# Reconstructing Phylogenies in Markov Models of Sequence Evolution

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy
by

Miklós Csűrös

Dissertation Directors: Dana Angluin
Ming-Yang Kao

December 2000

*Édesapámnak, édesanyámnak, és Katherine-nek.*

## Abstract

The path we follow in this dissertation leads from biomolecular sequences and mathematical sequence evolution models to the design of algorithms with superior efficiency for these models. We study the construction of evolutionary trees from sequences in a probabilistic framework. Our focal problem is that of learning evolutionary tree topologies from the sample sequences they generate in Markov models of evolution. We examine several models, most importantly, the i. i. d. Markov model, and some subclassess such as the Jukes-Cantor model and the Hasegawa-Kishino-Yano model.

We discuss the nature of evolutionary distances. We prove a novel result concerning the uniqueness of evolutionary distances as functionals of distributions over mutating sequences, namely, that distance functions differ only by a single factor in time-reversible models with constant substitution rates. We scrutinize methods for estimating evolutionary distances from sample sequences and derive novel upper bounds on the probabilities of large deviations in the cases of the Jukes-Cantor distance, Kimura's distance, the paralinear distance, and the LogDet metric. In each case we show that the probabilities decrease exponentially in sequence length and in the square of similarities between the sequences involved, where distance is the logarithm of similarity.

We offer a comprehensive overview of maximum likelihood, character-based, and distance-based topology reconstruction algorithms. We describe known theoretical guarantees for their success, and sources of computational and statistical difficulties. Building on our convergence analysis of distance functions, we extend existing results for popular distance-based algorithms, such as Neighbor-Joining, on the sample size they require for successful topology recovery.

The final chapter presents a family of novel distance-based algorithms on the principle of "Harmonic Greedy Triplets," originating from our analysis of the distance estimation error. We prove that the algorithms recover the correct topology from sample sequences that are polynomially long in tree size, while running in quadratic time in the number of leaves. Our algorithms are the fastest known to date with provable polynomial sample size bounds. We support our theoretical results with simulation experiments involving large, biologically motivated trees with up to 3135 leaves.

# Contents

# List of Figures

# Acknowledgments

It took me many years to arrive at this point. I have been very fortunate to have met many great pedagogues, from my first-grade teacher Mrs. Rózsika Daróczy to my dissertation advisors. I would like to express my greatest gratitude to all my teachers who taught me the beauty of knowledge and curiosity.

I would like to thank Ming-Yang Kao for suggesting the topic of the thesis and for fruitful discussions. I am greatly indebted to Dana Angluin for her support and guidance throughout my graduate studies. My thesis research has benefited from discussions with Balázs Kégl, James Aspnes, Péter Erdös, Daniel Huson, and Tandy Warnow.

I am very grateful to my parents and two brothers, Zoltán and Péter, for their loving support in all these years. I would like to thank my wife Katherine for the love and happiness she has brought to my life and her help in writing the thesis.

*New Haven, September 29, 2000.*

# Chapter 1

# Preliminaries

## 1.1 Introduction

Key characteristics of the human mind such as language and abstract thinking rely on its ability to explore similarities and differences between objects and events in the surrounding environment. The way we group objects in order to speak and think about them in an organized manner is a pivotal concern of philosophy, and many scientific disciplines, including biology. One of the earliest branches of biology originating from the ancient Greeks is taxonomy, the science of naming and classifying organisms. Biological classification gained a new, causal perspective with the introduction of evolutionary theories. The birth and rapid growth of molecular biology have enriched our understanding of evolution, and have provided a basis to mathematically sound and experimentally testable theories on how biomolecular sequences evolve. In this dissertation we study how evolutionary history can be retraced in various sequence evolution models. We examine the properties of relationships between evolving sequences in order to design efficient algorithms for successful recovery of evolutionary trees. We aim to focus on algorithmic issues; the nature of the problem, however, inevitably puts us at the intersection of statistics, computer science, and — to a lesser extent — molecular biology.

The thesis is structured as follows. The rest of Chapter 1 presents a summary of concepts in molecular biology pertaining to our study, as well as an introduction to evolutionary trees, graph-theoretical concepts, and probabilistic models of evolution. Chapter 2 discusses stochastic models of sequence evolution, concentrating on Markov models. Chapter 3 examines formal definitions of similarity measures between evolving sequences. Chapter 4 reviews existing algorithmic approaches to evolutionary tree reconstruction. Finally, Chapter 5 presents our Harmonic Greedy Triplets algorithms, theoretical results on their efficiency, and experimental results of reconstructing large evolutionary trees.

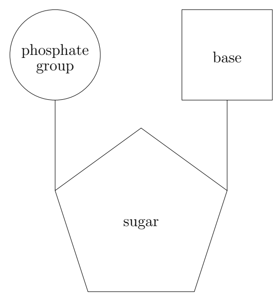
## 1.2 Biomolecular sequences

Before delving into the structure, function, and reconstruction of evolutionary trees, it is necessary to first elucidate the context in which they will be applied, namely, that of biomolecular sequences.

### 1.2.1 Nucleic acids and proteins

The study of molecular interactions within and between cells is ubiquitous in biology. In fact, hardly any part of biology can be effectively studied without it. Here we review some basic concepts of molecular biology. There are many good introductions to the field. In our presentation we rely on the canonic textbook of Lodish *et al.* (1995) and the review of Hunter (1999).

The complex functions of life are primarily the result of coordinated interactions between large molecules. The two groups of large molecules that are most important for our purposes are those of *nucleic acids* and *proteins*. Proteins are responsible for the structure, development, and functioning of the cells. Nucleic acids store and transfer the information necessary to build proteins. With a (not completely precise) analogy from computer science, proteins are dynamic entities corresponding to processes, and nucleic acids are static entities corresponding to stored programs. Nucleic acids and proteins are *linear polymers*. They consist of a sequence of building blocks, called *monomers*. There are only a few possible monomers, but the diversity of proteins and nucleic acids is enormous.




The monomers of nucleic acids are called *nucleotides*. A nucleotide consists of a phosphate group, a small sugar molecule, and an organic base. There are two types of nucleic acids: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). They differ in the types of sugar molecules in the nucleotides comprising them: DNA contains deoxyribose, and RNA contains ribose.

There are only a handful of bases occurring naturally. In DNA, the bases can be adenine (A), guanine (G), cytosine (C), or thymine (T). In RNA molecules, the bases may be A, G, C, or uracil (U) instead of T. Some special RNA molecules may contain also the base inosine (I). During polymerization, the phosphate groups and the sugars of the nucleotides attach to each other to form the DNA or RNA molecule (see Figure 1.1). Since the nucleotides only differ in their bases, a nucleic acid can be described as a sequence. The length of RNA sequences ranges from less than a hundred to many thousands. A DNA sequence may consist of up to a few hundred million nucleotides.

The monomers of proteins are called *amino acids*. There are twenty amino acids found naturally, which only differ in the so-called R-group, or side chain. Amino acids bond to each other in a peptide chain to form a protein (see Figure 1.1). A protein consists of up to a few thousand amino acids. The sequence information about proteins is stored by DNA molecules in the cell. In order to synthesize a protein, first a segment of DNA is transcribed into a specific type of RNA, called messenger RNA (mRNA), which is a polymer of four possible nucleotides: A, G, C, and U. Subsequently, the sequence of mRNA is translated into one or more protein sequences by a *ribosome*, which is a cell organelle responsible for protein synthesis. Ribosomes consist of ribosomal RNA (rRNA) and protein molecules. The mRNA transcribed from DNA may be modified before it reaches the ribosome, and in fact, the same primary mRNA is sometimes edited differently to yield different proteins. During synthesis, the ribosome adds one amino acid at a time proceeding physically along the mRNA in one direction. Thus, the translation from mRNA to proteins is sequential. The translation relies on the fixed encoding of amino acids by nucleotides, known as the *genetic code*. A block of nucleotides encoding an amino acid in the mRNA is called a *codon*.

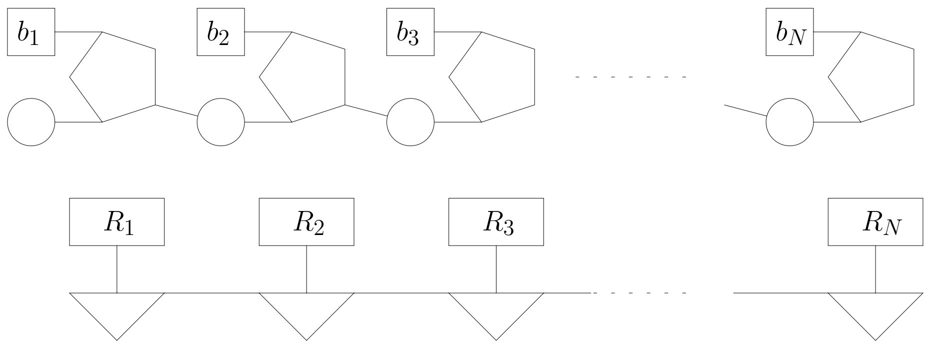



FIGURE 1.1: *Nucleic acids (upper image) and proteins (lower image) are linear polymers, consisting of a sequence of monomers. The nucleotides comprising the nucleic acid form a polynucleotide chain that is defined by the sequence of bases $b_1 b_2 \cdots b_N$. In a conceptually similar manner, the amino acids comprising the protein form a polypeptide chain that is defined by the sequence of R-groups $R_1 R_2 \cdots R_N$.*

A codon consists of three nucleotides, and thus there are 64 codons. With a few exceptions, the genetic code is universal among all organisms. In the universal genetic code there are three codons signaling the end of the protein sequence. The other 61 codons encode the 20 amino acids, and thus multiple codons may encode the same amino acid.

The amino acids are furnished to the ribosomes by transfer RNA (tRNA) molecules. Every tRNA has one specific amino acid that may be connected to it. In addition, every tRNA has a segment, called the *anticodon*, which is a block of three nucleotides. The ribosomes pair up codons with anticodons, separate the protein attached to the tRNA molecule and add it to the growing protein chain. Outside the ribosome, specific enzymes connect the amino acids to the tRNAs. There are 20 such enzymes, one for each amino acid, and 30-50 different tRNAs, depending on the organism. Thus, the success of the translation depends on the precise functioning of the ribosomes and the enzymes attaching amino acids to tRNAs. The codon-anticodon matching, the transcription of DNA to RNA, and even the physical structure of nucleic acids is the result of the chemical affinity of nucleotide bases to pair with each other. This phenomenon, known as *base pairing*, allows the formation of bonds between the following pairs of nucleotides: A with T or U, and G with C. The base inosine is only found in tRNAs, and can pair with C, A, or U. Finally, in case of the codon-anticodon matching, G sometimes pairs with U. Adenine and guanine are *purines*, and thymine, cytosine, and uracil are *pyrimidines*. As a general rule, pyrimidines pair with purines and vice versa.

The transcription of DNA to mRNA is carried out by a group of proteins, which read the DNA sequence in one direction, adding one nucleotide at a time. For each DNA nucleotide, one nucleotide is added to the growing RNA chain, guided by the rules of base pairing. The codon-anticodon matching also relies on the base pairing mechanism, with the "exceptional" events of matching G with U or I with another base occurring in the third position of the codon, removing some redundancy, since the same anticodon may be matched with more than one codon encoding the same amino acid.

The monomers of nucleic acids and proteins interact with each other. In their natural state many chemical bonds are formed between different parts determining the chemical properties of the molecules. In the case of nucleic acids, the physical shape of the molecule is primarily due to the base pairing mechanism. A single strand of nucleic acid, such as an mRNA or a tRNA, forms loops within itself. The structure of DNA is also the result of base pairing. In the native state of DNA, two polynucleotide strands are linked together. The bases in the two strands are complimentary. The two strands entwine to form the double helix structure discovered by Watson and Crick. During DNA replication, for instance, when the cell is replicating itself, the two strands of the DNA separate, and specific enzymes complement the single strands so that two identical DNAs with double strands are constructed.

The starting point for the synthesis of all molecules within the cell is DNA. Nucleic acids, such as tRNA and rRNA are directly transcribed from DNA. Proteins are translated from mRNAs that are transcribed from DNA. Other macromolecules and small molecules are synthesized by proteins. A region of DNA that is involved in the synthesis of a functional protein or RNA molecule is called a *gene*. Genes were actually discovered in the nineteenth century by Gregor Mendel as the discrete units of heredity. Genes are organized into chromosomes in the cell, so that each chromosome contains one double-stranded DNA molecule. In many organisms the chromosomes exist in multiple copies. Most normal human cells contain 22 pairs of so-called autosomes and two sex chromosomes. In eukaryotic organisms the chromosomes are found in the cell nucleus. Some other organelles, namely the mitochondrion and the chloroplast, contain their own DNA.

### 1.2.2 Molecular evolution

The evolution of organisms is the result of the variation in genes as they are passed to the descendants. The sources of the variations may be localized errors during DNA replication, or global rearrangements of genes within and between chromosomes. On a chain of descendants, the variations grow with the generations. Consequently, the difference between organisms at the genetic level can be used to infer their evolutionary relationships. Resolving the sequence of a particular gene, either as the genomic DNA, the transcribed RNA, or the translated protein is a fairly routine procedure by now (Smith *et al.* 1986; Bonfield and Staden 1995; Ewing *et al.* 1998). The challenge lies rather in the manner of comparing sequences between different organisms, which is the main theme of our dissertation.

Evolutionary ancestor-descendant relationships can be depicted by an evolutionary tree, or phylogeny. Homologous bimolecular sequences taken from different species can be used to reconstruct their evolutionary history. For example, Penny and Hasegawa (1997) used complete mitochondrial DNA sequences from a number of mammals to support the theory that the evolution of the platypus branched off from that of the marsupials after the evolution of marsupials and mammals took a different course. Sometimes there are DNA sequences available even from extinct species (Wayne *et al.* 1999), in which case paleontological theories can be evaluated through molecular evolutionary studies. An interesting example of one such study is that of Noro *et al.* (1998), who used mitochondrial gene sequences from a mammoth recovered from Siberian permafrost to establish the evolutionary relationships between the woolly mammoth and its extant relatives. Mitochondrial DNA is often used in molecular evolutionary studies (Brown 1985) since it is almost exclusively inherited in the maternal lineage. In contrast, nuclear DNA in diploid cells is inherited from both parents, and thus different trees may depict the evolution of a gene sequence through generations, depending on whether the maternal or paternal lineage is considered (Avise and Wollenberg 1997). Such problems can be resolved through a mathematically exact definition of the tree structure when finite populations are compared (e.g., Tavaré 1995), but it is beyond the scope of this dissertation. Mitochondrial sequences are favored because they mutate rapidly, and thus recent evolutionary events can be identified using them. For instance, evolutionary studies on the origin of different human populations (e.g., Maddison *et al.* 1992) often use mitochondrial DNA.

Phylogenetic analysis has many other applications besides retracing evolutionary history of species (Hillis 1997). Galtier *et al.* (1999) used rRNA sequences to hypothesize about the very origins of life, finding that the common ancestor to all extant life forms probably lived at a moderate temperature, refuting the "hot origin of life" conjecture. Gao *et al.* (1999) used HIV and SIV sequences to support the theory that HIV-1 originates from a subspecies of chimpanzees found in Gabon. A study in a similar vein by Bollyky and Holmes (1999) investigates several hypotheses about the origin of the hepatitis B virus. Comparison of viral sequences can be useful also for deducing the transmission chain of a particular disease, since the evolution of the viral sequences branches off at infection events. Ou *et al.* (1992) built a phylogeny based on HIV sequences to support the evidence that a Florida dentist infected several patients. Evolutionary trees can also be derived from the comparison of different genes within and between species, in which case the origin of the corresponding proteins can be studied. For example, phylogenies of globin and globin-like proteins (Suzuki and Imai 1998) furthered our understanding of how new genes emerge. Similar studies are useful also in predicting the function of novel genes (Eisen 1998) Finally, as a somewhat unexpected application, Matisoo-Smith *et al.* (1998) compared mitochondrial DNA sequences from Pacific rats on different islands in order to test hypotheses on how humans populated Polynesia. Since the rats accompanied the ancestral Polynesians on their voyages, and the evolution of the rats took different courses on different islands, the phylogeny provides information about prehistoric settlement events.

## 1.3 Graphs and trees

We start with an obligatory list of graph-theoretic terms used throughout this work, recalling some basic terminology from graph theory (see, for example, Bollobás 1979, Bondy and Murty 1976) and data structures (see for example, Cormen, Leiserson, and Rivest 1990). A *graph* $\mathcal{G} = (V, E)$ is defined by the set of *vertices* $V$ and the set of *edges* $E$. In an *undirected graph*, each element of $E$ is an unordered pair of vertices, whereas in a *directed graph*, each element is an ordered pair of vertices. If edge $e$ is defined by the pair $(u, v)$, then $u$ and $v$ are called the *endpoints* of $e$. Using a shorthand notation, we write $e = uv$ (which is equivalent to $e = vu$ in an undirected graph). The edge $uv$ is also said to *connect* $u$ to $v$.

The *degree* of a vertex $u$ in an undirected graph $(V, E)$ is the number of edges of the form $uv$. The *in-degree* of a vertex $u$ in a directed graph $(V, E)$ is the number of edges of the form $vu \in E$. The *out-degree* of $u$ is the number of edges of the form $uv \in E$.

A *path* is an alternating sequence of vertices and edges

$$v_1, e_1, v_2, e_2, \ldots, e_{k-1}, v_k$$

such that for every $i$, $e_i$ connects the vertex $v_{i-1}$ to $v_i$, and the vertices $v_1, \ldots, v_{k-1}$ are all different. If $v_1 = v_k$ and $k > 1$, then the path forms a *cycle*. The *length* of the path is the number of edges in it ($k - 1$ in this example). Notice that zero-length paths consisting of a single vertex are permitted.

A graph $\mathcal{G}$ is connected if there is a path between any two of its vertices. An undirected connected graph with no cycles is called an *unrooted tree*. A *rooted tree* $\mathcal{T} = (V, E)$ is a directed graph with vertex set $V$ and edge set $E$, in which there exists exactly one vertex $u \in V$, such that there is exactly one path from $u$ to any vertex in $\mathcal{T}$. The vertex $u$ is called the *root* of $\mathcal{T}$. By "tree" we will always mean "rooted tree". The vertices of a tree are also referred to as *nodes*. A rooted tree is obtained from an unrooted tree by directing its edges outwards from the root.

**Lemma 1.1.** *If the graph $\mathcal{T} = (V, E)$ is a tree, then every non-root node has in-degree one.*

Proof. Let $w \in V$ be an arbitrary non-root node. Since there is a path from the root to every node $v$ with $vw \in E$, which can be extended to reach $w$ through the edge $vw$, there can only be one edge $vw \in E$. ■

Let $\mathcal{T} = (V, E)$ be a tree. If $uv \in E$, then $u$ is the *parent* of $v$ and $v$ is a *child* of $u$. Lemma 1.1 states that every non-root node has exactly one parent. A node that has no children is called a *leaf*. A node that has a parent and at least one child also is called an *inner node*. Consequently, a node is either the root, an inner node, or a leaf. A tree in which every non-leaf node has two children is a *binary tree*. The *ancestors* of a node $v$ are the nodes on the path from the root to $v$. If a node $u$ is an ancestor of another node $v$, denoted by $u \prec v$, then $v$ is called a *descendant* of $u$. By definition, the parent of an ancestor of a node $v$ is also an ancestor of $v$. Similarly, the child of a descendant of a node $u$ is also a descendant of $u$. Every two nodes in the tree have at least one common ancestor, the root. The *lowest common ancestor*

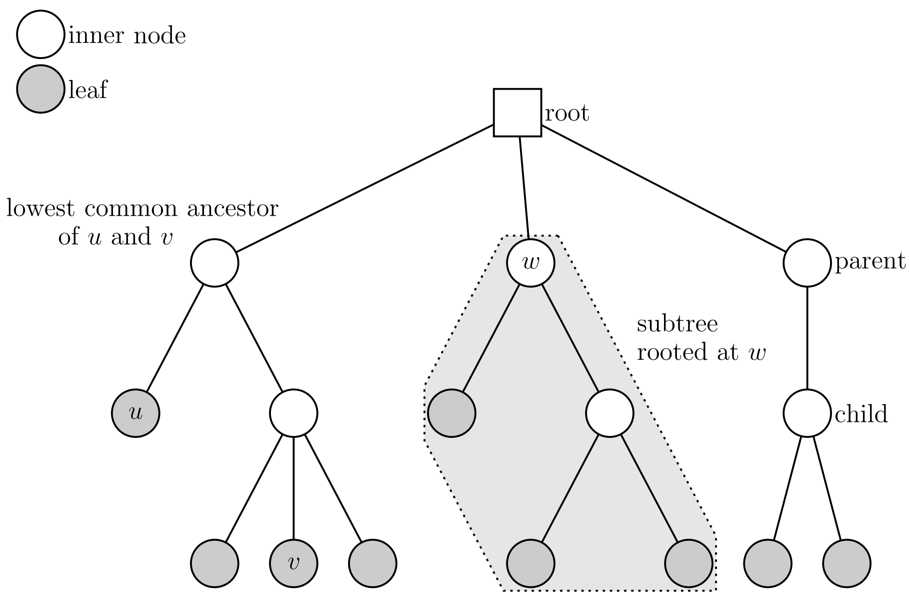



FIGURE 1.2: *Example for a rooted tree. Edges point downwards.*

of two nodes is their common ancestor that is reached by the longest path from the root. Every pair of nodes has only one lowest common ancestor; were this not the case, there would be more than one path leading from the root to either of the two nodes. Figure 1.2 illustrates some of the defined notions for trees.

