANCES IN COMPLEX SYSTEMS*, **7**(1), 39–46.

Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, **19**(3 Pt 1), 233.

Charleston, M. (1998). Jungles: A new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, **149**(2), 191–223.

Charleston, M. (2011). Treemap 3b. http://sites.google.com/site/cophylogeny.

Charleston, M. and Perkins, S. (2006). Traversing the tangle: Algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics*, **39**, 62–71.

Clewley, J. (1998). A user's guide to producing and interpreting tree diagrams in taxonomy and phylogenetics. part i. introduction and naming of parts. *Communicable disease and public health/PHLS*, **1**(1), 64.

Colless, D. (1982). Phylogenetics: The theory and practice of phylogenetic systematics. *Systematic Zoology*, **31**(1), 100–104.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, **5**(1), 16.

Darwin, C. (1859). *On the Origin of Species*. John Murray, 6th edition.

Diestel, R. (2006). *Graphentheorie*. Springer-Verlag Heidelberg, 3rd edition. (elektronische Fassung).

Dobzhansky, T. (1951). *Genetics and the Origin of Species*. Columbia Univ. Press, 3rd edition.

Doyon, J., Scornavacca, C., Gorbunov, K., Szöllösi, G., Ranwez, V., and Berry, V. (2011). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *Comparative Genomics*, pages 93–108.

Erwin, D. (2000). Macroevolution is more than repeated rounds of microevolution. *Evolution & development*, **2**(2), 78–84.

Felsenstein, J. (2004). Inferring phylogenies. *Sunderland, Massachusetts: Sinauer Associates*.

Fiala, K. and Sokal, R. (1985). Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. *Evolution*, pages 609–622.

Filipchenko, I. (1927). *Variabilität und Variation von Jur. Philiptschenko*. Gebrüder Borntraeger, Berlin.

Ford, D. (2005). Probabilities on cladograms: introduction to the alpha model.

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton University Press.

Gell-Mann, M. (1995). Plectics. *The third culture: Beyond the scientific revolution*, pages 316–332.

Gernhard, T., Hartmann, K., and Steel, M. (2008). Stochastic properties of generalised yule models, with biodiversity applications. *Journal of mathematical biology*, **57**(5), 713–735.

Gilinsky, N. and Good, I. (1989). Analysis of clade shape using queueing theory and the fast fourier transform. *Paleobiology*, pages 321–333.

Gould, S., Raup, D., Sepkoski, J., Schopf, T., and Simberloff, D. (1977). The shape of evolution; a comparison of real and random clades. *Paleobiology*, **3**, 23–40.

Gould, S. J. and Eldredge, N. (1993). Punctuated equilibrium comes of age. *Nature*, **366**, 223 – 227.

Gregory, T. (2008). Understanding evolutionary trees. *Evolution: Education and Outreach*, **1**(2), 121–137.

Guetz, A. and Holmes, S. (2010). Adaptive importance sampling for network growth models. *Annals of Operations Research*, pages 1–17.

Guibas, L. J. and Sedgewick, R. (1978). A dichromatic framework for balanced trees. In *SFCS '78: Proceedings of the 19th Annual Symposium on Foundations of Computer Science*, pages 8–21, Washington, DC, USA. IEEE Computer Society.

Guyer, C. and Slowinski, J. (1991). Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, pages 340–350.

Hafner, M. and Nadler, S. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, **332**, 258–259.

Hall, B., Hallgrímsson, B., and Strickberger, M. (2008). *Strickberger's evolution: the integration of genes, organisms and populations*. Jones & Bartlett Learning.

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

Harris, T. (1963). *The theory of branching processes*. Springer-Verlag, Berlin, and Prentice-Hall, Inc., Englewood Cliffs, N.J. Reprinted by Dover, NY, 1989 and 2002.

Heard, S. (1996). Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, pages 2141–2148.

Heard, S. and Hauser, D. (1995). Key evolutionary innovations and their ecological mechanisms. *Historical Biology*, **10**(2), 151–173.