Let $u$ be a node of tree $\mathcal{T} = (V, E)$. The graph defined by the descendants of $u$ and the edges between them is called the *subtree* of $\mathcal{T}$ rooted at $u$. In other words, the subtree of $\mathcal{T}$ rooted at $u$ is the graph $\mathcal{T}_u = (V_u, E_u)$ with

$$V_u = \Big\{v \in V \colon u \prec v\Big\};$$
$$E_u = \Big\{vw \in E \colon v, w \in V_u\Big\}.$$

Figure 1.2 shows an example for a subtree.

Two graphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_2, E_2)$ are *isomorphic* if there exist bijections $f \colon V_1 \mapsto V_2$ and $g \colon E_1 \mapsto E_2$ such that for each $uv \in E_1$,

$g(uv) = f(u)f(v)$. Isomorphism is denoted by

$$\mathcal{G}_1 \simeq \mathcal{G}_2.$$

If, in addition, there exists a node set $L \subseteq V_1$ such that for every $u \in L$, $f(u) = u$, then we say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are $L$-isomorphic, denoted by

$$\mathcal{G}_1 \underset{L}{\simeq} \mathcal{G}_2.$$

In particular, if $\mathcal{G}_1 \underset{V_1}{\simeq} \mathcal{G}_2$, then $\mathcal{G}_1 = \mathcal{G}_2$.

## 1.4 Stochastic models

### 1.4.1 Sequence evolution

Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be an alphabet of size $m > 1$. Let $\mathcal{A}^+$ denote the set of all sequences over $\mathcal{A}$ with positive length. Denote the set of sequences that may arise in the studied evolutionary process by $\mathcal{S} \subseteq \mathcal{A}^+$. From a theoretical viewpoint, $\mathcal{S} = \mathcal{A}^+$ is an obvious choice. However, "biological languages" may impose constraints, as they do for example in the study of a protein family characterized by some common features. We exclude the empty sequence from $\mathcal{S}$ because it would only complicate the discussion without any theoretical or practical advantage.

An *evolutionary tree*, or *phylogeny*, is defined by two components, a tree and a probability distribution over sequences associated with the tree nodes. Formally, a phylogeny is defined as a triple $\mathcal{P} = (V, E, \mathbb{P})$ with the following properties.

- The graph $\mathcal{T} = (V, E)$ is a rooted tree. The nodes of $\mathcal{T}$ are called *taxonomic units*, or shortly *taxa*.

- Every taxon $u \in V$ is associated with a random *taxon sequence* $X^{(u)}$, which is a random variable taking values on $\mathcal{S}$. The joint probability distribution for the vector of random sequences[1] $\langle X^{(u)} \colon u \in V \rangle$ is defined by $\mathbb{P}$.

[1]It is assumed without loss of generality that an ordering is fixed on $V$, say, by depth-first traversal of the nodes (see Cormen *et al.* 1990).

Since every taxon sequence can take at most countably infinite values, there are no measurability concerns. Thus, for any collection of sequence sets $\langle \mathcal{B}^{(u)} \subseteq \mathcal{S} \colon u \in V \rangle$, the probability

$$\mathbb{P} \bigcap_{u \in V} \left\{ X^{(u)} \in \mathcal{B}^{(u)} \right\}$$

is well-defined.

Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny. The underlying rooted tree $(V, E)$ describes the evolutionary ancestor-descendant relationships between the taxa. By removing the direction of the edges we get an unrooted tree that summarizes the "relatedness" of the taxa by taking away the aspect of time. Let $\Psi(\mathcal{P})$ denote the undirected graph obtained from the tree $(V, E)$ by removing the direction of the edges. The graph $\Psi(\mathcal{P})$ is called the *topology* of $\mathcal{P}$.

### 1.4.2 Evolutionary tree reconstruction

Let $\mathcal{P} = (V, E, \mathbb{P})$ be an evolutionary tree. The problem of *evolutionary tree reconstruction* is the one of deriving, or at least estimating, $\mathcal{P}$ from a set of observed sequences drawn according to $\mathbb{P}$. The problem is typically aggravated by the fact that one can observe only a subset of all the generated sequences. For example, it is rarely the case that one has access to gene sequences of extinct species in molecular evolution studies (with notable exceptions such as that of Noro *et al.* (1998) based on mammoth genes). Let $L \subseteq V$ denote the set of observable nodes. Generally, $L$ should include all the leaves lest we fail to have any information on some subtrees. In an epidemiological study where the nodes represent virus sequences in individuals at different times (as in the study of Leitner *et al.* 1996 on HIV), $L$ may include some inner nodes corresponding to sequences in samples taken from the same individual at earlier times. An *evolutionary tree reconstruction algorithm* is an algorithm that outputs a hypothetical evolutionary tree $\mathcal{P}^*$ based on a set of observed sequences associated with nodes in $L \subseteq V$.

The success of the algorithm is judged by how well it reconstructs the evolutionary relationships between nodes in $L$. There are many possible rooted trees in which those relationships are the same. In order to illustrate and clarify this statement we describe two relationship-preserving operations on rooted trees. Let $\mathcal{T} = (V, E)$ be a rooted tree. The first operation,

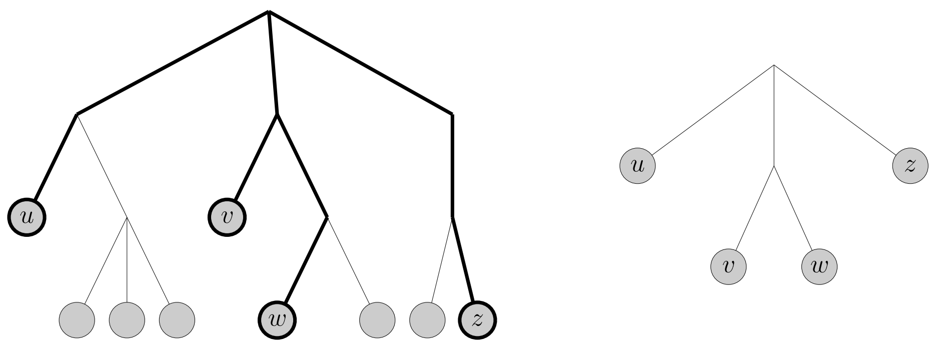



FIGURE 1.3: *The tree on the right-hand side is a topological minor of the tree on the left-hand side over the set $L = \{u, v, w, z\}$.*

$\texttt{INNER}(uv, w)$, is that of adding a new inner node $w \notin V$ on an arbitrary edge $uv \in E$, which results in the tree

$$\mathcal{T}' = \Big(V \cup \{w\}, \Big(E - \{uv\}\Big) \cup \{uw, wv\}\Big).$$

The second operation, $\texttt{LEAF}(u', w')$, is that of adding a new leaf $w' \notin V$ by connecting it to an arbitrary node $u' \in V$, which leads to the tree

$$\mathcal{T}'' = \Big(V \cup \{w'\}, E \cup \{u'w'\}\Big).$$

Clearly, no series of these two operations introduces new relationships between nodes of $V$. In other words, the ancestor-descendant relationships between the original nodes remain the same. For a formal treatment, we introduce the notion of topological minors. Let $\mathcal{T}_1 = (V_1, E_1)$ and $\mathcal{T}_2 = (V_2, E_2)$ be two rooted or unrooted trees. Let $f : V_2 \mapsto V_1$ be a mapping from nodes of $\mathcal{T}_2$ onto those of $\mathcal{T}_1$ such that if $u \neq u'$ then $f(u) \neq f(u')$ and if $uu' \in E_2$

then there is a path from $f(u)$ to $f(u')$ in $\mathcal{T}_1$. For every edge $uu' \in E_2$ define $f^*(uu')$ as the path from $f(u)$ to $f(u')$ in $\mathcal{T}_1$. The mapping $f$ *preserves the topology* of $\mathcal{T}_2$ if for all edges $e, e' \in E_2$, $f^*(e)$ and $f^*(e')$ share no edges. Now let $\mathcal{T}_1 = (V_1, E_1)$ be an arbitrary rooted (or unrooted) tree, and let $L \subseteq V_1$ be an arbitrary node set. A rooted (respectively, unrooted) tree $\mathcal{T}_2 = (V_2, E_2)$ is a *topological minor* of $\mathcal{T}_1$ over $L$ if $L \subseteq V_1 \cap V_2$ and there exists a topology preserving mapping $f$ from $V_2$ to $V_1$ such that for every $u \in L$, $f(u) = u$. We denote this fact by

$$\mathcal{T}_2 \underset{L}{\pitchfork} \mathcal{T}_1.$$

Figure 1.3 shows an example illustrating the concept of a topological minor. Returning to the operations `INNER` and `LEAF`, any series of their application to a tree $\mathcal{T} = (V, E)$ leads to a tree $\mathcal{T}'$ such that $\mathcal{T} \underset{V}{\pitchfork} \mathcal{T}'$. Using the minimal topological minors, we introduce an equivalency relation over trees. Let $\mathcal{T}_1 = (V_1, E_1)$ and $\mathcal{T}_2 = (V_2, E_2)$ be two rooted or unrooted trees and let $L \subseteq V_1 \cap V_2$ be an arbitrary nonempty node set. The trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are *topologically equivalent* over $L$ if there exists a tree $\mathcal{T}'$ with $\mathcal{T}' \underset{L}{\pitchfork} \mathcal{T}_1$ and $\mathcal{T}' \underset{L}{\pitchfork} \mathcal{T}_2$. The fact that $\mathcal{T}_1$ and $\mathcal{T}_2$ are topologically equivalent is denoted by

$$\mathcal{T}_1 \underset{L}{\sim} \mathcal{T}_2.$$

Now we are ready to discuss the success criterion for an evolutionary tree reconstruction algorithm. The algorithm *recovers the topology* if $\Psi(\mathcal{P}^*) \underset{L}{\sim} \Psi(\mathcal{P})$, i.e., if the evolutionary relationships between taxa of $L$ are derived correctly. The main focus of our study is topology recovery. A *topology reconstruction algorithm* outputs an unrooted tree $\Psi^*$, based on a set of observed sequences associated with nodes in $L \subseteq V$. The output $\Psi^*$ is the algorithm's prediction of $\Psi(\mathcal{P})$, so the algorithm is successful if $\Psi^* \underset{L}{\sim} \Psi(\mathcal{P})$. Obviously, every evolutionary tree reconstruction algorithm can be considered as a topology reconstruction algorithm.

In view of the general definition of phylogeny in §1.4.1, topology recovery seems hopeless, since there is no dependence imposed between the underlying tree and the taxon sequence distribution. Chapter 2 introduces several models in which there is such a dependence, making topology recovery possible. Such models imply that the reconstruction is restricted to a *hypothesis class* $\mathcal{C}$, i.e., that $\mathcal{P}$ belongs to $\mathcal{C}$, and that the reconstruction algorithm selects a topology based on that condition. Speaking generally, we hope that $\mathcal{C}$ is large enough to include a tree that models evolution well in reality, yet small enough to enable the use of an efficient algorithm. This philosophy is borrowed from computational learning theory (Kearns and Vazirani 1994) and statistical pattern recognition (Devroye, Györfi, and Lugosi 1996). This dissertation presents a family of novel efficient algorithms that work on large hypothesis classes used in molecular evolutionary studies. Leaving formal definitions of efficiency for later, we state in advance that the algorithms are efficient from both computational and statistical viewpoints — they run in polynomial time in the size of the input, and require small amounts of data for highly accurate topology recovery.

# Chapter 2

# Stochastic models of sequence evolution

## 2.1 Introduction

In §1.4 we outlined a general framework of stochastic sequence evolution. We noted that in order to achieve success in evolutionary tree reconstruction, there must be a known relationship between topology and sequence probability space. This chapter discusses some hypothesis classes that are widely used in molecular evolution studies, which incidentally do impose a relationship between topology and sequence probabilities. In fact, in every class studied, topology is a function of the random taxon sequence distribution.

## 2.2 Memoryless evolution

One particular feature of evolution is that inheritance depends solely on the parents and not on the entire series of ancestors. The evolutionary changes leading from the first mammals to *Homo sapiens* were not determined by those leading from the first eukaryotes to the dinosaurs. This memoryless feature suggests the use of Markov chain based models of evolution. Recalling the notation $X^{(u)}$ for the random sequence associated with node $u$, we thus impose that for every phylogeny $\mathcal{P} = (V, E, \mathbb{P})$, and every sequence set $\mathcal{B} \subseteq \mathcal{S}$, if $u \prec v \prec w$,

$$\mathbb{P}\Big\{X^{(w)} \in \mathcal{B} \;\Big|\; X^{(v)}, X^{(u)}\Big\} = \mathbb{P}\Big\{X^{(w)} \in \mathcal{B} \;\Big|\; X^{(v)}\Big\}$$

with probability 1. In other words, the random taxon sequences form a Markov chain on every path in an evolutionary tree.

Another particular feature of evolution is that evolutionary changes along different branches are more or less independent from each other. One can argue nonetheless that the evolution of ant-eaters does very much depend on the evolution of ants, but at the molecular level, the simplification is acceptable (Kimura (1983), among others, argues powerfully in support of this assumption).

Based on the above discussion, we augment the definition of phylogeny by the following — for lack of a better word — axiom[1].

**Axiom 2.1.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny. For every edge $uv \in E$, the following holds. Let $k > 1$, and let $w_1, w_2, \ldots, w_k \in V - \{u, v\}$ be a collection of nodes such that none of them is in the subtree rooted at $v$, i.e., $v \not\prec w_i$ for every $i = 1, 2, \ldots, k$. Then for arbitrary sequence sets $\mathcal{B}_u, \mathcal{B}_v \subseteq \mathcal{S}$ with $\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\} \neq 0$,*

$$\mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u, X^{(w_1)}, \ldots, X^{(w_k)}\Big\} = \mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\}$$

*with probability* 1.

**Definition 2.2.** *Define the following set of distributions for every evolutionary tree.*

- *The* root sequence distribution *is defined as the marginal distribution of the random taxon sequence associated with the root.*
- *The* sequence transition probabilities *assigned to each edge $uv$ are defined by the conditional probability distribution of $X^{(v)}$ given $X^{(u)}$.*

Axiom 2.1 reveals an important property of the joint probability distribution of random taxon sequences, namely, that the joint distribution is fully determined by root sequence distribution and sequence transition probabilities.

[1]This axiom was introduced in a less general version by Steel (1994b).

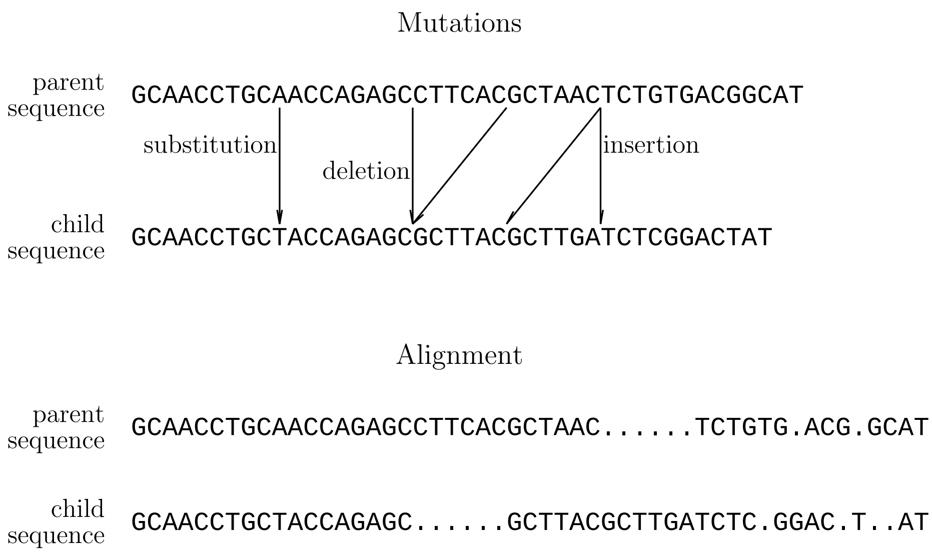



FIGURE 2.1: *Example of classifying sequence differences into three mutation categories to obtain an alignment.*

In light of the axiom, evolution can be interpreted as the broadcasting of a sequence from the root towards the leaves. The transmitted sequence undergoes certain random changes on each edge, governed by the sequence transition probabilities. The original sequence is randomly picked from the set $\mathcal{S}$ according to the root sequence distribution. The changes occurring to a sequence are called *mutations* in a biological context, and are usually classified in the following categories.

- *Substitutions* of characters in a sequence position.
- *Deletions* of characters from the parent sequence.
- *Insertions* of characters into the child sequence.

Decomposing the differences between two or more sequences into the above categories is a fundamental problem in computational biology. Such a decomposition results in an *alignment* of the sequences (see Figure 2.1). There is a vast body of research on sequence alignment problems; for a comprehensive overview see the work of Gusfield (1997) or Setubal and Meidanis (1997).

It is necessary to mention that there is a certain level of co-dependence between string alignment and evolutionary tree reconstruction. In order to align sequences generated by an evolutionary tree, we may need to recover the tree. On the other hand, most evolutionary tree reconstruction algorithms need an aligned set of sequences. Unfortunately, finding the optimal alignment for multiple sequences is NP-hard (Wang and Jiang 1994; Wang *et al.* 1996), with or without a phylogeny at hand, although there are many heuristics and suboptimal algorithms available (Gusfield 1997, Setubal and Meidanis 1997). In our study we simply assume that only substitutions occur on the edges, which corresponds to the practice of using aligned sequences as input to evolutionary tree reconstruction algorithms.

We show two properties of phylogenies in Lemma 2.1 and Theorem 2.2, which are useful in the later discussions. The following lemma generalizes the property described by Axiom 2.1 from parent-child to ancestor-descendant pairs.

**Lemma 2.1.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny. Let $u$, $v$ be an arbitrary ancestor-descendant pair, i.e., $u \prec v$, and let $u'$ denote the child of $u$ on the path towards $v$. Let $k > 1$ and let $w_1, w_2, \ldots, w_k \in V - \{u, v, u'\}$ be a collection of nodes such that none is in the subtree rooted at $u'$, i.e., $u' \not\prec w_i$ for every $i = 1, 2, \ldots, k$. Then for all sequence sets $\mathcal{B}_u, \mathcal{B}_v \subseteq \mathcal{S}$ with $\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\} \neq 0$,*

$$\mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u, X^{(w_1)}, \ldots, X^{(w_k)}\Big\} = \mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\}$$

*with probability* 1*.*

Proof. We prove the lemma by induction in the path length $l$ between $u$ and $v$.

*Base case*: $l = 1$. Then $u' = v$, and the lemma holds by Axiom 2.1.

*Induction step*: Let $l > 1$, and assume that the lemma holds for path lengths less than $l$. Let $\mathbf{X} = \langle X^{(w_1)}, \ldots, X^{(w_k)} \rangle$. By conditioning on $X^{(u')}$, and using the induction hypothesis,

$$\begin{aligned}
&\mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u, \mathbf{X}\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u')}, X^{(u)} \in \mathcal{B}_u, \mathbf{X}\Big\} \;\Big|\; X^{(u)} \in \mathcal{B}_u, \mathbf{X}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u')}, X^{(u)} \in \mathcal{B}_u, \Big\} \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big] \\
&= \mathbb{P}\Big\{X^{(v)} \in \mathcal{B}_v \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\}. \quad \blacksquare
\end{aligned}$$

The following theorem is crucial to our thesis. It establishes a relationship between the pairwise conditional probabilities of three random sequences associated with nodes that lie on a path in $\Psi(\mathcal{P})$.

**Theorem 2.2.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny. If $u, v, w \in V$ are three nodes such that $v$ lies on the path between $u$ and $w$ in $\Psi(\mathcal{P})$, then for all sequence sets $\mathcal{B}_u, \mathcal{B}_w \subseteq \mathcal{S}$ with $\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\} \neq 0$,*

$$\begin{aligned}
&\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\} \\
&= \sum_{s\colon \mathbb{P}\left\{X^{(v)}=s\right\}\neq 0} \mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(v)} = s\Big\}\mathbb{P}\Big\{X^{(v)} = s \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\}. \quad (2.1)
\end{aligned}$$

PROOF. Rewrite Equation (2.1) in the following equivalent form.

$$\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\} = \mathbb{E}\Big[\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(v)}\Big\} \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big].$$

If $w$ is a descendant of $v$, then

$$\begin{aligned}
&\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(v)}, X^{(u)} \in \mathcal{B}_u\Big\} \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big] \quad \text{(conditioning on } X^{(v)}\text{)} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \;\Big|\; X^{(v)}\Big\} \;\Big|\; X^{(u)} \in \mathcal{B}_u\Big]. \quad \text{(by Lemma 2.1)}
\end{aligned}$$

Otherwise, $u$ is a descendant of $v$. If $\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w\Big\} = 0$, then the theorem trivially holds. If $\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w\Big\} > 0$, then

$$\begin{aligned}
&\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \,\Big|\, X^{(u)} \in \mathcal{B}_u\Big\} \\
&= \frac{\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w\Big\}}{\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\}} \mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u \,\Big|\, X^{(w)} \in \mathcal{B}_w\Big\} \\
&= \frac{\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w\Big\}}{\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\}} \mathbb{E}\Big[\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u \,\Big|\, X^{(v)}, X^{(w)} \in \mathcal{B}_w\Big\} \,\Big|\, X^{(w)} \in \mathcal{B}_w\Big] \\
&= \frac{\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w\Big\}}{\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u\Big\}} \mathbb{E}\Big[\mathbb{P}\Big\{X^{(u)} \in \mathcal{B}_u \,\Big|\, X^{(v)}\Big\} \,\Big|\, X^{(w)} \in \mathcal{B}_w\Big] \quad \text{(by Lemma 2.1)} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_w \,\Big|\, X^{(v)}\Big\} \,\Big|\, X^{(u)} \in \mathcal{B}_u\Big]. \quad \blacksquare
\end{aligned}$$

## 2.3 The general Markov model

In the general Markov model the only mutations are independently distributed substitutions, i.e., each character of the broadcasted sequence undergoes changes independently. Deletions and insertions are not considered in this model. Consequently all character sequences have the same length $\ell$.