Hennig, W. (1966). *Phylogenetic Systematics*. University of Illinois Press, Urbana-Champaign.

Hernández-García, E., Tugrul, M., Herrada, A., Eguíluz, V., and Klemm, K. (2010). Simple models for scaling in phylogenetic trees. *International Journal of Bifurcation and Chaos*, **20**, 805–811.

Herrada, A., Tessone, C., K., K., Eguíluz, V., Hernández-García, E., and Duarte, C. (2008). Universal scaling in the branching of the tree of life. *PLoS One*, **3**(7).

Herrada, A., Eguíluz, V. M., Hernández-García, E., and Duarte, C. (2011). Scaling properties of protein family phylogenies. *BMC Evolutionary Biology*, **11**(1), 155.

Hey, J. (1992). Using phylogenetic trees to study speciation and extinction. *Evolution*, pages 627–640.

Himmelmann, L. and Metzler, D. (2007). A study on the empirical support for prior distributions on phylogenetic tree topologies. In *Lecture Notes in Informatics-Series of the Gesellschaft für Informatik (GI), Proceedings of the German Conference on Bioinformatics*, volume 115, pages 101–110.

Hughes, J., Kennedy, M., Johnson, K., Palma, R., and Page, R. (2007). Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and pectinopygus lice. *Systematic Biology*, **56**(2), 232–251.

Jackson, A. and Charleston, M. (2004). A cophylogenetic perspective of RNA–Virus evolution. *Molecular Biology and Evolution*, **21**(1), 45–57.

Jones, G. (2011). Tree models for macro-evolution and phylogenetic analysis. *Systematic Biology*.

Keller-Schmidt, S. and Klemm, K. (2011). A model of macro-evolution as a branching process based on innovations. *arXiv:1111.2608, Submitted to Advances in Complex Systems.*

Keller-Schmidt, S., Tuğrul, M., Eguíluz, V., Hernández-García, E., and Klemm, K. (2010). An age dependent branching model for macroevolution. *arXiv:1012.3298v1, Submitted.*, page 10.

Kendall, D. (1948). On the generalized" birth-and-death" process. *The annals of mathematical statistics*, **19**(1), 1–15.

Kikuchi, Y., Hosokawa, T., Nikoh, N., Meng, X., Kamagata, Y., and Fukatsu, T. (2009). Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. *BMC Biology*, **7**(1), 2.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kirkpatrick, M. and Slatkin, M. (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, **47**(4), 1171–1181.

Klemm, K. and Stadler, P. F. (2012). Rugged and elementary landscapes. In Y. Borenstein, editor, *Theory and Principled Methods for Designing Metaheustics*. accepted.

Lemey, P., Salemi, M., and Vandamme, A. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press Cambridge.

Libeskind-Hadas, R. (2010). Jane 2.0. http://www.cs.hmc.edu/ hadas/jane.

Liem, K. and Nitecki, M. (1990). *Key evolutionary innovations, differential diversity, and symecomorphosis*, pages 147–170. University of Chicago Press.

Liu, J. (2008). *Monte Carlo strategies in scientific computing*. Springer Verlag.

Malcolm, S. and Brower, L. (1989). Evolutionary and ecological implications of cardenolide sequestration in the monarch butterfly. *Cellular and Molecular Life Sciences*, **45**(3), 284–295.

Matsen, F. (2006). A geometrical approach to tree shape statistics. *Systematic Biology*, **55**(4), 652–661.

Matsen, F. (2007). Optimization over a class of tree shape statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**(3), 506–512.

McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences*, **164**(1), 81–92.

McPeek, M. (2008). The ecological dynamics of clade diversification and community assembly. *The American Naturalist*, **172**(6), E270–E284.

McPeek, M. and Brown, J. (2007). Clade age and not diversification rate explains species richness among animal taxa. *The American Naturalist*, **169**(4), E97–E106.