**Definition 2.3.** *The* general Markov class $\mathcal{C}_{\mathrm{M}}$ *is the set of every phylogeny* $\mathcal{P} = (V, E, \mathbb{P})$ *such that the sequence transition probabilities on each edge* $e = uv \in E$ *are defined by the* $m \times m$ edge mutation matrices $\{\mathbf{M}_e^{(k)} \colon k = 1, 2, \dots\}$ *in the following manner. For all* $\ell > 0$*, and length* $\ell$ *sequences* $s = s_1 \cdots s_\ell \in \mathcal{S}$ *and* $t = t_1 \cdots t_\ell \in \mathcal{S}$*,*

$$\mathbb{P}\Big\{X^{(v)} = t \,\Big|\, X^{(u)} = s\Big\} = \prod_{k=1}^{\ell} \mathbf{M}_e^{(k)}[s_k, t_k]. \tag{2.2}$$

We generalize the notion of a mutation matrix for every pair of nodes as the matrix of conditional probabilities on characters in the same positions of sequences associated with the two nodes.

**Definition 2.4.** *Let $u$ and $v$ be arbitrary nodes in an evolutionary tree $\mathcal{P}$. Let $X_k^{(u)}, X_k^{(u)}$ denote the $k$-th character of the random sequences associated with nodes $u$ and $v$, respectively. The $m \times m$ $\mathbf{M}_{uv}^{(k)}$ is defined by its entries for every $k > 0$ as*

$$\mathbf{M}_{uv}^{(k)}[i,j] = \begin{cases} \mathbb{P}\Big\{ X_k^{(v)} = j \,\Big|\, X_k^{(u)} = i \Big\} & \text{if } \mathbb{P}\left\{X_k^{(u)} = i\right\} \neq 0; \\ \mathbb{I}\{i = j\} & \text{otherwise.} \end{cases}$$

In particular, if $v$ is a child of $u$ on edge $e = uv$, then $\mathbf{M}_{uv}^{(k)}[i,j] = \mathbf{M}_e^{(k)}[i,j]$ in every sequence position $k$ and characters $i, j \in \mathcal{A}$ such that $\mathbb{P}\Big\{X_k^{(u)} = i\Big\} \neq 0$.

The next theorem is an application of Theorem 2.2 to the general Markov model, showing that mutation matrices multiply along paths.

**Theorem 2.3.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny of the general Markov class. Let $u, v, w \in V$ be three nodes such that $v$ lies on a path between $u$ and $w$ in $\Psi(\mathcal{P})$. In every sequence position $k > 0$, if $\mathbb{P}\Big\{X_k^{(u)} = i\Big\} \neq 0$ for every symbol $i \in \mathcal{A}$, then*

$$\mathbf{M}_{uv}^{(k)} \mathbf{M}_{vw}^{(k)} = \mathbf{M}_{uw}^{(k)}.$$

Proof. Let $i, i'$ be two arbitrary symbols of $\mathcal{A}$. We show that

$$\sum_{j \in \mathcal{A}} \mathbf{M}_{uv}^{(k)}[i,j] \mathbf{M}_{vw}^{(k)}[j,i'] = \mathbf{M}_{uw}^{(k)}[i,i'].$$

Define $\mathcal{B}_{k,j}$ for every $j \in \mathcal{A}$ to be the set of sequences in $\mathcal{S}$ in which the $k$-th character is $j$. By Theorem 2.2,

$$\begin{aligned}
\mathbf{M}_{uw}^{(k)}[i,i'] &= \mathbb{P}\Big\{X_k^{(w)} = i' \,\Big|\, X_k^{(u)} = i\Big\} = \mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_{k,i'} \,\Big|\, X^{(u)} \in \mathcal{B}_{k,i}\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_{k,i'} \,\Big|\, X^{(v)}\Big\} \,\Big|\, X^{(u)} \in \mathcal{B}_{k,i}\Big] \\
&= \sum_{j \in \mathcal{A}} \mathbb{P}\Big\{X^{(w)} \in \mathcal{B}_{k,i'} \,\Big|\, X_k^{(v)} = j\Big\} \mathbb{P}\Big\{X_k^{(v)} = j \,\Big|\, X^{(u)} \in \mathcal{B}_{k,i}\Big\} \\
&= \sum_{j \in \mathcal{A}} \mathbf{M}_{vw}^{(k)}[j,i'] \mathbf{M}_{uv}^{(k)}[i,j]. \quad \blacksquare
\end{aligned}$$

By expanding Theorem 2.3 to every node on a path we obtain the following corollary.

**Corollary 2.4.** *Let $u'$ be a descendant of $u$, and let $u, e_1, u_1, \ldots, e_l, u_l = u'$ be the path from $u$ to $u'$. In every sequence position $k > 0$, the following holds. Let $\mathbf{M}^{(k)} = \prod_{j=1}^{l} \mathbf{M}_{e_j}^{(k)}$. For every character pair $i, i' \in \mathcal{A}$,*

$$\mathbf{M}_{uu'}^{(k)}[i, i'] = \begin{cases} \mathbf{M}^{(k)}[i, i'] & \text{if } \mathbb{P}\Big\{X_k^{(u)} = i\Big\} \neq 0; \\ \mathbb{I}\{i = i'\} & \text{otherwise.} \end{cases} \tag{2.3}$$

*Consequently, for all $k, k' > 0$ and characters $i, i' \in \mathcal{A}$, if $\mathbf{M}^{(k)} = \mathbf{M}^{(k')}$, then*

$$\mathbf{M}_{uu'}^{(k)}[i, i'] = \mathbf{M}_{uu'}^{(k')}[i, i'], \tag{2.4}$$

*given that $\mathbb{P}\Big\{X_k^{(u)} = i\Big\}$ and $\mathbb{P}\Big\{X_{k'}^{(u)} = i\Big\}$ are both positive or both equal to zero.*

## 2.4 The i. i. d. Markov model

Even though Corollary 2.4 suggests that the transition matrices may not need to be indexed by the sequence position if the edge mutation matrices are constant, it is important to notice that Equation (2.4) does not hold generally unless $u \prec u'$. It does hold in general, however, if the root sequence consists of independent identically distributed (i.i.d.) characters. The class of phylogenies for which this is true is called the i. i. d. Markov class. This general model was introduced by Steel (1994b); its origins have been credited to Farris (1973) and Cavender (1978).

**Definition 2.5.** *The* i. i. d. Markov *model is a subclass of the general Markov model, in which for every phylogeny $\mathcal{P} = (V, E, \mathbb{P})$ the following hold:*

*(i) The edge mutation matrices are constant across sequence positions. In other words, on each edge $e \in E$ there exists $\mathbf{M}_e$ such that in every sequence position $k$, $\mathbf{M}_e^{(k)} = \mathbf{M}_e$.*

*(ii) The root sequence distribution is defined by a* sequence length distribution *over the positive integers, and fixed* root symbol frequencies *over the alphabet given by the vector*

$$\boldsymbol{\pi}^{(0)} = \langle \pi_1^{(0)} \cdots \pi_m^{(0)} \rangle.$$

*Each character of the root sequence $X^{(0)}$ is independent identically distributed according to $\boldsymbol{\pi}^{(0)}$. In other words, for every sequence length $\ell > 0$, and sequence $s_1 \cdots s_\ell \in \mathcal{S}$,*

$$\mathbb{P}\Big\{X^{(0)} = s_1 \cdots s_\ell\Big\} = \mathbb{P}\Big\{\Big|X^{(0)}\Big| = \ell\Big\} \prod_{k=1}^{\ell} \pi_{s_k}^{(0)}. \tag{2.5}$$

*(iii) For every sequence position $k > 0$, if $\mathbb{P}\Big\{\Big|X^{(0)}\Big| \geq k\Big\} \neq 0$, then for each node $u \in V$ and symbol $i \in \mathcal{A}$, $\mathbb{P}\Big\{X_k^{(u)} = i\Big\} \neq 0$.*

Condition (iii) is included in the definition in order to avoid unnecessary complications due to degenerate cases. It is satisfied, for example, if all the root base frequencies are positive and none of the edge mutation matrices have all-zero columns. When the root sequence has length $\ell$, the generated sequences can be viewed as the result of $\ell$ *random node labelings*. In each labeling the nodes are labeled by random symbols of the alphabet $\mathcal{A}$. The $k$-th character of each sequence is generated by the $k$-th labeling. Every labeling is carried out starting from the root and proceeding towards the leaves in the following manner. The root is labeled by a symbol drawn according to the distribution defined by the root symbol frequencies, so that for every $i \in \mathcal{A}$, $\mathbb{P}\{\text{root label} = i\} = \pi_i^{(0)}$. On edge $e$, the child's label is randomly selected based on the parent's label so that

$$\mathbb{P}\Big\{\text{child's label is } j \;\Big|\; \text{parent's label is } i\Big\} = \mathbf{M}_e[i,j].$$

In the i. i. d. Markov model, Equation (2.4) can be extended to any node pair as shown by the next lemma. Consequently, the mutation matrices do not need to be indexed by sequence position.

**Lemma 2.5.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. For all nodes $u, v \in V$, and sequence lengths $k' > k > 0$, if $\mathbb{P}\Big\{\Big|X^{(u)}\Big| \geq k'\Big\} \neq 0$, then*

$$\mathbf{M}_{uv}^{(k)} = \mathbf{M}_{uv}^{(k')}.$$

Proof. Let $w$ be the root of $\mathcal{P}$. We prove the lemma in three cases, depending on whether $u \prec v$, $v \prec u$, or neither.

*Case I.* Node $v$ is a descendant of node $u$. By Equation (2.5), and Definition 2.4,

$$\mathbb{P}\Big\{X_k^{(u)} = i\Big\} = \mathbb{P}\Big\{\Big|X^{(w)}\Big| \geq k\Big\} \sum_{j \in \mathcal{A}} \pi_j^{(0)} \mathbf{M}_{wu}^{(k)}[j, i];$$
$$\mathbb{P}\Big\{X_{k'}^{(u)} = i\Big\} = \mathbb{P}\Big\{\Big|X^{(w)}\Big| \geq k'\Big\} \sum_{j \in \mathcal{A}} \pi_j^{(0)} \mathbf{M}_{wu}^{(k')}[j, i].$$

Since $\Big|X^{(w)}\Big| = \Big|X^{(u)}\Big|$, and $\mathbb{P}\Big\{\Big|X^{(w)}\Big| \geq k\Big\} \geq \mathbb{P}\Big\{\Big|X^{(w)}\Big| \geq k'\Big\}$, $\mathbb{P}\Big\{X_k^{(u)} = i\Big\}$ and $\mathbb{P}\Big\{X_{k'}^{(u)} = i\Big\}$ are either both positive or both equal to zero. Hence $\mathbf{M}_{uv}^{(k)} = \mathbf{M}_{uv}^{(k')}$ is implied by Corollary 2.4. Notice that the equations hold even if $u$ is the root.

*Case II.* Node $u$ is a descendant of $v$. We show that for every two symbols $i, j \in \mathcal{A}$, $\mathbf{M}_{uv}^{(k)}[i, j] = \mathbf{M}_{uv}^{(k')}[i, j]$. Since $\mathbb{P}\Big\{\Big|X^{(u)}\Big| \geq k'\Big\} \neq 0$, and $k < k'$, $\mathbb{P}\Big\{X_k^{(u)} = i\Big\} \neq 0$ because $\mathbb{P}\Big\{X_{k'}^{(u)} = i\Big\} \neq 0$. Thus, by Definition 2.4,

$$\begin{aligned} \mathbf{M}_{uv}^{(k)}[i, j] &= \mathbb{P}\Big\{X_k^{(v)} = j \;\Big|\; X_k^{(u)} = i\Big\} \\ &= \frac{\mathbb{P}\Big\{X_k^{(u)} = i \;\Big|\; X_k^{(v)} = j\Big\}\mathbb{P}\Big\{X_k^{(v)} = j\Big\}}{\mathbb{P}\Big\{X_k^{(u)} = i\Big\}} \\ &= \mathbf{M}_{vu}^{(k)}[j, i] \frac{\sum_{j' \in \mathcal{A}} \pi_{j'}^{(0)} \mathbf{M}_{wv}^{(k)}[j', j]}{\sum_{j' \in \mathcal{A}} \pi_{j'}^{(0)} \mathbf{M}_{wu}^{(k)}[j', i]}. \end{aligned} \tag{$*$}$$

Similarly,

$$\mathbf{M}_{uv}^{(k')}[i, j] = \mathbf{M}_{vu}^{(k')}[j, i] \frac{\sum_{j' \in \mathcal{A}} \pi_{j'}^{(0)} \mathbf{M}_{wv}^{(k')}[j', j]}{\sum_{j' \in \mathcal{A}} \pi_{j'}^{(0)} \mathbf{M}_{wu}^{(k')}[j', i]}. \tag{$**$}$$

By Case I, and since $u$ and $v$ are descendants of the root $w$,

$$\mathbf{M}_{wu}^{(k)} = \mathbf{M}_{wu}^{(k')}, \qquad \mathbf{M}_{wv}^{(k)} = \mathbf{M}_{wv}^{(k')}, \quad \text{and} \quad \mathbf{M}_{vu}^{(k)} = \mathbf{M}_{vu}^{(k')}.$$

Thus by Equations ($*$) and ($**$), $\mathbf{M}_{uv}^{(k)}[i, j] = \mathbf{M}_{uv}^{(k')}[i, j]$.

*Case III.* Neither $u \prec v$ nor $v \prec u$. Let $v'$ be the lowest common ancestor of $u$ and $v$. By Cases I, II, and Theorem 2.3,

$$\mathbf{M}_{uv}^{(k)} = \mathbf{M}_{uv'}^{(k)} \mathbf{M}_{v'v}^{(k)} = \mathbf{M}_{uv'}^{(k')} \mathbf{M}_{v'v}^{(k')} = \mathbf{M}_{uv}^{(k')}. \quad \blacksquare$$

Subsequently we omit indexing on the sequence position.

**Definition 2.6.** *Define the* random taxon label $\xi^{(u)}$ *associated with each node $u$, as a random variable such that $\langle \xi^{(u)} \colon u \in V \rangle$ is distributed identically to $\langle X_1^{(u)} \colon u \in V \rangle$.*

*Furthermore, for every node pair $u, v \in V$, define the $m \times m$* mutation matrix *by its entries as*

$$\mathbf{M}_{uv}[i,j] = \mathbf{M}_{uv}^{(1)}[i,j] = \mathbb{P}\Big\{ \xi^{(v)} = j \;\Big|\; \xi^{(u)} = i \Big\}.$$

*Finally, for every node $w \in V$, define the* symbol frequencies

$$\pi_i^{(w)} = \mathbb{P}\Big\{ \xi^{(w)} = i \Big\},$$

*forming the vector of* taxon label distribution

$$\boldsymbol{\pi}^{(w)} = \langle \pi_1^{(w)}, \ldots, \pi_m^{(w)} \rangle.$$

By property (iii) of Definition 2.5, $\pi_i^{(w)} > 0$ for every node $w \in V$ and symbol $i \in \mathcal{A}$. Theorem 2.3 has the following corollaries in the i. i. d. Markov model.

**Corollary 2.6 (of Theorem 2.3).** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. Let $u, v, w \in V$ be three nodes such that $v$ lies on the path between $u$ and $w$ in $\Psi(\mathcal{P})$. Then*

$$\mathbf{M}_{uv} \mathbf{M}_{vw} = \mathbf{M}_{uw}.$$

**Corollary 2.7 (of Theorem 2.3).** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. Let $u, u' \in V$ be an arbitrary ancestor-descendant pair, and let $u, e_1, u_1, \ldots, e_l, u_l = u'$ be the path from $u$ to $u'$. Then*

$$\mathbf{M}_{uu'} = \prod_{k=1}^{l} \mathbf{M}_{e_k}.$$

## 2.5 Subclasses of the i. i. d. Markov model

Every edge mutation matrix is defined by $m(m-1)$ parameters in the i. i. d. Markov model. Taking the root symbol frequencies also into account, a phylogeny $\mathcal{P} = (V, E, \mathbb{P})$ is defined by $\Big(|E|m(m-1) + (m-1)\Big)$ free parameters. Several models have been proposed in which evolutionary trees are defined by fewer parameters. A survey and comparison of many such models are given among others by Swofford *et al.* (1996) and Zharkikh (1994). The main themes in the subclasses are (1) restricting the set of possible edge mutation matrices to a subset of stochastic matrices, which form an algebraic group, and (2) restricting the root label distribution. This section reviews the most common restrictions made in the literature of molecular evolution.

**Constant substitution rates** From the evolutionary biologist's viewpoint, it is important to relate the mutation matrices to a time scale. Phylogenies used in evolutionary biology have weighted edges, and the edge weights correspond to time between speciation events. Evolutionary time is incorporated into the i. i. d. Markov model by defining the mutation matrix on each edge $e$ as the transition matrix of a time-invariant Markov process running for a certain time $\tau_e$. When the Markov processes running on the edges have the same instantaneous transition matrix $\mathbf{Q}$, then $\mathbf{M}_e = \exp(\mathbf{Q}\tau_e)$ on every edge $e$. The matrix $\mathbf{Q}$ is the *constant substitution rate* matrix. This assumption lies at the heart of the *molecular clock* theory, originating from Zuckerkandl and Pauling (1962, 1965) and Margoliash (1963). The theory has since raised much debate (e.g., Goodman 1976; Wilson *et al.* 1977; Kimura 1981b; Li and Gojobori 1983; Wu and Li 1985; Ayala 1997). Many special subclasses of the i. i. d. Markov model have been defined by specifying constant substitution rates with some particular features. A discussion of such models focusing on substitution rates was conducted by Tavaré (1986) and Rodríguez *et al.* (1990). Constant substitution rates with no restriction on the rate matrix were introduced by Lanave *et al.* (1984) and Barry and Hartigan (1987).

In some cases canonical substitution rate matrices are measured in laboratory conditions and then used for estimating evolutionary time. The most important examples are those of the "point accepted mutation" or PAM matrices (Dayhoff, Schwartz, and Orcutt 1978), and the BLOSUM matrices (Henikoff and Henikoff 1992), both for amino acid sequences.

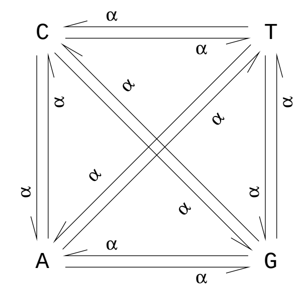
**Stationary taxon label distribution** An interesting special case of the i. i. d. Markov model occurs when the edge transition matrices share the same stationary distribution, which is in turn the root symbol distribution; i.e., $\boldsymbol{\pi}^{(0)}\mathbf{M}_e = \boldsymbol{\pi}^{(0)}$ on each edge $e$. As a consequence, the random labels are identically distributed in the phylogeny. This assumption is usually made in conjunction with the constant substitution rate assumption, in which case $\boldsymbol{\pi}^{(0)}\mathbf{Q} = \mathbf{0}$ holds.

**General time-reversible model** A subclass of the i. i. d. Markov model often found in the literature (see, e.g., Lanave *et al.* 1984 and Tavaré 1986) comprises phylogenies in which the random taxon labels form a time-reversible Markov chain along any path. In particular, for a phylogeny $\mathcal{P} = (V, E, \mathbb{P})$ in the general time-reversible model,

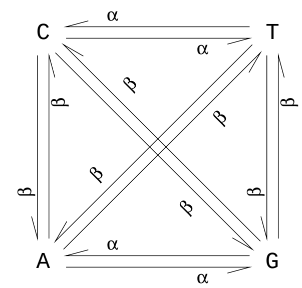
$$\mathbf{M}_{uv}[i,j]\pi_i^{(u)} = \mathbf{M}_{uv}[j,i]\pi_j^{(u)}$$

for every edge $uv \in E$ and all symbols $i, j \in \mathcal{A}$. Consequently, every mutation matrix can be written as $\mathbf{M}_{uv} = \mathbf{M}^*_{uv} \operatorname{diag}\left(\boldsymbol{\pi}^{(u)}\right)$, where $\operatorname{diag}\left(\boldsymbol{\pi}^{(u)}\right)$ is the diagonal matrix derived from $\boldsymbol{\pi}^{(u)}$ and $\mathbf{M}^*_{uv}$ is a symmetric matrix (Zharkikh 1994). Furthermore, time reversibility implies that on every edge $uv$, $\boldsymbol{\pi}^{(u)} = \boldsymbol{\pi}^{(v)}$, i.e., that the random taxon labels are distributed identically.