Mebs, D. (1994). Anemonefish symbiosis: vulnerability and resistance of fish to the toxin of the sea anemone. *Toxicon*, **32**(9), 1059–1068.

Merkl, R. and Waack, S. (2003). *Bioinformatik Interaktiv - Algorithmen und Praxis*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

Merkle, D. and Middendorf, M. (2005). Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, **123**(4), 277–299.

Merkle, D., Middendorf, M., and Wieseke, N. (2010). A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, **11**(Suppl 1), S60.

Mooers, A. and Heard, S. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, **72**, 31–54.

Moran, P. A. P. (1958). Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge University Press.

Mount, D. W. (2004). *Bioinformatics - Sequence and Genome Analysis*. Cold Spring Harbor Lab Press, 2nd edition.

Nee, S. (2004). Extinct meets extant: simple models in paleontology and molecular phylogenetics. *Paleobiology*, **30**(2), 172.

Nee, S. (2006). Birth-death models in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 1–17.

Nee, S., Mooers, A., and Harvey, P. (1992). Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences*, **89**(17), 8322.

Nee, S., May, R., and Harvey, P. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **344**(1309), 305–311.

Nievergelt, J. (1974). Binary search trees and file organization. *ACM Computing Surveys*, **6**(3), 195–207.

O'Brien, H., Miadlikowska, J., and Lutzoni, F. (2005). Assessing host specialization in symbiotic cyanobacteria associated with four closely related species of the lichen fungus peltigera. *European Journal of Phycology*, **40**(4), 363–378.

Ovadia, Y., Fielder, D., Conow, C., and Libeskind-Hadas, R. (2011). The cophylogeny reconstruction problem is NP-complete. *Journal of Computational Biology*, **18**(1), 59–65.

Page, R. and Holmes, E. (1998). *Molecular evolution: a phylogenetic approach*. Wiley-Blackwell.

Pavlopoulos, G., Soldatos, T., Barbosa-Silva, A., and Schneider, R. (2010). A reference guide for tree analysis and visualization. *BioData mining*, **3**(1), 1.

Per, B. (1996). *How Nature Works: The Science of Self-Organised Criticality*. Copernicus Press, New York, NY.

Pfaff, B. (2004). Performance analysis of bsts in system software. *ACM SIGMETRICS - Performance Evaluation Review*, **32**(1), 410–411.

Pigliucci, M. (2008). What, if anything, is an evolutionary novelty? *Philosophy of Science*, **75**, 887–898.

Pigolotti, S., Flammini, A., Marsili, M., and Maritan, A. (2005). Species lifetime distribution for simple models of ecologies. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(44), 15747–15751.

Pinelis, I. (2003). Evolutionary models of phylogenetic trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**(1522), 1425–1431.

Ramsden, C., Holmes, E., and Charleston, M. (2009). Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence. *Molecular Biology and Evolution*, **26**(1), 143–153.

Raup, D., Gould, S., Schopf, T., and Simberloff, D. (1973). Stochastic models of phylogeny and the evolution of diversity. *The Journal of Geology*, pages 525–542.

Reed, D., Light, J., Allen, J., and Kirchman, J. (2007). Pair of lice lost or parasites regained: The evolutionary history of anthropoid primate lice. *BMC Biology*, **7**(1), 5–7.

Refrégier, G., Le Gac, M., Jabbour, F., Widmer, A., Shykoff, J., Yockteng, R., Hood, M., and Giraud, T. (2008). Cophylogeny of the anther smut fungi and their caryophyllaceous

hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evolutionary Biology*, **8**(1), 100.

Reznick, D. N. and Ricklefs, R. E. (2009). Darwin's bridge between microevolution and macroevolution. *Nature*, **457**(7231), 837–842.

Ricklefs, R. (2007). Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution*, **22**(11), 601–610.

Ripley, B. (1987). *Stochastic simulation*, volume 183. Wiley Online Library.