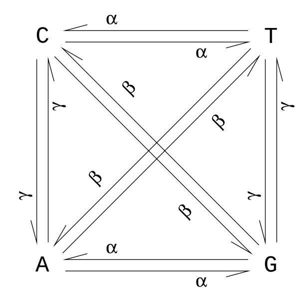
**Transition matrices for modeling molecular evolution** Molecular evolutionary models often reduce the number of free parameters in transition matrices, based on either biological or computational considerations. We have already mentioned the PAM matrices (Dayhoff *et al.* 1978) that model amino acid substitutions. There is also a number of nucleotide substitution rate models discussed in the literature and routinely used in practice. Such models were introduced predominantly with the assumptions of constant substitution rates and a stationary taxon label distribution. These assumptions are not always necessary, and we attempt to generalize the models to the time-independent framework used so far, similarly to Barry and Hartigan (1987) and Zharkikh (1994). Figure 2.2 summarizes the particularities of many substitution rate models. The set of models shown in the figure is not exhaustive, but it includes the most commonly used models incorporated in standard molecular phylogeny packages such as Phylip (Felsenstein 1993) and PAUP (Swofford 1990).



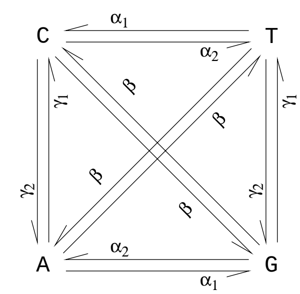

Jukes-Cantor (JC) model (Jukes and Cantor 1969).



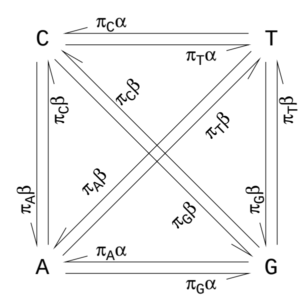

Kimura's two parameter (K2P) model (Kimura 1980).



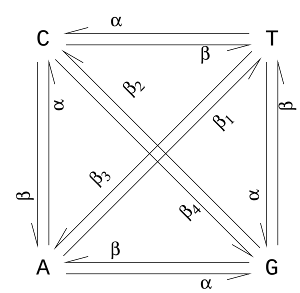

Kimura's three parameter (K3P) model (Kimura 1981a).




Five parameter (TK) model of (Takahata and Kimura 1981).




Five parameter (HKY) model with $\pi_{\mathtt{A}} + \pi_{\mathtt{G}} + \pi_{\mathtt{T}} + \pi_{\mathtt{C}} = 1$ (Hasegawa, Kishino, and Yano 1985).




Six parameter (GIN) model (Gojobori, Ishii, and Nei 1982; Kimura 1981a).

FIGURE 2.2: *Common nucleotide substitution rate models.*

### 2.5.1 Jukes-Cantor model

The first nucleotide substitution rate model was proposed by Jukes and Cantor (1969). It has later been generalized by Neyman (1971) to an arbitrary alphabet, omitting the constant substitution rate assumption. We follow Neyman's proposal and present his generalized model. Each edge mutation matrix in this model has solely one free parameter, so that there exists $0 \le p_e \le 1$ for each edge $e$ such that

$$\mathbf{M}_e[i,j] = \begin{cases} \frac{p_e}{m-1} & \text{if } i \neq j; \\ 1 - p_e & \text{if } i = j. \end{cases}$$

Despite its simplicity, the Jukes-Cantor model is very well-suited to discuss the characteristics of the i. i. d. Markov model (of which we do take advantage). Due to the fact that one parameter defines the mutation matrix, many calculations are easier to carry out than in the case of more general models. To illustrate our point, we prove the following lemma (see also, e.g., Farach and Kannan 1999).

**Lemma 2.8.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the Jukes-Cantor model. For every ancestor-descendant node pair $u \prec v$, there exists $0 \le p_{uv} \le 1$ such that*

$$\mathbf{M}_{uv}[i,j] = \begin{cases} \frac{p_{uv}}{m-1} & \textit{if } i \neq j; \\ 1 - p_{uv} & \textit{if } i = j. \end{cases}$$

*In particular, if the path leading from $u$ to $v$ is $u, e_1, u_1, e_2, \dots, e_l, u_l = v$, then*

$$1 - \frac{m}{m-1} p_{uv} = \prod_{k=1}^{l} \Bigl(1 - \frac{m}{m-1} p_{e_k}\Bigr). \tag{2.6}$$

PROOF. We prove the lemma by induction in the path length $l$.

*Base case.* If $l = 1$, then $v$ is a child of $u$ and the lemma is true by definition of the edge mutation matrices in this model.

*Induction step.* If $l > 1$, then let $v'$ be the parent of $v$, $e = uv'$, and assume that the lemma holds for $u \prec v'$. Let $i \neq j$ be two symbols of the alphabet. Since $\mathbb{P}\left\{\xi^{(u)} = i\right\} \neq 0$,

$$\begin{aligned}
\mathbf{M}_{uv}[i,j] &= \mathbb{P}\left\{\xi^{(v)} = j \;\middle|\; \xi^{(u)} = i\right\} \\
&= \sum_{k\in\mathcal{A}} \mathbb{P}\left\{\xi^{(v)} = j \;\middle|\; \xi^{(v')} = k\right\}\mathbb{P}\left\{\xi^{(v')} = k \;\middle|\; \xi^{(u)} = i\right\} \\
&= (1 - p_{uv'})\frac{p_e}{m-1} \qquad (k = i, k \neq j) \\
&\quad + \frac{p_{uv'}}{m-1}(1 - p_e) \qquad (k \neq i, k = j) \\
&\quad + (m-2)\frac{p_{uv'}}{m-1}\frac{p_e}{m-1} \qquad (k \neq i, k \neq j) \\
&= \frac{p_{uv'}}{m-1} + \frac{p_e}{m-1} - m\frac{p_{uv'}p_e}{(m-1)^2}.
\end{aligned}$$

Thus

$$\left(1 - \frac{m}{m-1}\mathbf{M}_{uv}[i,j]\right) = \left(1 - \frac{m}{m-1}p_{uv'}\right)\left(1 - \frac{m}{m-1}p_e\right),$$

and the lemma follows from the induction hypothesis. ■

**Definition 2.7.** *Let $\mathcal{M}_{\mathrm{JC}}$ be the class of $m \times m$ stochastic matrices defined as follows. A stochastic matrix $\mathbf{M}$ belongs to $\mathcal{M}_{\mathrm{JC}}$ if and only if there exists $0 \leq p \leq 1$ such that the entries of $\mathbf{M}$ can be written as*

$$\mathbf{M}[i,j] = \begin{cases} \frac{p}{m-1} & \text{if } i \neq j; \\ 1 - p & \text{if } i = j. \end{cases}$$

*The hypothesis class $\mathcal{C}_{\mathrm{JC}}$ is the set of phylogenies in the i. i. d. Markov model in which every edge mutation matrix belongs to $\mathcal{M}_{\mathrm{JC}}$.*

**Lemma 2.9.** *The set $\mathcal{M}_{\mathrm{JC}}$ is closed under matrix multiplication.*

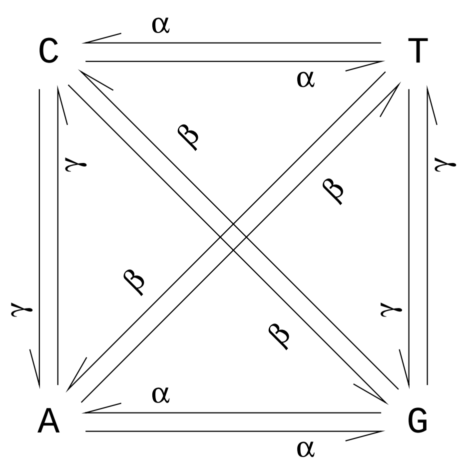
Proof. The lemma follows from Lemma 2.8. ■

Another useful implication of Lemma 2.8 is the following corollary.

**Corollary 2.10.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the Jukes-Cantor* *model. For every three nodes $u, v, w \in V$, if $u \prec v \prec w$, then*

$$1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(w)}\Big\} = \left(1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}\right)\left(1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(v)} \neq \xi^{(w)}\Big\}\right).$$

### 2.5.2 Kimura's two and three parameter models




Kimura's three parameter model (Kimura 1981a) distinguishes three types of substitutions: purine-purine or pyrimidine-pyrimidine transitions, and two types of purine-pyrimidine transversions, with substitution rates $\alpha$, $\beta$, and $\gamma$, respectively. In Kimura's two parameter model, all transversions are equivalent, i.e., $\beta = \gamma$. The Jukes-Cantor model can also be considered as a specific case of the three parameter model with $\alpha = \beta = \gamma$.

Zharkikh (1994) points out a weakness in Kimura's models and the Jukes-Cantor model: namely, that if the mutation matrices are symmetric, then the stationary taxon label distribution is uniform at every node. In other words, the eigenvector of the substitution rate matrix with eigenvalue 0 is $\langle 1/4, 1/4, 1/4, 1/4 \rangle$.

**Fact 2.11.** *Let $\alpha > 0$, $\beta > 0$, $\gamma > 0$, and define the matrix $\mathbf{Q}$ as*

$$\mathbf{Q} = \begin{bmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{bmatrix}.$$

*The spectral decomposition of* $\mathbf{Q}$ *can then be written as*

$$\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}$$

*with*

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix},$$

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(\alpha+\beta) & 0 & 0 \\ 0 & 0 & -2(\alpha+\gamma) & 0 \\ 0 & 0 & 0 & -2(\beta+\gamma) \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{bmatrix}.$$

Kimura's three parameter model gives a good opportunity to illustrate that the molecular clock assumption, i.e., the use of substitution rate matrices, results in loss of generality. Given the spectral decomposition of a substitution rate matrix $\mathbf{Q}$ as shown by Fact 2.11, the corresponding mutation matrices $\{\mathbf{M}_\tau \colon \tau \geq 0\}$ can be written as

$$\mathbf{M}_\tau = e^{\mathbf{Q}\tau} = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix} \tag{2.7a}$$

with

$$\begin{aligned} p &= \frac{1}{4}\Big(1 - e^{-2(\alpha+\beta)\tau} - e^{-2(\alpha+\gamma)\tau} + e^{-2(\beta+\gamma)\tau}\Big), \\ q &= \frac{1}{4}\Big(1 - e^{-2(\alpha+\beta)\tau} + e^{-2(\alpha+\gamma)\tau} - e^{-2(\beta+\gamma)\tau}\Big), \\ r &= \frac{1}{4}\Big(1 + e^{-2(\alpha+\beta)\tau} - e^{-2(\alpha+\gamma)\tau} - e^{-2(\beta+\gamma)\tau}\Big). \end{aligned} \tag{2.7b}$$

If complex substitution rates are not allowed, then not every stochastic matrix of the form on the right-hand side of Equation(2.7a) has a corresponding substitution rate matrix. For example, Equation (2.7b) implies that $p + q + r < 3/4$.

The extensions of Kimura's models in the next two definitions and the lemma are self-explanatory.

**Definition 2.8.** *Let $\mathcal{M}_{\mathrm{K2P}}$ be the class of $4 \times 4$ stochastic matrices defined as follows. A matrix $\mathbf{M}$ belongs to $\mathcal{M}_{\mathrm{K2P}}$ if and only if there exists $p, q \geq 0$ with $p + 2q \leq 1$ such that*

$$\mathbf{M} = \begin{bmatrix} 1-p-2q & p & q & q \\ p & 1-p-2q & q & q \\ q & q & 1-p-2q & p \\ q & q & p & 1-p-2q \end{bmatrix}.$$

*The hypothesis class $\mathcal{C}_{\mathrm{K2P}}$ is the set of phylogenies in the i. i. d. Markov model in which every edge mutation matrix belongs to $\mathcal{M}_{\mathrm{K2P}}$.*

**Definition 2.9.** *Let $\mathcal{M}_{\text{K3P}}$ be the class of $4 \times 4$ stochastic matrices defined as follows. A matrix $\mathbf{M}$ belongs to $\mathcal{M}_{\text{K3P}}$ if and only if there exists $p, q, r \geq 0$ with $p + q + r \leq 1$ such that*

$$\mathbf{M} = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix}$$

*The hypothesis class $\mathcal{C}_{\text{K3P}}$ is the set of phylogenies in the i. i. d. Markov model in which every edge mutation matrix belongs to $\mathcal{M}_{\text{K3P}}$.*

**Lemma 2.12.** *The sets $\mathcal{M}_{\text{K2P}}$ and $\mathcal{M}_{\text{K3P}}$ are closed under matrix multiplication.*

Proof. Let $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{\text{K3P}}$ with

$$\mathbf{M} = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix},$$

$$\mathbf{M}' = \begin{bmatrix} 1-p'-q'-r' & p' & q' & r' \\ p' & 1-p'-q'-r' & r' & q' \\ q' & r' & 1-p'-q'-r' & p' \\ r' & q' & p' & 1-p'-q'-r' \end{bmatrix}.$$

Using the spectral decomposition of $\mathbf{M}$ and $\mathbf{M}'$,

$$\mathbf{M} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U}$$

with

$$\mathbf{U} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{bmatrix},$$

$$\mathbf{\Lambda} = \mathrm{diag}(1, 1-2p-2q, 1-2p-2r, 1-2q-2r),$$

and

$$\mathbf{M} = \mathbf{U}^{-1}\mathbf{\Lambda}'\mathbf{U}$$

with

$$\mathbf{\Lambda}' = \mathrm{diag}(1, 1-2p'-2q', 1-2p'-2r', 1-2q'-2r').$$

Consequently,

$$\mathbf{M}'' = \mathbf{M}\mathbf{M}' = \mathbf{U}^{-1}(\mathbf{\Lambda}\mathbf{\Lambda}')\mathbf{U}.$$

Expanding the matrix multiplication on the right-hand side,

$$\mathbf{M}'' = \begin{bmatrix} 1-p''-q''-r'' & p'' & q'' & r'' \\ p'' & 1-p''-q''-r'' & r'' & q'' \\ q'' & r'' & 1-p''-q''-r'' & p'' \\ r'' & q'' & p'' & 1-p''-q''-r'' \end{bmatrix}$$

with

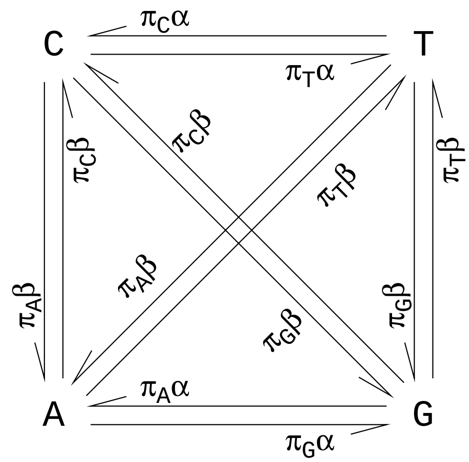
$$\begin{aligned} (1-2p''-2q'') &= (1-2p-2q)(1-2p'-2q'), \\ (1-2p''-2r'') &= (1-2p-2r)(1-2p'-2r'), \\ (1-2q''-2r'') &= (1-2q-2r)(1-2q'-2r'), \end{aligned}$$

and thus $\mathbf{M}'' \in \mathcal{M}_{\mathrm{K3P}}$. The claim for $\mathcal{M}_{\mathrm{K2P}}$ is proven similarly. ■

### 2.5.3 Asymmetric mutation models

As we mentioned before, the major drawback of substitution rate models with symmetric matrices is that the stationary taxon label distribution is uniform. Since nucleotide frequencies are often unequal in a gene, several models have been proposed that allow for more general stationary distributions. Hasegawa, Kishino, and Yano (1985), for example, proposed a model based on Kimura's two parameter model that explicitly incorporates arbitrary base frequencies. The five parameter nucleotide substitution model of Takahata and Kimura (1981) on the other hand implies a stationary distribution in which $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ and $\pi_{\mathtt{G}} = \pi_{\mathtt{C}}$, i.e., adenine and thymine have the same frequencies, and guanine and cytosine frequencies are also equal. In practice the substitution rates are estimated from the mutation matrices, which are in turn estimated from the sequence samples at hand, in a more or less straightforward manner. The dependencies between the substitution rates and the mutation matrix on an edge are implied by the substitution rate model. With the assumption of time-reversibility such dependencies apply to any node pair, since the mutation matrices multiply along any path (cf. Corollary 2.7), and the set $\{\exp(\mathbf{Q}\tau)\colon \tau \geq 0\}$ is closed under matrix multiplication. The closedness of the class is always desirable with any substitution rate model so that the particularities of the model can be exploited for any node pair and not only for neighbors. By Lemmas 2.9 and 2.12, the constant substitution rate assumption can be omitted in the Jukes-Cantor model and in Kimura's two and three parameter models, while preserving the closeness of the matrix classes. Relaxing the assumption in more complicated substitution models, however, is not always straightforward. We discuss the difficulties in detail in conjunction with the Hasegawa-Kishino-Yano and Gojobori-Ishii-Nei models. The substitution model of Takahata and Kimura (1981) does not seem to offer an elegant generalization and we prefer to avoid a lengthy technical analysis.

### 2.5.4 Hasegawa-Kishino-Yano model




The five parameter model of Hasegawa, Kishino, and Yano (1985) generalizes Kimura's two parameter model so that arbitrary stationary base frequencies $(\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}})$ are allowed. Two types of mutations are considered: transitions $(\alpha)$ and transversions $(\beta)$.

**Fact 2.13.** *Let $\alpha, \beta > 0$,*

$$\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}} > 0, \qquad \textit{and} \qquad \pi_{\mathtt{A}} + \pi_{\mathtt{G}} + \pi_{\mathtt{T}} + \pi_{\mathtt{C}} = 1,$$

*and define the matrix $\mathbf{Q}$ as*

$$\mathbf{Q} = \begin{bmatrix} \mu_1 & \pi_{\mathtt{G}}\alpha & \pi_{\mathtt{T}}\beta & \pi_{\mathtt{C}}\beta \\ \pi_{\mathtt{A}}\alpha & \mu_2 & \pi_{\mathtt{T}}\beta & \pi_{\mathtt{C}}\beta \\ \pi_{\mathtt{A}}\beta & \pi_{\mathtt{G}}\beta & \mu_3 & \pi_{\mathtt{C}}\alpha \\ \pi_{\mathtt{A}}\beta & \pi_{\mathtt{G}}\beta & \pi_{\mathtt{T}}\alpha & \mu_4 \end{bmatrix}$$

*with $\mu_i = -\sum_{j \neq i} \mathbf{Q}[i,j]$. The spectral decomposition of $\mathbf{Q}$ can then be written as*

$$\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}$$

*with*

$$\mathbf{V} = \begin{bmatrix} 1 & \pi_{\mathtt{T}} + \pi_{\mathtt{C}} & \frac{\pi_{\mathtt{G}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & 0 \\ 1 & \pi_{\mathtt{T}} + \pi_{\mathtt{C}} & -\frac{\pi_{\mathtt{A}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & 0 \\ 1 & -(\pi_{\mathtt{A}} + \pi_{\mathtt{G}}) & 0 & \frac{\pi_{\mathtt{C}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} \\ 1 & -(\pi_{\mathtt{A}} + \pi_{\mathtt{G}}) & 0 & -\frac{\pi_{\mathtt{T}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} \end{bmatrix},$$

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\beta & 0 & 0 \\ 0 & 0 & -(\pi_{\mathtt{A}} + \pi_{\mathtt{G}})\alpha - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})\beta & 0 \\ 0 & 0 & 0 & -(\pi_{\mathtt{A}} + \pi_{\mathtt{G}})\beta - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})\alpha \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} \pi_{\mathtt{A}} & \pi_{\mathtt{G}} & \pi_{\mathtt{T}} & \pi_{\mathtt{C}} \\ \frac{\pi_{\mathtt{A}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & \frac{\pi_{\mathtt{G}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & -\frac{\pi_{\mathtt{T}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} & -\frac{\pi_{\mathtt{C}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

When the substitution rate matrix has the form given by Fact 2.13, the corresponding edge mutation matrices $\{\mathbf{M}_\tau : \tau \geq 0\}$ can be written as

$$\mathbf{M}_\tau = e^{\mathbf{Q}\tau} = \begin{bmatrix} \nu_1 & \pi_{\mathtt{G}} p_1 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} p_1 & \nu_2 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \nu_3 & \pi_{\mathtt{C}} p_2 \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \pi_{\mathtt{T}} p_2 & \nu_4 \end{bmatrix}, \tag{2.8a}$$

$$\nu_i = 1 - \sum_{j \neq i} \mathbf{M}_\tau[i, j],$$

with

$$
\begin{aligned}
p_1 &= \frac{(\pi_{\mathtt{A}} + \pi_{\mathtt{G}}) + (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})e^{-\beta\tau} - e^{-((\pi_{\mathtt{A}}+\pi_{\mathtt{G}})\alpha + (\pi_{\mathtt{T}}+\pi_{\mathtt{C}})\beta)\tau}}{\pi_{\mathtt{A}} + \pi_{\mathtt{G}}}, \\
p_2 &= \frac{(\pi_{\mathtt{T}} + \pi_{\mathtt{C}}) + (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})e^{-\beta\tau} - e^{-((\pi_{\mathtt{T}}+\pi_{\mathtt{C}})\alpha + (\pi_{\mathtt{A}}+\pi_{\mathtt{G}})\beta)\tau}}{\pi_{\mathtt{T}} + \pi_{\mathtt{C}}}, \\
q &= 1 - e^{-\beta\tau}.
\end{aligned}
\tag{2.8b}
$$

We mention in passing that Zharkikh (1994) makes a mistake in presenting the model by asserting $p_1 = p_2$, which is not true in general. Using the model without assuming constant substitution rates presents some difficulties: namely, that stochastic matrices in the form of the right-hand side in Equation (2.8a) do not form a closed set under matrix multiplication. On the other hand, since the eigenvectors depend only on the stationary distribution, matrices with fixed $\langle \pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}} \rangle$ do form a closed set.