Ronquist, F. (1998). Three-dimensional cost-matrix optimization and maximum cospeciation. *Cladistics*, **14**(2), 167–172.

Rosen, D. (1978). Vicariant patterns and historical explanation in biogeography. *Systematic Zoology*, **27**(2), 159–188.

Sackin, M. (1972). Good and bad phenograms. *Systematic Biology*, **21**(2), 225–226.

Salisbury, B. (1999). Misinformative characters and phylogeny shape. *Systematic biology*, **48**(1), 153.

Sanderson, M., Donoghue, M., Piel, W., and Eriksson, T. (1994). TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany*, **81**(6), 183.

Savage, H. (1983). The shape of evolution: systematic tree topology. *Biological Journal of the Linnean Society*, **20**(3), 225–244.

Semple, C. and Steel, M. (2003). *Phylogenetics*, volume 24. Oxford University Press, USA.

Sepkoski, J. J. (1993). Ten years in the library; new data confirm paleontological patterns. *Paleobiology*, **19**(1), 43–51.

Shao, K. (1990). Tree balance. *Systematic Biology*, **39**(3), 266.

Simberloff, D. (1987). Calculating probabilities that cladograms match: A method of biogeographical inference. *Systematic Biology*, **36**(2), 175–195.

Simberloff, D., Heck, K. L., McCoy, E. D., Conner, E. F., Nelson, G., and Rosen, D. E. (1981). *There have been no statistical tests of cladistic biogeographic hypotheses*, pages 40–63. Columbia University Press, New York. Book, Section.

Slowinski, J. (1990). Probabilities of n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. *Systematic Biology*, **39**(1), 89.

Sneppen, K., Bak, P., Flyvbjerg, H., and Jensen, M. (1995). Evolution as a self-organized critical phenomenon. *Proceedings of the National Academy of Sciences*, **92**(11), 5209–5213.

Steel, M. and McKenzie, A. (2001). Properties of phylogenetic trees generated by yule-type speciation models. *Mathematical Biosciences*, **170**(1), 91–112.

Steel, M. and McKenzie, A. (2002). The 'shape' of phylogenies under simple random speciation models. In M. Lässig and A. Valleriani, editors, *Biological Evolution and Statistical Physics*, volume 585 of *Lecture Notes in Physics, Berlin Springer Verlag*, pages 162–180. Springer.

Stich, M. and Manrubia, S. C. (2009). Topological properties of phylogenetic trees in evolutionary models. *European Physical Journal B*, **70**(4), 583–592.

Stiles, F. (1981). Geographical aspects of bird-flower coevolution, with particular reference to central america. *Annals of the Missouri Botanical Garden*, pages 323–351.

Thomson, K. (1992). Macroevolution: the morphological problem. *American Zoologist*, **32**(1), 106.

Ungar, P. (2010). *Mammal Teeth: Origin, Evolution, and Diversity*. The Johns Hopkins University Press.

Venditti, C., Meade, A., and Pagel, M. (2009). Phylogenies reveal new interpretation of speciation and the red queen. *Nature*, **463**(7279), 349–352.

West, G., Brown, J., and Enquist, B. (1997). A general model for the origin of allometric scaling laws in biology. *Science*, **276**, 122–126.

Whelan, S., de Bakker, P., Quevillon, E., Rodriguez, N., and Goldman, N. (2006). Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Research*, **34**(suppl 1), D327.

Wieseke, N., Merkle, D., and Middendorf, M. (2010). CoRe-PA. http://pacosy.informatik.uni-leipzig.de/core-pa.

Willis, J. and Yule, G. (1922). Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, **109**(2728), 177–179.

Wiuf, C., Brameier, M., Hagberg, O., and Stumpf, M. (2006). A likelihood approach to analysis of network data. In *Proceedings of the National Academy of Science*, volume 103, pages 7566–7570.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, New York. Brooklyn Botanic Gardens.

Yule, G. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, **213**, 21–87.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Leipzig, den 08. März 2012

(Stephanie Keller-Schmidt)