**Definition 2.10.** *Let*

$$\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}} > 0, \qquad \pi_{\mathtt{A}} + \pi_{\mathtt{G}} + \pi_{\mathtt{T}} + \pi_{\mathtt{C}} = 1, \qquad \boldsymbol{\pi} = \langle \pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}} \rangle.$$

*Let* $\mathcal{M}_{\mathrm{HKY}(\boldsymbol{\pi})}$ *be the class of* $4 \times 4$ *matrices defined as follows. A stochastic matrix* $\mathbf{M}$ *belongs to* $\mathcal{M}_{\mathrm{HKY}(\boldsymbol{\pi})}$ *if and only if there exists* $p_1, p_2, q \geq 0$ *such that*

$$\mathbf{M} = \begin{bmatrix} \nu_1 & \pi_{\mathtt{G}} p_1 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} p_1 & \nu_2 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \nu_3 & \pi_{\mathtt{C}} p_2 \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \pi_{\mathtt{T}} p_2 & \nu_4 \end{bmatrix},$$

*with* $\nu_i = 1 - \sum_{j \neq i} \mathbf{M}[i,j]$.

**Lemma 2.14.** *The set* $\mathcal{M}_{\mathrm{HKY}(\boldsymbol{\pi})}$ *is closed under matrix multiplication.*

Proof. Let $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{\mathrm{HKY}(\boldsymbol{\pi})}$ be arbitrary matrices with

$$\mathbf{M} = \begin{bmatrix} \nu_1 & \pi_{\mathtt{G}} p_1 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} p_1 & \nu_2 & \pi_{\mathtt{T}} q & \pi_{\mathtt{C}} q \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \nu_3 & \pi_{\mathtt{C}} p_2 \\ \pi_{\mathtt{A}} q & \pi_{\mathtt{G}} q & \pi_{\mathtt{T}} p_2 & \nu_4 \end{bmatrix}, \text{where } \nu_i = 1 - \sum_{j \neq i} \mathbf{M}[i,j], \text{ and}$$

$$\mathbf{M}' = \begin{bmatrix} \nu_1 & \pi_{\mathtt{G}} p_1' & \pi_{\mathtt{T}} q' & \pi_{\mathtt{C}} q' \\ \pi_{\mathtt{A}} p_1' & \nu_2 & \pi_{\mathtt{T}} q' & \pi_{\mathtt{C}} q' \\ \pi_{\mathtt{A}} q' & \pi_{\mathtt{G}} q' & \nu_3 & \pi_{\mathtt{C}} p_2' \\ \pi_{\mathtt{A}} q' & \pi_{\mathtt{G}} q' & \pi_{\mathtt{T}} p_2' & \nu_4 \end{bmatrix}, \text{where } \nu_i' = 1 - \sum_{j \neq i} \mathbf{M}'[i,j].$$

Using the spectral decomposition of $\mathbf{M}$ and $\mathbf{M}'$,

$$\mathbf{M} = \mathbf{U}^{-1} \boldsymbol{\Lambda} \mathbf{U}$$

with

$$\mathbf{U} = \begin{bmatrix} \pi_{\mathtt{A}} & \pi_{\mathtt{G}} & \pi_{\mathtt{T}} & \pi_{\mathtt{C}} \\ \frac{\pi_{\mathtt{A}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & \frac{\pi_{\mathtt{G}}}{\pi_{\mathtt{A}}+\pi_{\mathtt{G}}} & -\frac{\pi_{\mathtt{T}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} & -\frac{\pi_{\mathtt{C}}}{\pi_{\mathtt{T}}+\pi_{\mathtt{C}}} \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\Lambda} = \mathrm{diag}\Big(1, 1-q, 1-(\pi_{\mathtt{A}}+\pi_{\mathtt{G}})p_1 - (\pi_{\mathtt{T}}+\pi_{\mathtt{C}})q, 1-(\pi_{\mathtt{T}}+\pi_{\mathtt{C}})p_2 - (\pi_{\mathtt{A}}+\pi_{\mathtt{G}})q\Big),$$

and

$$\mathbf{M}' = \mathbf{U}^{-1} \boldsymbol{\Lambda}' \mathbf{U},$$

with

$$\boldsymbol{\Lambda}' = \mathrm{diag}\Big(1, 1-q', 1-(\pi_{\mathtt{A}}+\pi_{\mathtt{G}})p_1' - (\pi_{\mathtt{T}}+\pi_{\mathtt{C}})q', 1-(\pi_{\mathtt{T}}+\pi_{\mathtt{C}})p_2' - (\pi_{\mathtt{A}}+\pi_{\mathtt{G}})q'\Big).$$

Subsequently

$$\mathbf{M}'' = \mathbf{M}\mathbf{M}' = \mathbf{U}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')\mathbf{U}.$$

Expanding the matrix multiplication on the right-hand side,

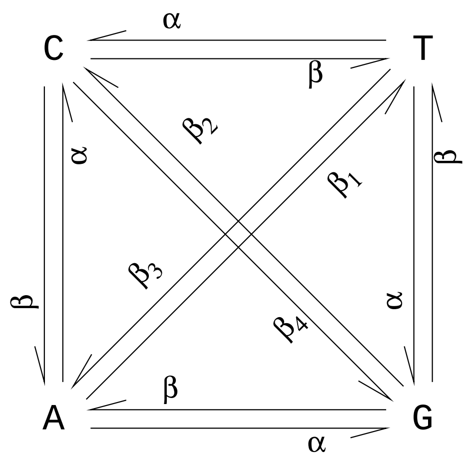
$$\mathbf{M}'' = \begin{bmatrix} \nu_1 & \pi_{\mathtt{G}}p_1'' & \pi_{\mathtt{T}}q'' & \pi_{\mathtt{C}}q'' \\ \pi_{\mathtt{A}}p_1'' & \nu_2 & \pi_{\mathtt{T}}q'' & \pi_{\mathtt{C}}q'' \\ \pi_{\mathtt{A}}q'' & \pi_{\mathtt{G}}q'' & \nu_3 & \pi_{\mathtt{C}}p_2'' \\ \pi_{\mathtt{A}}q'' & \pi_{\mathtt{G}}q'' & \pi_{\mathtt{T}}p_2'' & \nu_4 \end{bmatrix}, \text{where } \nu_i'' = 1 - \sum_{j \neq i} \mathbf{M}''[i,j]$$

with

$$\begin{aligned}
&(1 - q'') = (1 - q)(1 - q'), \\
&\Big(1 - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})p_1'' - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})q''\Big) \\
&= \Big(1 - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})p_1 - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})q\Big)\Big(1 - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})p_1' - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})q'\Big), \\
&\Big(1 - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})p_2'' - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})q''\Big) \\
&= \Big(1 - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})p_2 - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})q\Big)\Big(1 - (\pi_{\mathtt{T}} + \pi_{\mathtt{C}})p_2' - (\pi_{\mathtt{A}} + \pi_{\mathtt{G}})q'\Big).
\end{aligned}$$

Hence the stochastic matrix $\mathbf{M}''$ is an element of $\mathcal{M}_{\mathrm{HKY}(\boldsymbol{\pi})}$. ■

### 2.5.5 Gojobori-Ishii-Nei model




Kimura (1981a) introduced a substitution rate model with six parameters that was further analyzed later by Gojobori, Ishii, and Nei (1982). The latter study reported the eigenvalues and eigenvectors of the substitution rate matrix and related them to the entries of the corresponding mutation matrix.

**Fact 2.15.** *Let $\alpha, \beta, \beta_1, \beta_2, \beta_3, \beta_4 > 0$, and define the matrix* $\mathbf{Q}$ *as*

$$\mathbf{Q} = \begin{bmatrix} -(2\alpha+\beta_1) & \alpha & \beta_1 & \alpha \\ \beta & -(2\beta+\beta_2) & \beta & \beta_2 \\ \beta_3 & \alpha & -(2\alpha+\beta_3) & \alpha \\ \beta & \beta_4 & \beta & -(2\beta+\beta_4) \end{bmatrix}.$$

*The spectral decomposition of* $\mathbf{Q}$ *can be written as*

$$\mathbf{Q} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U}$$

*with*

$$\mathbf{U} = \begin{bmatrix} \frac{(\alpha+\beta_3)\beta}{(2\alpha+\beta_1+\beta_3)(\alpha+\beta)} & \frac{(\beta+\beta_4)\alpha}{(2\beta+\beta_2+\beta_4)(\alpha+\beta)} & \frac{(\alpha+\beta_1)\beta}{(2\alpha+\beta_1+\beta_3)(\alpha+\beta)} & \frac{(\beta+\beta_2)\alpha}{(2\beta+\beta_2+\beta_4)(\alpha+\beta)} \\ \frac{\beta-\beta_3}{2\beta-\beta_1-\beta_3} & -\frac{\alpha-\beta_4}{2\alpha-\beta_2-\beta_4} & \frac{\beta-\beta_1}{2\beta-\beta_1-\beta_3} & -\frac{\alpha-\beta_2}{2\alpha-\beta_2-\beta_4} \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix},$$

$$\mathbf{\Lambda} = \mathrm{diag}(0, -2\alpha-2\beta, -2\alpha-\beta_1-\beta_3, -2\beta-\beta_2-\beta_4).$$

Based on the spectral decomposition given by Fact 2.15 we can obtain the set of edge mutation matrices

$$\left\{\mathbf{U}^{-1}e^{\mathbf{\Lambda}\tau}\mathbf{U} \colon \tau \geq 0\right\}$$

corresponding to the rate matrix $\mathbf{Q} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U}$. Carrying out the expansion shows that the edge mutation matrices have the form

$$\mathbf{M} = \begin{bmatrix} 1-p_1-p_2-r_1 & p_1 & r_1 & p_2 \\ q_1 & 1-q_1-q_2-r_2 & q_2 & r_2 \\ r_3 & p_1 & 1-p_1-p_2-r_3 & p_2 \\ q_1 & r_4 & q_2 & 1-q_1-q_2-r_4 \end{bmatrix}. \tag{2.9}$$

The parameters $p_i, q_i, r_i$ are not independent, each one of them is a function of seven variables — the six parameters of $\mathbf{Q}$ and the time factor $\tau$. We mention that Zharkikh (1994) errs again in the presentation of the model by asserting $p_1 = p_2$ and $q_1 = q_2$, which does not hold in general. Even though the parameters on the right-hand side of Equation 2.9 are neither completely independent, nor arbitrary, it is natural to relax the constant substitution rate assumption by defining a class of stochastic matrices in that form.

**Definition 2.11.** *Let $\mathcal{M}_{\text{GIN}}$ be the class of stochastic matrices defined as follows. A matrix $\mathbf{M}$ belongs to $\mathcal{M}_{\text{GIN}}$ if and only if there exist*

$$p_1, p_2, q_1, q_2, r_1, r_2, r_3, r_4 \geq 0$$

*such that*

$$\mathbf{M} = \begin{bmatrix} 1-p_1-p_2-r_1 & p_1 & r_1 & p_2 \\ q_1 & 1-q_1-q_2-r_2 & q_2 & r_2 \\ r_3 & p_1 & 1-p_1-p_2-r_3 & p_2 \\ q_1 & r_4 & q_2 & 1-q_1-q_2-r_4 \end{bmatrix}.$$

**Lemma 2.16.** *The class $\mathcal{M}_{\text{GIN}}$ is closed under matrix multiplication.*

Proof. By taking two arbitrary matrices $\mathbf{M}, \mathbf{M}' \in \mathcal{M}_{\text{GIN}}$, one can verify that for their product $\mathbf{M}'' = \mathbf{M}\mathbf{M}'$,

$$\mathbf{M}''[1,2] = \mathbf{M}''[3,2], \qquad \mathbf{M}''[1,4] = \mathbf{M}''[3,4],$$
$$\mathbf{M}''[2,1] = \mathbf{M}''[4,1], \qquad \mathbf{M}''[2,3] = \mathbf{M}''[4,3]$$

do hold. Consequently $\mathbf{M}'' \in \mathcal{M}_{\mathrm{GIN}}$. ■

### 2.5.6 Reconstructible mutation matrices

Consider the mutation matrix

$$\mathbf{M} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

which may be a mutation matrix in the Jukes-Cantor model for a binary alphabet. If $uv$ and $vw$ are two edges in $\Psi(\mathcal{P})$ for an evolutionary tree $\mathcal{P}$ and $\mathbf{M}_{uv} = \mathbf{M}_{vw} = \mathbf{M}$, then by Corollary 2.6,

$$\mathbf{M}_{uw} = \mathbf{M}_{uv}\mathbf{M}_{vw} = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix}.$$

On the other hand, the same mutation matrix arises also if

$$\mathbf{M}_{uv} = \mathbf{M}_{vw} = \mathbf{M}' = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

This trivial example illustrates that edge mutation matrices are not necessarily determined by a subset of the mutation matrices. Additional assumptions based on, say, biological assumptions may restrict the number of solutions. In the case of $\mathbf{M}'$ the probability of mutations is twice as large as that of unchanging characters, which is unlikely in most contexts. A mild condition suggested by Chang (1996) on a subclass of the i. i. d. Markov model is offered by the next definition.

**Definition 2.12.** (Chang 1996) *A class* $\mathcal{M}$ *of stochastic matrices is* reconstructible from rows *if for each* $\mathbf{M} \in \mathcal{M}$*, and permutation matrix* $\mathbf{\Pi}$ *different from the identity matrix,* $\mathbf{\Pi}\mathbf{M} \notin \mathcal{M}$.

Definition 2.12 is useful for setting a condition on the uniqueness of the edge mutation matrices in evolutionary tree reconstruction (see Theorem 5.1). In our example the matrices $\mathbf{M}$ and $\mathbf{M}'$ cannot both be members of a class $\mathcal{M}$ if $\mathcal{M}$ is reconstructible from rows. Chang (1996) offers the example of the matrix class $\mathcal{M}_{\mathrm{DLC}}$, with DLC standing for "diagonal largest in column." For each $\mathbf{M} \in \mathcal{M}_{\mathrm{DLC}}$, $\mathbf{M}[j,j] > \mathbf{M}[i,j]$ for all $i \neq j$; $\mathcal{M}_{\mathrm{DLC}}$ is thus reconstructible from rows. Similarly, we define the class $\mathcal{M}_{\mathrm{DLR}}$ of matrices with the "diagonal largest in row" property, where for each $\mathbf{M} \in \mathcal{M}_{\mathrm{DLR}}$, $\mathbf{M}[i,i] > \mathbf{M}[i,j]$ for all $i \neq j$. Obviously, $\mathcal{M}_{\mathrm{DLR}}$ is also reconstructible from rows. In the case of most biomolecular applications, mutations occur typically with small probabilities. Consequently, in each row $i$ of an edge mutation matrix $\mathbf{M}$,

$$\sum_{j \neq i} \mathbf{M}[i,j] < \mathbf{M}[i,i].$$

All such matrices belong to $\mathcal{M}_{\mathrm{DLR}}$. For instance, edge mutation matrices in the Jukes-Cantor model with mutation probabilities less than $(1 - 1/m)$ also belong to $\mathcal{M}_{\mathrm{DLR}}$.

# Chapter 3

# Similarity and evolutionary distance

## 3.1 Introduction

A number of evolutionary tree reconstruction algorithms are based on a notion of *evolutionary distance* between nodes of the phylogeny. Evolutionary distance measures the dissimilarity between two random taxon sequences. When analyzing the evolutionary relationships between a set of species, for example, distance between two species is often directly related to the time passed since their evolution took a different course. Distance is a standard term in the literature of evolutionary trees, but the notion of *similarity* varies with different authors (see, for example, Swofford *et al.* 1996). In fact, similarity rarely enters the discussion in most studies, and if so, the term is cursorily used as an intermediate tool in analyzing various features of evolutionary distances. We choose a different path, by defining similarity first, and then defining distance as a function of similarity. The main justification for this choice is that the particular notion of similarity we use is more general than that of distance. For example, many applied studies encounter the undesirable situation where the estimated distance is infinite or negative between two taxon sequences. Furthermore, estimation problems are more manifest when using similarities, simplifying our discussion a great deal. The next series of definitions (3.1, 3.2, 3.3) defines similarity and distance as functions of sequence distributions over $\mathcal{S} \times \mathcal{S}$. They are applied to joint distributions of taxon sequence pairs.

**Definition 3.1.** *Let $S$ be a function that maps distributions over $\mathcal{S} \times \mathcal{S}$ onto the closed interval $[-1, 1]$. Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny, and for each node pair $\langle u, v\rangle \in V^2$, let $\mathbb{P}_{uv}$ denote the joint probability distribution of their random taxon sequences $\langle X^{(u)}, X^{(v)}\rangle$. The function $S$ is a* similarity metric *over $\mathcal{P}$ if and only if it has the following three properties.*

*(S) The function $S$ is symmetric; i.e., for all nodes $u, v \in V$,*

$$S(\mathbb{P}_{uv}) = S(\mathbb{P}_{vu}). \tag{3.1a}$$

*(M) The function $S$ is multiplicative; i.e., for any three nodes $u$, $v$, and $w$ where $v$ lies on the path from $u$ to $w$ in $\Psi(\mathcal{P})$,*

$$S(\mathbb{P}_{uw}) = S(\mathbb{P}_{uv})S(\mathbb{P}_{vw}). \tag{3.1b}$$

*(I) For two arbitrary nodes $u$ and $v$ where $\mathbb{P}\left\{X^{(u)} = X^{(v)}\right\} = 1$,*

$$S(\mathbb{P}_{uv}) = 1. \tag{3.1c}$$

*In particular, $S(\mathbb{P}_{uu}) = 1$.*

**Definition 3.2.** *Let $S$ be a function that maps distributions over $\mathcal{S} \times \mathcal{S}$ onto $[-1, 1]$. Define the function $D$ mapping distributions over $\mathcal{S}\times\mathcal{S}$ onto $[0, \infty]$ as*

$$D(\mathbb{P}) = \begin{cases} -\ln|S(\mathbb{P})| & \text{if } S(\mathbb{P}) \neq 0, \\ \infty & \text{if } S(\mathbb{P}) = 0. \end{cases}$$

*Let $\mathcal{P}$ be an arbitrary evolutionary tree. The function $D$ is a* distance metric *(corresponding to $S$) over $\mathcal{P}$ if and only if $S$ is a similarity metric over $\mathcal{P}$.*

**Definition 3.3.** *Let $\mathcal{C}$ be a set of phylogenies. If $S$ is a similarity metric over every $\mathcal{P} \in \mathcal{C}$, then $S$ is a similarity metric over $\mathcal{C}$, and the corresponding function $D$ defined by Definition 3.2 is a distance metric over $\mathcal{C}$.*

In what follows we use the shorthand notations

$$S(u, v) = S(\mathbb{P}_{u,v}), \text{and}$$
$$D(u, v) = D(\mathbb{P}_{uv}),$$

if the evolutionary tree to which the nodes belong is understood from the context.

**Fact 3.1.** *Let $D$ be a distance metric corresponding to similarity metric $S$ over a phylogeny $\mathcal{P} = (V, E, \mathbb{P})$. By the properties of $S$ listed in Definition 3.1, $D$ has the following properties.*

*(S) $D$ is symmetric; i.e., for all nodes $u, v \in V$,*

$$D(u, v) = D(v, u). \tag{3.2a}$$

*(A) $D$ is additive; i.e., for any three nodes $u, v, w \in V$ where $v$ lies on the path from $u$ to $w$ in $\Psi(\mathcal{P})$,*

$$D(u, w) = D(u, v) + D(v, w). \tag{3.2b}$$

*(In case of infinite distances, the conventional arithmetic extensions are applied: $\infty + x = x + \infty = \infty + \infty = \infty$.)*

*(O) For arbitrary nodes $u, v \in V$ where $\mathbb{P}\Big\{X^{(u)} = X^{(v)}\Big\} = 1$,*

$$D(u, v) = 0. \tag{3.2c}$$

*In particular, $D(u, u) = 0$.*

In this chapter we show several examples of similarity metrics over classes of phylogenies. In order to show that a particular function $S$ is a similarity metric, we have to show that $S$ satisfies Properties (S), (M), and (I) of Definition 3.1. Analyzing whether Properties (S) and (I) hold is usually straightforward. The challenge arises from proving that Property (M) holds for $S$. The next lemma eases this task.

**Lemma 3.2.** *Let $S$ be a function that maps distributions over $\mathcal{S} \times \mathcal{S}$ onto the interval $[-1, 1]$. Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny. Assume that the following hold for $S$ and $\mathcal{P}$.*

*(M*) For three arbitrary nodes $u, v, w \in V$,*

*(M*.1) if $u \prec v \prec w$, then*

$$S(u, w) = S(u, v)S(v, w); \tag{3.3a}$$

*(M*.2) if $v$ is the lowest common ancestor of $u$ and $w$, then*

$$S(u, w) = S(u, v)S(v, w). \tag{3.3b}$$

*If, in addition, Properties (S) and (I) of Definition 3.1 hold, then Property (M) also holds, and therefore $S$ is a similarity metric over $\mathcal{P}$.*

PROOF. Let $u, v, w$ be three nodes such that $v$ lies on the path between $u$ and $w$ in $\Psi(\mathcal{P})$. We prove that Equation (3.1b) holds in four cases.

*Case I:* $u \prec v \prec w$.

*Case II:* $v$ is the lowest common ancestor of $u$ and $w$.

*Case III:* $w \prec v \prec u$.

*Case IV:* None of the above.

Property (M) holds in Cases I and II by assumption of the lemma. In Case III, by Properties (S) and (M*.1),

$$S(u,v)S(v,w) = S(v,u)S(w,v) = S(w,u) = S(u,w).$$

In Case IV, we assume without loss of generality that $v \prec u$. Let $v'$ be the lowest common ancestor of $v$ and $w$. By previous cases, and Property (M*),

$$\begin{aligned} S(u,v)S(v,w) &= S(u,v)\Big(S(v,v')S(v',w)\Big) \\ &= \Big(S(u,v)S(v,v')\Big)S(v',w) \\ &= S(u,v')S(v',w) \\ &= S(u,w). \end{aligned}$$

Consequently, Property (M) holds for $S$ and $\mathcal{P}$ in all four cases, and thus $S$ is a similarity metric over $\mathcal{P}$. ■

Definitions 3.1, 3.2, and 3.3 permit many similarity and distance metrics to apply to the same evolutionary tree or class of evolutionary trees. As a trivial result, similarities and distances can be arbitrarily scaled while preserving the properties prescribed by the definitions.

**Fact 3.3.** *If $S$ is a similarity metric on an evolutionary tree $\mathcal{P}$, and the corresponding distance metric is $D$, then*

1. *for every positive integer $k$, $S^k$ is also a similarity metric on $\mathcal{P}$, with corresponding distance metric $k \cdot D$;*

2. *for every positive number $c$, $|S|^c$ is also a similarity metric on $\mathcal{P}$, with corresponding distance metric $c \cdot D$.*

## 3.2 Distance metrics

### 3.2.1 Jukes-Cantor distance

Recall that $\mathcal{C}_{\mathrm{JC}}$ is the class of phylogenies in which every edge mutation matrix belongs to $\mathcal{M}_{\mathrm{JC}}$ (see Definition 2.7). Corollary 2.10 suggests that

$$S(u,v) = \mathbb{P}\Big\{\xi^{(u)} = \xi^{(v)}\Big\} - \frac{1}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}$$

is a good candidate for a similarity metric over $\mathcal{C}_{\mathrm{JC}}$. Properties (S) and (I) required by Definition 3.1 are obviously satisfied. Corollary 2.10 proves that Property (M*.1) of Lemma 3.2 is satisfied. Theorem 3.4 below proves that Property (M) is satisfied in general for three nodes on a path.

**Definition 3.4.** *Define the functions $S_{\mathrm{JC}}$ and $D_{\mathrm{JC}}$ on distributions over $\mathcal{S}\times\mathcal{S}$ as follows. Let $\mathbb{P}$ be an arbitrary distribution over $\mathcal{S}\times\mathcal{S}$, let the random sequence pair $\langle X, X'\rangle$ be distributed according to $\mathbb{P}$ and let $X_1, X'_1$ denote the first characters of $X, X'$, respectively.*

$$\begin{aligned} S_{\mathrm{JC}}(\mathbb{P}) &= \mathbb{P}\Big\{X_1 = X'_1\Big\} - \frac{1}{m-1}\mathbb{P}\Big\{X_1 \neq X'_1\Big\}; \\ D_{\mathrm{JC}}(\mathbb{P}) &= \begin{cases} -\ln\Big|S_{\mathrm{JC}}(\mathbb{P})\Big| & \text{if } S_{\mathrm{JC}}(\mathbb{P}) \neq 0; \\ \infty & \text{if } S_{\mathrm{JC}}(\mathbb{P}) = 0. \end{cases} \end{aligned} \tag{3.4a}$$

*If, in particular, $\mathcal{P} \in \mathcal{C}_{\mathrm{JC}}$ is an arbitrary phylogeny in the Jukes-Cantor model, and $\mathbb{P}$ is the joint distribution of the random taxon sequences associated with two arbitrary nodes $u, v$ in $\mathcal{P}$, then Equation* (3.4a) *can be written equivalently as*

$$\begin{aligned} S_{\mathrm{JC}}(u,v) &= \mathbb{P}\Big\{\xi^{(u)} = \xi^{(v)}\Big\} - \frac{1}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}; \\ D_{\mathrm{JC}}(u,v) &= \begin{cases} -\ln\Big|S_{\mathrm{JC}}(u,v)\Big| & \text{if } S_{\mathrm{JC}}(u,v) \neq 0; \\ \infty & \text{if } S_{\mathrm{JC}}(u,v) = 0. \end{cases} \end{aligned} \tag{3.4b}$$

*The function* $D_{\mathrm{JC}}$ *is referred to as* Jukes-Cantor distance.

**Theorem 3.4.** *The function* $S_{\mathrm{JC}}$ *is a similarity metric over* $\mathcal{C}_{JC}$.

PROOF. Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the Jukes-Cantor model. We verify that Properties (S), (I), and (M) of Definition 3.1 hold for $S_{\mathrm{JC}}$ and $\mathcal{P}$. By Equation (3.4b) in the definition of $S_{\mathrm{JC}}$, for every node pair $u, v \in V$,

$$S_{\mathrm{JC}}(u, v) = S_{\mathrm{JC}}(v, u);$$

thus Property (S) is satisfied. If $\mathbb{P}\Big\{X^{(u)} = X^{(v)}\Big\} = 1$, then $\mathbb{P}\Big\{\xi^{(u)} = \xi^{(v)}\Big\} = 1$, and thus Property (I) also holds.

We take advantage of Lemma 3.2 to prove that Property (M) holds. The function $S_{\mathrm{JC}}$ satisfies Property (M*.1) by Corollary 2.10. We prove that Property (M*.2) is also satisfied. Let $u, v, w \in V$ be three nodes such that $v$ is the lowest common ancestor of $u$ and $w$. Let

$$\begin{aligned} p_{uv} &= \mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}; \\ p_{vw} &= \mathbb{P}\Big\{\xi^{(v)} \neq \xi^{(w)}\Big\}. \end{aligned}$$

By Lemma 2.8, for every three symbols $i, j, k \in \mathcal{A}$,

$$\begin{aligned} \mathbb{P}\Big\{\xi^{(u)} = i \,\Big|\, \xi^{(v)} = k\Big\} &= \begin{cases} \frac{p_{uv}}{m-1} & \text{if } i \neq k, \\ 1 - p_{uv} & \text{if } i = k; \end{cases} \\ \mathbb{P}\Big\{\xi^{(w)} = j \,\Big|\, \xi^{(v)} = k\Big\} &= \begin{cases} \frac{p_{vw}}{m-1} & \text{if } j \neq k, \\ 1 - p_{vw} & \text{if } j = k. \end{cases} \end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(w)}\Big\} &= \sum_{k\in\mathcal{A}} \mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(w)} \;\Big|\; \xi^{(v)} = k\Big\}\mathbb{P}\Big\{\xi^{(v)} = k\Big\} \\
&= \sum_{k\in\mathcal{A}} \mathbb{P}\Big\{\xi^{(v)} = k\Big\} \sum_{\substack{i,j\in\mathcal{A} \\ i\neq j}} \mathbb{P}\Big\{\xi^{(u)} = i, \xi^{(w)} = j \;\Big|\; \xi^{(v)} = k\Big\} \\
&= \sum_{k\in\mathcal{A}} \mathbb{P}\Big\{\xi^{(v)} = k\Big\}\bigg((m-1)(1-p_{uv})\frac{p_{vw}}{m-1} \\
&\qquad\qquad + (m-1)\frac{p_{uv}}{m-1}(1-p_{vw}) \\
&\qquad\qquad + (m-1)(m-2)\frac{p_{uv}}{m-1}\frac{p_{vw}}{m-1}\bigg) \\
&= p_{uv} + p_{vw} - \frac{m}{m-1}p_{uv}p_{vw}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_{\mathrm{JC}}(u,w) &= \mathbb{P}\Big\{\xi^{(u)} = \xi^{(v)}\Big\} - \frac{1}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(w)}\Big\} \\
&= 1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(w)}\Big\} \\
&= \left(1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}\right)\left(1 - \frac{m}{m-1}\mathbb{P}\Big\{\xi^{(v)} \neq \xi^{(w)}\Big\}\right) \\
&= S_{\mathrm{JC}}(u,v)S_{\mathrm{JC}}(v,w).
\end{aligned}$$

Property (M*) of Lemma 3.2 is thus satisfied by $S_{\mathrm{JC}}$ and $\mathcal{P}$. By Lemma 3.2, Property (M) is also satisfied. Therefore, $S_{\mathrm{JC}}$ is a similarity metric over $\mathcal{C}_{\mathrm{JC}}$. ∎

We pointed out in Fact 3.3 that distances and similarities can be scaled in a practically arbitrary manner. As an example, Jukes and Cantor (1969) originally defined their evolutionary distance for the DNA alphabet as

$$D^*(u,v) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\}\right),$$

whenever $\mathbb{P}\Big\{\xi^{(u)} \neq \xi^{(v)}\Big\} < 3/4$. The reason for such scaling can be understood in the context of constant substitution rates. Let $\alpha > 0$, and

$$\mathbf{Q} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}.$$

By Equation (2.7),

$$\mathbf{M}_\tau = e^{\mathbf{Q}\tau} = \begin{bmatrix} 1-p & \frac{p}{3} & \frac{p}{3} & \frac{p}{3} \\ \frac{p}{3} & 1-p & \frac{p}{3} & \frac{p}{3} \\ \frac{p}{3} & \frac{p}{3} & 1-p & \frac{p}{3} \\ \frac{p}{3} & \frac{p}{3} & \frac{p}{3} & 1-p \end{bmatrix}$$

with

$$p = \frac{3}{4}\left(1 - e^{-4\alpha\tau}\right).$$

If two nodes $u, v$ are separated by time $\tau$, then

$$D^*(u, v) = 3\alpha\tau,$$

which equals the expected number of substitutions over time $\tau$ for a Markov process with instantaneous transition matrix $\mathbf{Q}$.

A value of particular interest to evolutionary biologists is the *time of divergence* between species (see Figure 3.1). Let the leaves $u$ and $v$ represent two species in a phylogeny. It is assumed that their associated sequences evolve independently with the same mutation rate $\mathbf{Q}$ over the same time $\tau_{uv}$ from the sequence associated with the lowest common ancestor of the nodes. The value $\tau_{uv}$ gives the time of divergence for the two species. From the

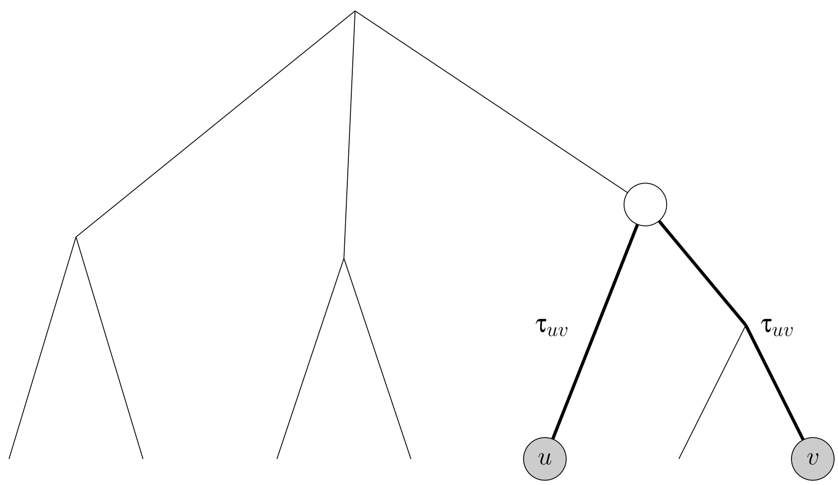



FIGURE 3.1: *Time of divergence $\tau_{uv}$ between two species $u$ and $v$. The associated sequences $X^{(u)}$ and $X^{(v)}$ evolve independently for the same time $\tau_{uv}$, with the same constant substitution rate. Consequently, $u$ and $v$ are separated by $(2\tau_{uv})$ time.*

above calculations $\tau_{uv}$ can be obtained as

$$\begin{aligned}\tau_{uv} =& \tau/2 = -\frac{1}{8\alpha}\ln\left(1-\frac{4}{3}\mathbb{P}\left\{\xi^{(u)} \neq \xi^{(v)}\right\}\right)\\ =&\frac{1}{6\alpha}D^*(u,v)\\ =&\frac{1}{8\alpha}D_{\mathrm{JC}}(u,v).\end{aligned}$$

One can measure the substitution rate $\alpha$, for example, in laboratory conditions The measured substitution rate can be used to estimate the time of divergence. Since it is necessary in any case to apply additional scaling factors to the distance, we see little theoretical advantage in using $D^*$ instead of $D_{\mathrm{JC}}$ in this study.

### 3.2.2 Kimura's distance

Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in Kimura's three parameter model. Each edge mutation matrix $\mathbf{M}_e$ with $e = uv \in E$ can be written as

$$\mathbf{M}_e = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix}.$$

As a specific consequence,

$$\begin{aligned} &\mathbb{P}\Big\{\xi^{(v)} \in \{\mathtt{A},\mathtt{G}\} \;\Big|\; \xi^{(u)} \in \{\mathtt{T},\mathtt{C}\}\Big\} = \mathbb{P}\Big\{\xi^{(v)} \in \{\mathtt{T},\mathtt{C}\} \;\Big|\; \xi^{(u)} \in \{\mathtt{A},\mathtt{G}\}\Big\} \\ &= q + r. \end{aligned} \tag{3.5}$$

Equation (3.5) implies that if we use a purine-pyrimidine encoding (`A` or `G` vs. `T` or `C`), then the encoded sequences generated by $\mathcal{P}$ have the same distribution as if they had been generated by an evolutionary tree in the Jukes-Cantor model with a binary alphabet! Consequently, the function

$$\begin{aligned} S(\mathbb{P}_{uv}) = 1 - 2\bigg(&\mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{A},\mathtt{G}\}, \xi^{(v)} \in \{\mathtt{T},\mathtt{C}\}\Big\} \\ &+ \mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{T},\mathtt{C}\}, \xi^{(v)} \in \{\mathtt{A},\mathtt{G}\}\Big\}\bigg) \end{aligned}$$

is a similarity metric over $\mathcal{P}$. Applying the same reasoning to the other two groupings of the nucleotides, we obtain Theorem 3.5.

**Definition 3.5.** *Define the functions $S_1$, $S_2$, $S_3$, $S_{\mathrm{K3}}$, and $D_{\mathrm{K3}}$ on distributions over $\mathcal{S} \times \mathcal{S}$ with $\mathcal{S} \subseteq \{\mathtt{A},\mathtt{G},\mathtt{T},\mathtt{C}\}^+$ as follows. Let $\mathbb{P}$ be an arbitrary distribution over $\mathcal{S} \times \mathcal{S}$, let the random sequence pair $\langle X, X' \rangle$ be distributed according to $\mathbb{P}$, and let $X_1, X'_1$ denote the first characters of $X, X'$, respec-*

*tively.*

$$
\begin{aligned}
S_1(\mathbb{P}) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{X_1 \in \{\mathtt{A}, \mathtt{G}\}, X_1' \in \{\mathtt{T}, \mathtt{C}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{X_1 \in \{\mathtt{T}, \mathtt{C}\}, X_1' \in \{\mathtt{A}, \mathtt{G}\}\Big\}\Bigg); \\
S_2(\mathbb{P}) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{X_1 \in \{\mathtt{A}, \mathtt{T}\}, X_1' \in \{\mathtt{G}, \mathtt{C}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{X_1 \in \{\mathtt{G}, \mathtt{C}\}, X_1' \in \{\mathtt{A}, \mathtt{T}\}\Big\}\Bigg); \\
S_3(\mathbb{P}) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{X_1 \in \{\mathtt{A}, \mathtt{C}\}, X_1' \in \{\mathtt{T}, \mathtt{G}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{X_1 \in \{\mathtt{T}, \mathtt{G}\}, X_1' \in \{\mathtt{A}, \mathtt{C}\}\Big\}\Bigg); \\
S_{\mathrm{K3}}(\mathbb{P}) = {} & S_1(\mathbb{P}) S_2(\mathbb{P}) S_3(\mathbb{P}); \\
D_{\mathrm{K3}}(\mathbb{P}) = {} & \begin{cases} -\ln\Big|S_{\mathrm{K3}}(\mathbb{P})\Big| & \textit{if } S_{\mathrm{K3}}(\mathbb{P}) \neq 0, \\ \infty & \textit{if } S_{\mathrm{K3}}(\mathbb{P}) = 0. \end{cases}
\end{aligned}
\tag{3.6a}
$$

*If, in particular,* $\mathcal{P} \in \mathcal{C}_{\mathrm{K3P}}$ *is an arbitrary phylogeny in Kimura's three parameter model, and* $\mathbb{P}$ *is the joint distribution of the random taxon labels associated with two arbitrary nodes* $u, v$ *in* $\mathcal{P}$*, then Equation* (3.6a) *can be* *written equivalently as*

$$
\begin{aligned}
S_1(u,v) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{A},\mathtt{G}\}, \xi^{(v)} \in \{\mathtt{T},\mathtt{C}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{T},\mathtt{C}\}, \xi^{(v)} \in \{\mathtt{A},\mathtt{G}\}\Big\}\Bigg); \\
S_2(u,v) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{A},\mathtt{T}\}, \xi^{(v)} \in \{\mathtt{G},\mathtt{C}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{G},\mathtt{C}\}, \xi^{(v)} \in \{\mathtt{A},\mathtt{T}\}\Big\}\Bigg); \\
S_3(u,v) = {} & 1 - 2\Bigg(\mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{A},\mathtt{C}\}, \xi^{(v)} \in \{\mathtt{T},\mathtt{G}\}\Big\} \\
& \qquad + \mathbb{P}\Big\{\xi^{(u)} \in \{\mathtt{T},\mathtt{G}\}, \xi^{(v)} \in \{\mathtt{A},\mathtt{C}\}\Big\}\Bigg); \\
S_{\mathrm{K3}}(u,v) = {} & S_1(u,v)S_2(u,v)S_3(u,v); \\
D_{\mathrm{K3}}(u,v) = {} & \begin{cases} -\ln\Big|S_{\mathrm{K3}}(u,v)\Big| & \text{if } S_{\mathrm{K3}}(u,v) \neq 0, \\ \infty & \text{if } S_{\mathrm{K3}}(u,v) = 0. \end{cases}
\end{aligned}
\tag{3.6b}
$$

*The function* $D_{\mathrm{K3}}$ *is referred to as* Kimura's three parameter distance.

**Theorem 3.5.** *All four functions* $S_1$*,* $S_2$*,* $S_3$*, and* $S_{\mathrm{K3}}$ *are similarity metrics over* $\mathcal{C}_{\mathrm{K3}}$*.*

Proof. We use a coupling argument and Theorem 3.4. Let $\mathcal{P} = (V, E, \mathbb{P})$ be an arbitrary phylogeny in $\mathcal{C}_{\mathrm{K3P}}$ and define the evolutionary tree $\mathcal{P}_1 = (V, E, \mathbb{P}_1)$, encoding sequences generated by $\mathcal{P}$ as follows. Denote the random taxon sequences generated by $\mathcal{P}$ as $\langle X^{(u)} \colon u \in V\rangle$, and those generated by $\mathcal{P}_1$ as $\langle Y^{(u)} \colon u \in V\rangle$. The random taxon sequence generation by $\mathcal{P}_1$ is coupled with that of $\mathcal{P}$ so that

$$
Y_i^{(u)} = \begin{cases} \mathtt{U} & \text{if } X_i^{(u)} \in \{\mathtt{A},\mathtt{G}\}; \\ \mathtt{Y} & \text{if } X_i^{(u)} \in \{\mathtt{T},\mathtt{C}\}. \end{cases}
$$

Subsequently, $\mathcal{P}_1$ is in the Jukes-Cantor model with the alphabet $\{\mathtt{U},\mathtt{Y}\}$. Applying $S_1$ to $\mathcal{P}$ gives the same result as applying $S_{JC}$ to $\mathcal{P}_1$. Consequently, $S_1$ is a similarity metric over $\mathcal{P}$ by Theorem 3.4. We can prove analogously that $S_2$ and $S_3$ are similarity metrics. Hence $S_{\mathrm{K3}} = S_1 S_2 S_3$ is also a similarity metric over $\mathcal{P}$. ■

### 3.2.3 Paralinear distance

Definitions 3.4 and 3.5 showed examples of similarity metrics over restricted classes of evolutionary trees. In general, however, neither of the functions $S_{\mathrm{JC}}$ and $S_{\mathrm{K3}}$ is a similarity metric over an arbitrary phylogeny in the i. i. d. Markov model. Corollary 2.6 suggests a way to arrive at such a general similarity metric. By Corollary 2.6,

$$f(u,v) = \det \mathbf{M}_{uv}$$

satisfies Property (M) of Definition 3.1. In addition, the function $f$ also satisfies Property (I) since if $\mathbb{P}\Big\{X^{(u)} = X^{(v)}\Big\} = 1$, then $\mathbf{M}_{uv}$ is the identity matrix. Unfortunately, Property (S) is not satisfied in general, because $\det \mathbf{M}_{uv} = \det \mathbf{M}_{vu}$ does not always hold in the i. i. d. Markov model. A convenient solution to this problem is to use $f(u,v)f(v,u)$ or $\sqrt{f(u,v)f(v,u)}$ as a similarity metric.

**Definition 3.6.** *Define the functions $S_{\mathrm{L}}$ and $D_{\mathrm{L}}$ on distributions over $\mathcal{S}\times\mathcal{S}$ as follows. Let $\mathbb{P}$ be an arbitrary distribution over $\mathcal{S}\times\mathcal{S}$, let the random sequence pair $\langle X, X'\rangle$ be distributed according to $\mathbb{P}$, and let $X_1, X'_1$ denote the first characters of $X, X'$, respectively. Define the $m \times m$ transition matrices $\mathbf{M}_{\mathbb{P}}$, $\mathbf{M}'_{\mathbb{P}}$ by their entries as*

$$\mathbf{M}_{\mathbb{P}}[i,j] = \mathbb{P}\Big\{X'_1 = j \;\Big|\; X_1 = i\Big\};$$
$$\mathbf{M}'_{\mathbb{P}}[i,j] = \mathbb{P}\Big\{X_1 = j \;\Big|\; X'_1 = i\Big\}.$$

*Then*

$$\begin{aligned} S_{\mathrm{L}}(\mathbb{P}) &= \sqrt{\Big(\det \mathbf{M}_{\mathbb{P}}\Big)\Big(\det \mathbf{M}'_{\mathbb{P}}\Big)}, \textit{ and} \\ D_{\mathrm{L}}(\mathbb{P}) &= \begin{cases} -\ln S_{\mathrm{L}}(\mathbb{P}) & \textit{if } S_{\mathrm{L}}(\mathbb{P}) \neq 0 \\ \infty & \textit{if } S_{\mathrm{L}}(\mathbb{P}) = 0. \end{cases} \end{aligned} \tag{3.7a}$$

*If, in particular, $\mathcal{P}$ is an arbitrary phylogeny in the i. i. d. Markov model, and $\mathbb{P}$ is the joint distribution of the random taxon labels associated with two arbitrary nodes $u, v$ in $\mathcal{P}$, then Equation* (3.7a) *can be written equivalently as*

$$\begin{aligned} S_{\mathrm{L}}(u,v) &= \sqrt{\Big(\det \mathbf{M}_{uv}\Big)\Big(\det \mathbf{M}_{vu}\Big)}; \\ D_{\mathrm{L}}(u,v) &= \begin{cases} -\ln S_{\mathrm{L}}(u,v) & \text{if } S_{\mathrm{L}}(u,v) \neq 0, \\ \infty & \text{if } S_{\mathrm{L}}(u,v) = 0. \end{cases} \end{aligned} \tag{3.7b}$$

*The function $D_{\mathrm{L}}$ is referred to as* paralinear distance.

**Theorem 3.6.** (Chang and Hartigan 1991, Lake 1994, and Chang 1996.) *$S_{\mathrm{L}}$ is a similarity metric in the i. i. d. Markov model.*

Proof. Properties (S) and (I) of Definition 3.1 are trivially satisfied. By Corollary 2.6, if the nodes $u, v, w$ of a phylogeny $\mathcal{P}$ in the i. i. d. Markov model lie on a path in $\Psi(\mathcal{P})$, then $\mathbf{M}_{uw} = \mathbf{M}_{uv}\mathbf{M}_{vw}$ and $\mathbf{M}_{wu} = \mathbf{M}_{wv}\mathbf{M}_{vu}$. Thus

$$\begin{aligned} S_{\mathrm{L}}(u,w) &= \sqrt{\Big(\det \mathbf{M}_{uw}\Big)\Big(\det \mathbf{M}_{wu}\Big)} \\ &= \sqrt{\Big(\det \mathbf{M}_{uv}\Big)\Big(\det \mathbf{M}_{vw}\Big)\Big(\det \mathbf{M}_{wv}\Big)\Big(\det \mathbf{M}_{vu}\Big)} \\ &= \left(\sqrt{\Big(\det \mathbf{M}_{uv}\Big)\Big(\det \mathbf{M}_{vu}\Big)}\right)\left(\sqrt{\Big(\det \mathbf{M}_{vw}\Big)\Big(\det \mathbf{M}_{wv}\Big)}\right) \\ &= S_{\mathrm{L}}(u,v)S_{\mathrm{L}}(v,w). \quad \blacksquare \end{aligned}$$

An alternative definition of $D_{\mathrm{L}}$ is offered by the next lemma.

**Lemma 3.7.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be an evolutionary tree in the i. i. d. Markov model. For every node pair $u, v \in V$, define the $m \times m$* joint probability matrix *by its entries as*

$$\mathbf{J}_{uv}[i,j] = \mathbb{P}\Big\{\xi^{(u)} = i, \xi^{(v)} = j\Big\}.$$

*We claim the following.*

$$\det \mathbf{J}_{uv} = \det \mathbf{J}_{vu}; \tag{3.8a}$$

$$\Big(\det \mathbf{M}_{uv}\Big) \prod_{i\in\mathcal{A}} \pi_i^{(u)} = \det \mathbf{J}_{uv}; \tag{3.8b}$$

$$\Big(\det \mathbf{M}_{uv}\Big) \prod_{i\in\mathcal{A}} \pi_i^{(u)} = \Big(\det \mathbf{M}_{vu}\Big) \prod_{i\in\mathcal{A}} \pi_i^{(v)}. \tag{3.8c}$$

PROOF. Equation (3.8a) holds because $\mathbf{J}_{vu}$ is the transpose of $\mathbf{J}_{uv}$. Since

$$\mathbb{P}\Big\{\xi^{(u)} = i, \xi^{(v)} = j\Big\} = \pi_i^{(u)} \mathbb{P}\Big\{\xi^{(v)} = j \;\Big|\; \xi^{(u)} = i\Big\},$$

the matrix $\mathbf{J}_{uv}$ can be obtained from $\mathbf{M}_{uv}$ by multiplying the rows of $\mathbf{M}_{uv}$ with $\pi_1^{(u)}, \ldots, \pi_m^{(u)}$, respectively. Consequently Equation (3.8b) holds. Switching the roles of $u$ and $v$ in Equation (3.8b),

$$\Big(\det \mathbf{M}_{vu}\Big) \prod_{i\in\mathcal{A}} \pi_i^{(v)} = \det \mathbf{J}_{vu}. \tag{$*$}$$

Equations (3.8a), (3.8b), and ($*$) together imply Equation (3.8c). ■

Lemma 3.7 also shows that

$$\Big(\det \mathbf{M}_{uv}\Big)\Big(\det \mathbf{M}_{vu}\Big) = \frac{\Big(\det \mathbf{J}_{uv}\Big)^2}{\Big(\prod_{i\in\mathcal{A}} \pi_i^{(u)}\Big)\Big(\prod_{i\in\mathcal{A}} \pi_i^{(v)}\Big)},$$

so the square roots on the right-hand sides of Equations (3.7a) and (3.7b) always exist. In fact, Lake (1994) originally defines the distance as

$$D_{\mathrm{L}}(u,v) = -\ln \frac{|\det \mathbf{J}_{uv}|}{\Big(\prod_{i\in\mathcal{A}} \pi_i^{(u)}\Big)^{1/2}\Big(\prod_{i\in\mathcal{A}} \pi_i^{(v)}\Big)^{1/2}}, \tag{3.9}$$

while Equation (3.7b) is the preferred formula of Chang and Hartigan (1991) and Chang (1996). Yet another equivalent definition is

$$D_{\mathrm{L}}(u,v) = -\ln|\det \mathbf{M}_{uv}| + \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u)} - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(v)}.$$

**Corollary 3.8.** *If $\mathcal{P}$ is a phylogeny in the time-reversible model, then for every node pair $u, v$,*

$$D_{\mathrm{L}}(u, v) = -\ln |\det \mathbf{M}_{uv}|. \tag{3.10}$$

This formula was first given by Barry and Hartigan (1987), who called it "asynchronous distance", since $\det \mathbf{M}_{uv} \neq \det \mathbf{M}_{vu}$ in general.

## 3.3 Uniqueness of evolutionary distances

There may be many similarity metrics defined over the same class of phylogenies. In addition to the scaling described by Fact 3.3, substantially different similarity metrics may also exist. For example, in the Jukes-Cantor model, the functions $S_{\mathrm{L}}$, $S_{\mathrm{K3}}$, and $S_{JC}$ are all similarity metrics, yielding different values in general. However, Properties (S), (I), and especially (M) of Definition 3.1 restrict the set of possible similarity and distance metrics over a class. We show that in the class of phylogenies in the i. i. d. Markov model with the assumptions of time-reversibility and constant substitution rates, all distance metrics can be written in the same specific format.

Consider the class $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$ of all phylogenies in the i. i. d. Markov model with time-reversibility and constant substitution rate matrix $\mathbf{Q}$. Since every edge mutation matrix in a phylogeny of $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$ is a power of $e^{\mathbf{Q}}$, and the taxon label distributions are stationary, any distance metric over $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$ is a function of one free parameter, evolutionary time. Specifically, let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$, and let $u, v \in V$. Then $\mathbf{M}_{uv} = \exp(\mathbf{Q}\tau)$ for some $\tau \geq 0$, and for any distance metric $D$,

$$D(u, v) = \varphi_{\mathbf{Q}}(\tau)$$

with some function $\varphi_{\mathbf{Q}} \colon [0, \infty) \mapsto [0, \infty]$. By Property (A) of Definition 3.2,

$$\varphi_{\mathbf{Q}}(\tau) + \varphi_{\mathbf{Q}}(\tau') = \varphi_{\mathbf{Q}}(\tau + \tau').$$

Imposing continuity on $\varphi_{\mathbf{Q}}$, it is not difficult to show that it must be a linear function.

**Theorem 3.9.** *Let* $\mathbf{Q}$ *be a substitution rate matrix such that exactly one of its eigenvalues equals zero. Let* $D$ *be a distance metric over* $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$*, the subclass of the i. i. d. Markov model with time-reversibility and constant substitution rate matrix* $\mathbf{Q}$*. We claim the following.*

*(1) There exists a function* $\varphi_{\mathbf{Q}}\colon [0,\infty) \mapsto [0,\infty]$ *such that for every node pair* $u, v$ *of a phylogeny belonging to* $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$*,*

$$D(u,v) = \varphi_{\mathbf{Q}}\left(\frac{\ln\det \mathbf{M}_{uv}}{\ln\det e^{\mathbf{Q}}}\right).$$

*(2) If* $\varphi_{\mathbf{Q}}$ *is continuous at* $0$*, i.e.,* $\lim_{x\to+0}\varphi_{\mathbf{Q}}(x) = \varphi_{\mathbf{Q}}(0)$*, then for every* $x, \tau \geq 0$*,*

$$\varphi_{\mathbf{Q}}(x\tau) = x\varphi_{\mathbf{Q}}(\tau). \tag{3.11}$$

*In particular,*

$$\varphi_{\mathbf{Q}}(\tau) = \tau\varphi_{\mathbf{Q}}(1).$$

Proof. The joint distribution of $\langle \xi^{(u)}, \xi^{(v)}\rangle$ is defined by $\boldsymbol{\pi}^{(u)}$ and $\mathbf{M}_{uv}$. Since the substitution rates are constant, there exists $\tau \geq 0$ such that

$$\mathbf{M}_{uv} = e^{\mathbf{Q}\tau}.$$

The vector $\boldsymbol{\pi}^{(u)}$ is independent from $u$ by time-reversibility and the fact that $\mathbf{Q}$ has one eigenvector with eigenvalue 0. Consequently, $D(u,v)$ depends on the one free parameter $\tau$, i.e., there exists $\varphi_{\mathbf{Q}}$ such that

$$D(u,v) = \varphi_{\mathbf{Q}}(\tau) = \varphi_{\mathbf{Q}}\left(\frac{\ln\det \mathbf{M}_{uv}}{\ln\det e^{\mathbf{Q}}}\right),$$

which is Claim (1). We prove Claim (2) of the theorem in three cases.

*Case I.* Equation (3.11) is proven by induction for non-negative integer $x$. For $x = 1$, the equation holds trivially. For $x = 0$, the equation holds by Property (O) of distance metrics. Assume that $x > 1$ and

$$\varphi_{\mathbf{Q}}((x-1)\tau) = (x-1)\varphi_{\mathbf{Q}}(\tau).$$

Let $\mathcal{P} \in \mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$ be a phylogeny such that it contains three nodes $u, v, w$, where $u \prec v \prec w$, and $\mathbf{M}_{uv} = \exp(\mathbf{Q}\tau)$, $\mathbf{M}_{vw} = \exp(\mathbf{Q}(x-1)\tau)$. By Corollary 2.6, $\mathbf{M}_{uw} = \exp(\mathbf{Q}x\tau)$. Subsequently, the additivity of $D$ and the induction hypothesis imply that

$$\varphi_{\mathbf{Q}}(x\tau) = \varphi_{\mathbf{Q}}((x-1)\tau) + \varphi_{\mathbf{Q}}(\tau) = ((x-1)+1)\varphi_{\mathbf{Q}}(\tau) = x\varphi_{\mathbf{Q}}(\tau).$$

Thus $\varphi_{\mathbf{Q}}(x\tau) = x\varphi_{\mathbf{Q}}(\tau)$ for every non-negative integer $x$.

*Case II.a.* Let $x = 1/k$ with a positive integer $k$. By Case I, $\varphi_{\mathbf{Q}}(\tau) = k\varphi_{\mathbf{Q}}(\tau/k)$, and thus $x\varphi_{\mathbf{Q}}(\tau) = \varphi_{\mathbf{Q}}(x\tau)$.

*Case II.b.* Let $x = k'/k$ with positive integers $k, k'$. By Cases I and II.a,

$$\varphi_{\mathbf{Q}}(x\tau) = k'\varphi_{\mathbf{Q}}(\tau/k) = \frac{k'}{k}\varphi_{\mathbf{Q}}(\tau).$$

*Case III.* Let $x$ be a positive irrational number. Let $\{x_k\colon k = 1, 2, \ldots\}$ be an infinite decreasing series of positive rational numbers such that they converge to $x$, i.e., $\lim_{k\to\infty} x_k = x$. By previous cases, Property (A), and the continuity assumption,

$$\begin{aligned}
\varphi_{\mathbf{Q}}(0) &= \lim_{k\to\infty} \varphi_{\mathbf{Q}}(x\tau - x_k\tau) \\
&= \varphi_{\mathbf{Q}}(x\tau) - \lim_{k\to\infty} x_k\varphi_{\mathbf{Q}}(\tau) \\
&= \varphi_{\mathbf{Q}}(x\tau) - x\varphi_{\mathbf{Q}}(\tau).
\end{aligned}$$

Since $\varphi_{\mathbf{Q}}(0) = 0$, $\varphi_{\mathbf{Q}}(x\tau) = x\varphi_{\mathbf{Q}}(\tau)$. ■

Theorem 3.9 shows that although many distance metrics may be defined for the same phylogeny, they only differ by scaling factors in the class $\mathcal{C}_{\mathrm{TR}(\mathbf{Q})}$. For example, Gojobori *et al.* (1982) and Hasegawa *et al.* (1985) define distance metrics for the Gojobori-Ishii-Nei and Hasegawa-Kishino-Yano models with the assumptions of time-reversibility and constant substitution rates. The rather complicated distance metrics turn out to be identical to the paralinear distance with a scaling factor that depends on the eigenvalues of the substitution rate matrix.

## 3.4 Empirical distance and similarity

Thus far the theme of our discussion has been the distribution of random taxon sequences. With the introduction of evolutionary distance and similarity, an explicit connection has been established between distribution and topology in large classes of evolutionary trees. Given the set of all pairwise distances between nodes, reconstructing the topology is simply the problem of finding a minimal spanning tree in a graph, which can be solved very efficiently by the well-known algorithms of Kruskal (1956) and Prim (1957). Formally, let $\mathcal{P} = (V, E, \mathbb{P})$ be an evolutionary tree with a distance metric $D$. Construct the full graph $\mathcal{G}$ over $V$ with edge weights equal to the pairwise distances between the endpoints on all edges. If all the distances are positive, then the minimum length spanning tree of $\mathcal{G}$ is uniquely $\Psi(\mathcal{P})$.

In practice it is impossible to obtain all the distances between nodes. The difficulties stem already from the fact that non-leaf nodes in a phylogeny frequently represent extinct species and thus our knowledge about them is limited. What we can reasonably hope for, instead, is knowledge about the leaf nodes. Fortunately, the pairwise distances between leaves are already sufficient to recover the topology of the evolutionary tree, due to additivity of the distances. There are many algorithms that recover the topology of a phylogeny $\mathcal{P} = (V, E, \mathbb{P})$ with leaves $L \subset V$ from distances between leaves rapidly, such as the $O\left(|L|^2\right)$ algorithms of Bandelt (1990) and Gusfield (1997), or our Fit-Tree algorithm in §5.1.2. If we know the distribution of the random taxon sequences, we can carry out the exact calculation of pairwise distances ideally required by these topology reconstruction algorithms. When we have to reconstruct the topology from a sample of sequences instead, the distribution and therefore the distances are not exactly calculated but only estimated from the sample. The simplest way to estimate the distance (or similarity) of two nodes is to substitute the distribution of their taxon sequences in the definition of the distance (or similarity) with the empirical distribution calculated from their sample sequences. The estimates obtained in this way are called *empirical distances* (or *empirical similarities*). For example, let us suppose that the binary sample sequences $X^{(u)} = s_1 \cdots s_\ell$ and $X^{(v)} = t_1 \cdots t_\ell$ are observed, associated with leaves $u$ and $v$ of an evolutionary tree in the Jukes-Cantor model. The probability $\mathbb{P}\left\{\xi^{(u)} \neq \xi^{(v)}\right\}$ can be estimated as

$$\hat{p}_{uv} = \frac{1}{\ell} \sum_{k=1}^{\ell} \mathbb{I}\{s_k \neq t_k\},$$

which in turn can be plugged into Equation (3.4) as

$$\hat{D}_{\mathrm{JC}}(u, v) = -\ln\left|1 - 2\hat{p}_{uv}\right|$$

to obtain an estimate of $D_{\mathrm{JC}}(u, v)$. Proceeding similarly with each leaf pair we should obtain fairly good distance estimates if the sequences are "long enough." The pivotal question is how long the sequences have to be in order to obtain the correct topology. Alternatively, we may wonder how successful an evolutionary tree reconstruction algorithm can be if the available sample sequences are not very long. Molecular evolution studies often build trees from aligned sequences corresponding to the same gene in different species. Gene sequence lengths typically fall between a few hundred and a few thousand base pairs. The theoretically maximal amount of data available is limited by the size of the genomes. Therefore arbitrary precision in distance estimation cannot be assumed. The distance estimates usually do not correspond to any evolutionary tree in the sense that the distance metric used would not give node distances equal to the estimates on any evolutionary tree in the hypothesis class. Searching for a tree that would give distances that are "closest" to the estimates is a tempting idea, but with most sensible notions of "closest," even the approximation version constitutes an NP-hard problem by becoming a variant of the Steiner-tree problem (Day 1987; Agarwala *et al.* 1999). We choose a different route, which begins with analyzing the speed with which various empirical distances converge to the true distances. The insight into the statistical nature of the problem gained from this analysis allows us to design efficient algorithms that recover the topology with high success.

In order to recover the evolutionary relationships between taxa, an algorithm typically requires that the distances be estimated within a certain error. Specifically, the success of the algorithm is guaranteed if for every node pair $(u, v)$ there exists an error bound $\epsilon_D$, which possibly depends on the node pair, such that

$$\left|\hat{D}(u,v) - D(u,v)\right| < \epsilon_D. \tag{3.12}$$

Assuming that $\hat{D}(u,v) = -\ln \hat{S}(u,v)$, Equation (3.12) can be rewritten in terms of the estimated similarity as

$$e^{-\epsilon_D} < \frac{\hat{S}(u,v)}{S(u,v)} < e^{\epsilon_D}. \tag{3.13}$$

In the rest of the chapter we derive explicit upper bounds on large deviations of the empirical similarity. Our main tool in obtaining the bounds is Chernoff's method. Chernoff's bounding method originates from his 1952 paper and is now routinely used in many areas of computer science and discrete mathematics (Alon and Spencer 1992). We illustrate the method in the proof of the following lemma.

**Lemma 3.10.** *Let $\eta$ be a binomially distributed random variable with parameters $\ell$ and $p$; i.e., for every $k = 0, \ldots, \ell$,*

$$\mathbb{P}\Big\{\eta = k\Big\} = \binom{\ell}{k} p^k (1-p)^{\ell-k}.$$

*If $p < \frac{1}{2}$, then*

$$\mathbb{P}\left\{\eta \geq \frac{\ell}{2}\right\} \leq \Big(1 - (1-2p)^2\Big)^{\ell/2}.$$

Proof. Let $\eta_1, \eta_2, \ldots, \eta_\ell$ be independent identically distributed random variables such that

$$\mathbb{P}\Big\{\eta_i = 1\Big\} = 1 - \mathbb{P}\Big\{\eta_i = 0\Big\} = p, \qquad i = 1, \ldots, \ell.$$

Let $c > 0$ be a positive number. Since $\eta$ and $\sum_{i=1}^{\ell} \eta_i$ are identically distributed,

$$\mathbb{P}\left\{\eta \geq \frac{\ell}{2}\right\} = \mathbb{P}\left\{\sum_{i=1}^{\ell} \eta_i \geq \frac{\ell}{2}\right\} = \mathbb{P}\left\{e^{c\sum_{i=1}^{\ell}\eta_i} \geq e^{c\ell/2}\right\}. \tag{3.14}$$

By Markov's inequality (see, e.g., Rényi 1970),

$$\mathbb{P}\left\{e^{c\sum_{i=1}^{\ell}\eta_i} \geq e^{c\ell/2}\right\} \leq \frac{\mathbb{E}e^{c\sum_{i=1}^{\ell}\eta_i}}{e^{c\ell/2}}. \tag{$*$}$$

Furthermore, by the independence of the random variables $\eta_i$,

$$\mathbb{E}e^{c\sum_{i=1}^{\ell}\eta_i} = \prod_{i=1}^{\ell} \mathbb{E}e^{c\eta_i} = \Big((1-p) + pe^c\Big)^{\ell}. \tag{$**$}$$

Therefore, by Equations (3.14), $(*)$, and $(**)$,

$$\mathbb{P}\left\{\eta \geq \frac{\ell}{2}\right\} \leq e^{-c\ell/2}\Big((1-p) + pe^c\Big)^{\ell}.$$

The right-hand side is minimized by choosing

$$c = \ln \frac{1-p}{p},$$

thus

$$\mathbb{P}\Big\{\eta \geq \frac{\ell}{2}\Big\} \leq \Big(\frac{p}{1-p}\Big)^{\ell/2}\Big(2(1-p)\Big)^{\ell} = \Big(4p(1-p)\Big)^{\ell/2} = \Big(1-(1-2p)^2\Big)^{\ell/2},$$

proving the lemma. ■

Chernoff's method for bounding sums of random variables consists of the exponential transformation in Equation (3.14) with a suitable choice of $c$ in the exponent.

### 3.4.1 Jukes-Cantor distance

**Definition 3.7.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the Jukes-Cantor model. Let $u, v \in V$ be two nodes with associated taxon sequences $X^{(u)}, X^{(v)} \in \mathcal{S}$. The empirical similarity $\hat{S}_{\mathrm{JC}}$ between $u$ and $v$ is defined as*

$$\hat{S}_{\mathrm{JC}}(u,v) = \frac{1}{\ell}\sum_{i=1}^{\ell}\Big(\mathbb{I}\big\{X_i^{(u)} = X_i^{(v)}\big\} - \frac{1}{m-1}\mathbb{I}\big\{X_i^{(u)} \neq X_i^{(v)}\big\}\Big), \tag{3.15a}$$

*with*

$$\ell = \Big|X^{(u)}\Big| = \Big|X^{(v)}\Big|.$$

*The empirical distance $\hat{D}_{\mathrm{JC}}$ between the two nodes is defined as*

$$\hat{D}_{\mathrm{JC}}(u,v) = \begin{cases} -\ln\Big|\hat{S}_{\mathrm{JC}}(u,v)\Big| & \text{if } \hat{S}_{\mathrm{JC}}(u,v) \neq 0; \\ \infty & \text{if } \hat{S}_{\mathrm{JC}}(u,v) = 0. \end{cases} \tag{3.15b}$$

Sometimes it is also known that $S_{\mathrm{JC}}(u,v) > 0$ for all nodes $u \neq v$. For example, if all the edge mutation probabilities are less than $(1 - 1/m)$, then all the pairwise similarities are positive. Although with exponentially small probability (bounded from above by $\Big(1 - S_{\mathrm{JC}}^2(u,v)\Big)^{\ell/2}$ in case of $m = 2$ by Lemma 3.10), Equation (3.15a) may give $\hat{S}_{\mathrm{JC}}(u,v) \leq 0$. Negative estimated similarities are often discounted, especially by biologists. More preferable estimators can be defined in this case as, e.g.,

$$\tilde{S}_{\mathrm{JC}}(uv) = \begin{cases} \hat{S}_{\mathrm{JC}}(u,v) & \text{if } \hat{S}_{\mathrm{JC}}(u,v) > 0, \\ \kappa_\ell & \text{if } \hat{S}_{\mathrm{JC}}(u,v) \le 0; \end{cases} \tag{3.16a}$$

$$\tilde{D}_{\mathrm{JC}}(u,v) = \begin{cases} \hat{D}_{\mathrm{JC}}(u,v) & \text{if } \hat{S}_{\mathrm{JC}}(u,v) > 0, \\ K_\ell & \text{if } \hat{S}_{\mathrm{JC}}(u,v) \le 0; \end{cases} \tag{3.16b}$$

where $\kappa_\ell$ is a small constant that is less than or equal to the smallest positive value of $\hat{S}_{\mathrm{JC}}$, and $K_\ell$ is a large constant that is larger than or equal to the maximum finite value of $\hat{D}_{\mathrm{JC}}$ on sample length $\ell$. From Equation (3.15a), it follows that if $\hat{S}_{\mathrm{JC}}(u,v) > 0$, then

$$\hat{S}_{\mathrm{JC}}(u,v) \ge \ell^{-1} \frac{m - (\ell \bmod m)}{m-1}.$$

Consequently, we can pick any $K_\ell \ge (\ln \ell + \ln(m-1))$, and the values

$$\kappa_\ell = \frac{1}{\ell(m-1)}, \quad \text{and} \quad K_\ell = \ln \ell + \ln m$$

are viable choices in Equation (3.16). Trivially, $\hat{S}_{\mathrm{JC}}$ is an unbiased estimator of $S_{\mathrm{JC}}$ while $\mathbb{E}\tilde{S}_{\mathrm{JC}}(u,v)$ is larger than $S_{\mathrm{JC}}(u,v)$ unless $S_{\mathrm{JC}}(u,v) = 1$.

**Lemma 3.11.** *Let $\hat{S}_{\mathrm{JC}}$ be the estimate of $S_{\mathrm{JC}}$ given by Equation* (3.15a)*. Let*

$$\gamma = \frac{m}{m-1}.$$

*For all nodes $u, v$ of a phylogeny in the Jukes-Cantor model, if $S_{\mathrm{JC}}(u,v) \ne 0$, then for every sample length $\ell$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{ \frac{\hat{S}_{\mathrm{JC}}(u,v)}{S_{\mathrm{JC}}(u,v)} \le 1 - \epsilon \right\} \le \exp\left( -\frac{2}{\gamma^2} \ell S_{\mathrm{JC}}^2(u,v) \epsilon^2 \right); \tag{3.17a}$$

$$\mathbb{P}\left\{ \frac{\hat{S}_{\mathrm{JC}}(u,v)}{S_{\mathrm{JC}}(u,v)} \ge 1 + \epsilon \right\} \le \exp\left( -\frac{2}{\gamma^2} \ell S_{\mathrm{JC}}^2(u,v) \epsilon^2 \right). \tag{3.17b}$$

The lemma follows from Hoeffding's inequality (1963), a Chernoff bound.

**Theorem 3.12.** (Hoeffding 1963)

*Let $\eta_1, \eta_2, \ldots, \eta_\ell$ be independent random variables such that for every $i$, there exists $a_i, b_i \in \mathbb{R}$ with $\mathbb{P}\{\eta_i \in [a_i, b_i]\} = 1$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\Bigg\{\sum_{i=1}^{\ell}\Big(\eta_i - \mathbb{E}\eta_i\Big) \geq \epsilon\Bigg\} \leq \exp\Bigg(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell}(b_i - a_i)^2}\Bigg); \tag{3.18a}$$

$$\mathbb{P}\Bigg\{\sum_{i=1}^{\ell}\Big(\eta_i - \mathbb{E}\eta_i\Big) \leq -\epsilon\Bigg\} \leq \exp\Bigg(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell}(b_i - a_i)^2}\Bigg). \tag{3.18b}$$

Proof of Lemma 3.11. Define the random variables $\{\eta_i \colon i = 1, \ldots \ell\}$ as

$$\eta_i = \begin{cases} -\frac{1}{m-1} & \text{if } X_i^{(u)} \neq X_i^{(v)}; \\ 1 & \text{if } X_i^{(u)} = X_i^{(v)}. \end{cases}$$

Equation (3.15a) can be rewritten as

$$\hat{S}_{\mathrm{JC}}(u, v) = \frac{1}{\ell}\sum_{i=1}^{\ell}\eta_i.$$

The random variables $\{\eta_i\}$ are independent and identically distributed, and

$$\mathbb{E}\eta_i = S_{\mathrm{JC}}(u, v).$$

Subsequently, for $S_{\mathrm{JC}}(u, v) > 0$, the probabilities on the right-hand side of Equation (3.17) can be rearranged as

$$\mathbb{P}\Bigg\{\frac{\hat{S}_{\mathrm{JC}}(u, v)}{S_{\mathrm{JC}}(u, v)} \leq 1 - \epsilon\Bigg\} = \mathbb{P}\Bigg\{\sum_{i=1}^{\ell}(\eta_i - \mathbb{E}\eta_i) \leq -\ell S_{\mathrm{JC}}(u, v)\epsilon\Bigg\};$$

$$\mathbb{P}\Bigg\{\frac{\hat{S}_{\mathrm{JC}}(u, v)}{S_{\mathrm{JC}}(u, v)} \geq 1 + \epsilon\Bigg\} = \mathbb{P}\Bigg\{\sum_{i=1}^{\ell}(\eta_i - \mathbb{E}\eta_i) \geq \ell S_{\mathrm{JC}}(u, v)\epsilon\Bigg\}.$$

The lemma follows now from Hoeffding's inequality applied to the random variables $\eta_i$, which take their values in the interval $\Big[-\frac{1}{m-1}, 1\Big]$.

If $S_{\mathrm{JC}}(u,v) < 0$, then the rearrangement results in the equations

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{JC}}(u,v)}{S_{\mathrm{JC}}(u,v)} \leq 1-\epsilon\right\} = \mathbb{P}\left\{\sum_{i=1}^{\ell}(\eta_i - \mathbb{E}\eta_i) \geq -\ell S_{\mathrm{JC}}(u,v)\epsilon\right\};$$
$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{JC}}(u,v)}{S_{\mathrm{JC}}(u,v)} \geq 1+\epsilon\right\} = \mathbb{P}\left\{\sum_{i=1}^{\ell}(\eta_i - \mathbb{E}\eta_i) \leq \ell S_{\mathrm{JC}}(u,v)\epsilon\right\}.$$

Applying Hoeffding's inequality to the right-hand sides symbolically yields the same upper bounds as before. ■

### 3.4.2 Kimura's distance

**Definition 3.8.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in Kimura's three parameter model. Let $u, v \in V$ be two nodes with associated random sequences $X^{(u)}, X^{(v)}$. Let*

$$\ell = \left|X^{(u)}\right| = \left|X^{(v)}\right|.$$

*Define the indicator variables $\{P_i, Q_i, R_i \colon i = 1, \ldots, \ell\}$ for the following events.*

$$\begin{aligned} X_i^{(u)} &= \texttt{A G T C}\quad \texttt{A G T C}\quad \texttt{A G T C} \\ X_i^{(v)} &= \underbrace{\texttt{G A C T}}_{P_i}\quad \underbrace{\texttt{T C A G}}_{Q_i}\quad \underbrace{\texttt{C T G A}}_{R_i} \end{aligned}$$

*The empirical similarity $\hat{S}_{\mathrm{K3}}$ between $u$ and $v$ is defined as*

$$\hat{S}_{\mathrm{K3}}(u,v) = f_{\mathrm{K3}}\Big(X^{(u)}, X^{(v)}\Big)$$
$$= \frac{\Big(\ell - 2\sum_{i=1}^{\ell}(P_i+Q_i)\Big)\Big(\ell - 2\sum_{i=1}^{\ell}(P_i+R_i)\Big)\Big(\ell - 2\sum_{i=1}^{\ell}(Q_i+R_i)\Big)}{\ell}. \tag{3.19a}$$

*The empirical distance $\hat{D}_{\mathrm{K3}}$ between the two nodes is defined as*

$$\hat{D}_{\mathrm{K3}}(u,v) = \begin{cases} -\ln\left|\hat{S}_{\mathrm{K3}}(u,v)\right| & \textit{if } \hat{S}_{\mathrm{K3}}(u,v) \neq 0, \\ \infty & \textit{if } \hat{S}_{\mathrm{K3}}(u,v) = 0. \end{cases} \tag{3.19b}$$

Let

$$P = \sum_{i=1}^{\ell} P_i \qquad Q = \sum_{i=1}^{\ell} Q_i \qquad R = \sum_{i=1}^{\ell} R_i. \tag{3.20}$$

Equation (3.19a) can be rewritten as

$$\begin{aligned}\hat{S}_{\mathrm{K3}}(u,v) =& 1 - \frac{4}{\ell}(P+Q+R) + \frac{4}{\ell^2}(P^2+Q^2+R^2) + \frac{12}{\ell^2}(PQ+PR+QR) \\ & - \frac{8}{\ell^3}(P^2Q+PQ^2+P^2R+PR^2+Q^2R+QR^2) - \frac{16}{\ell^3}PQR.\end{aligned} \tag{3.21}$$

Notice that when

$$\mathbf{M}_{uv} = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix},$$

the vector $\langle P, Q, R, \ell - P - Q - R\rangle$ in Equation (3.21) has a multinomial distribution with parameters $(\ell, p, q, r, 1-p-q-r)$. In order to analyze the statistical properties of $\hat{S}_{\mathrm{K3}}$, we use the following general lemma on multinomially distributed random variables.

**Lemma 3.13.** *Let $k > 1$ be a positive integer, and let $p_1, p_2, \ldots, p_k$ be non-negative numbers such that $\sum_{i=1}^{k} p_i = 1$. For all non-negative integers $\ell, \alpha_1, \alpha_2, \ldots, \alpha_k$, define*

$$E(\ell, \alpha_1, \alpha_2, \ldots, \alpha_k) = \mathbb{E} \prod_{i=1}^{k} \eta_i^{\alpha_i},$$

*where the random vector $\langle \eta_1, \ldots, \eta_k\rangle$ has a multinomial distribution with parameters $(\ell, p_1, p_2, \ldots, p_k)$ so that*

$$\mathbb{P}\Big\{\eta_1 = i_1, \eta_2 = i_2, \ldots, \eta_k = i_k\Big\} = \binom{\ell}{i_1\, i_2 \cdots i_k} p_1^{i_1} p_2^{i_2} \ldots p_k^{i_k}$$

*for all* $i_1 + i_2 + \cdots + i_k = \ell$. *The following induction rules hold.*

$$E(\ell, 0, 0, \ldots, 0) = 1; \tag{3.22a}$$

$$E(\ell, \alpha_1, \ldots, \alpha_k) = 0 \qquad \textit{if } \ell < \sum_{i=1}^{k} \mathbb{I}\{\alpha_i > 0\}; \tag{3.22b}$$

$$\begin{aligned} E(\ell, \alpha_1, \ldots, \alpha_{j-1}, \alpha_j + 1, \alpha_{j+1} \ldots, \alpha_k) &= \mathbb{E}\left[\eta_j \prod_{i=1}^{k} \eta_i^{\alpha_i}\right] \\ = \ell p_j \sum_{i=0}^{\alpha_j} \binom{\alpha_j}{i} E(\ell - 1, \alpha_1, \ldots, \alpha_{j-1}, i, \alpha_{j+1}, \ldots, \alpha_k) & \qquad \textit{otherwise.} \end{aligned} \tag{3.22c}$$

Proof. Equations (3.22a) and (3.22b) are trivial. We have to prove Equation (3.22c) only for $j = 1$ since the claim follows from symmetry for $j > 1$. Assume that $\ell > 0$ (for $\ell = 0$, Equation (3.22b) applies). By the definition of the expected value,

$$\begin{aligned} &\mathbb{E}\left[\eta_1 \prod_{i=1}^{k} \eta_i^{\alpha_i}\right] \\ &\quad = \sum_{i_1 + \cdots + i_k = \ell} \binom{\ell}{i_1\, i_2 \cdots i_k} i_1^{\alpha_1 + 1} i_2^{\alpha_2} \cdots i_k^{\alpha_k} p_1^{i_1} p_2^{i_2} \cdots p_k^{i_k} \\ &\quad = \ell p_1 \sum_{i_1 + \cdots + i_k = \ell} \frac{(\ell - 1)!}{(i_1 - 1)! i_2! \cdots i_k!} (i_1 - 1 + 1)^{\alpha_1} i_2^{\alpha_2} \cdots i_k^{\alpha_k} p_1^{i_1 - 1} p_2^{i_2} \cdots p_k^{i_k} \\ &\quad = \ell p_1 \sum_{i + i_2 + \cdots + i_k = \ell - 1} \binom{\ell - 1}{i\, i_2 \cdots i_k} \left(\sum_{j=0}^{\alpha_1} \binom{\alpha_1}{j} i^j\right) i_2^{\alpha_2} \cdots i_k^{\alpha_k} p_1^{i} p_2^{i_2} \cdots p_k^{i_k} \\ &\quad = \ell p_1 \sum_{j=0}^{\alpha_1} \binom{\alpha_1}{j} E(\ell - 1, j, \alpha_2, \ldots, \alpha_k). \quad \blacksquare \end{aligned}$$

**Corollary 3.14.** *Let* $\langle \eta_1, \eta_2, \ldots, \eta_k \rangle$ *be a multinomially distributed random* *vector with parameters* $(\ell, p_1, p_2, \ldots, p_k)$. *Then*

$$\begin{aligned}
\mathbb{E}\eta_1 &= \ell p_1; && \text{(3.23a)}\\
\mathbb{E}\eta_1\eta_2 &= \ell(\ell-1)p_1p_2; && \text{(3.23b)}\\
\mathbb{E}\eta_1^2 &= \ell(\ell-1)p_1^2 + \ell p_1; && \text{(3.23c)}\\
\mathbb{E}\eta_1\eta_2\eta_3 &= \ell(\ell-1)(\ell-2)p_1p_2p_3; && \text{(3.23d)}\\
\mathbb{E}\eta_1^2\eta_2 &= \ell(\ell-1)(\ell-2)p_1^2p_2 + \ell(\ell-1)p_1p_2; && \text{(3.23e)}\\
\mathbb{E}\eta_1^3 &= \ell(\ell-2)(\ell-2)p_1^3 + 3\ell(\ell-1)p_1^2 + \ell p_1; && \text{(3.23f)}\\
\mathbb{E}\prod_{i=1}^{k'}\eta_i &= \ell(\ell-1)\cdots(\ell-k'+1)p_1p_2\cdots p_{k'} \qquad \textit{for } k' \le k. && \text{(3.23g)}
\end{aligned}$$

**Lemma 3.15.** *Let* $\hat{S}_{\mathrm{K3}}$ *be the estimator for* $S_{\mathrm{K3}}$ *defined by Definition 3.8. Then for all nodes* $u, v$ *of a phylogeny in Kimura's three parameter model,*

$$0 \le \mathbb{E}\hat{S}_{\mathrm{K3}}(u,v) - S_{\mathrm{K3}}(u,v) \le \frac{1}{\ell} + \frac{1}{\ell^2} \le \frac{2}{\ell}. \tag{3.24}$$

PROOF. Let

$$\mathbf{M}_{uv} = \begin{bmatrix} 1-p-q-r & p & q & r \\ p & 1-p-q-r & r & q \\ q & r & 1-p-q-r & p \\ r & q & p & 1-p-q-r \end{bmatrix}.$$

By Equations (3.21) and (3.23),

$$\begin{aligned}
\mathbb{E}\hat{S}_{\mathrm{K3}}(u,v) - S_{\mathrm{K3}}(u,v) = \frac{4}{\ell}\Big(&(p+q+r) - 7(pq+pr+qr) - (p^2+q^2+r^2)\\
&+ 6(p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2) + 12pqr\Big)\\
+ \frac{16}{\ell^2}\Big(&(pq+pr+qr) - 2pqr\\
&- (p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2)\Big).
\end{aligned}$$

The maxima and minima of the two terms on the right-hand side of the equation can be found by derivation, yielding Equation (3.24). ■

**Lemma 3.16.** *Let $\hat{S}_{\mathrm{K3}}$ be the estimator for $S_{\mathrm{K3}}$ defined by Definition 3.8. The following inequalities hold when the sample sequences have length $\ell$. For all nodes $u, v$ of a phylogeny in Kimura's three parameter model, and for all $\epsilon > 0$,*

*if $S_{\mathrm{K3}}(u,v) > 0$, then*

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \le 1-\epsilon\right\} \le \exp\left(-\frac{1}{2}\ell S^2_{\mathrm{K3}}(u,v)\epsilon^2\right), \tag{3.25a}$$

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \ge 1+\epsilon\right\} \le \exp\left(-\frac{1}{2}\ell\left(\epsilon S_{\mathrm{K3}}(u,v) - \frac{2}{\ell}\right)^2\right); \tag{3.25b}$$

*and if $S_{\mathrm{K3}}(u,v) < 0$, then*

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \le 1-\epsilon\right\} \le \exp\left(-\frac{1}{2}\ell\left(\epsilon S_{\mathrm{K3}}(u,v) - \frac{2}{\ell}\right)^2\right), \tag{3.25c}$$

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \ge 1+\epsilon\right\} \le \exp\left(-\frac{1}{2}\ell S^2_{\mathrm{K3}}(u,v)\epsilon^2\right). \tag{3.25d}$$

The lemma follows from McDiarmid's inequality (1989), a Chernoff bound.

**Theorem 3.17.** (McDiarmid 1989)

*Let $\eta_1, \eta_2, \ldots, \eta_\ell$ be independent random variables taking values in a set $W$. Let $f\colon W^\ell \mapsto \mathbb{R}$ be a function such that for $i = 1, 2, \ldots, \ell$,*

$$\sup_{\substack{x_1,\ldots,x_\ell \\ x'_i \in W}} \Big| f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_\ell) - f(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_\ell)\Big| \le c_i.$$

*Let $c^2 = \sum_{i=1}^{\ell} c_i^2$. For every $\epsilon > 0$,*

$$\mathbb{P}\Big\{f(\eta_1,\ldots,\eta_\ell) - \mathbb{E}f(\eta_1,\ldots,\eta_\ell) \ge \epsilon\Big\} \le e^{-2\epsilon^2/c^2}; \tag{3.26a}$$

$$\mathbb{P}\Big\{f(\eta_1,\ldots,\eta_\ell) - \mathbb{E}f(\eta_1,\ldots,\eta_\ell) \le -\epsilon\Big\} \le e^{-2\epsilon^2/c^2}. \tag{3.26b}$$

PROOF OF LEMMA 3.16. By Equations (3.24), if $S_{\mathrm{K3}}(u,v) > 0$, then

$$\begin{aligned}\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \le 1-\epsilon\right\} &= \mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) \le S_{\mathrm{K3}}(u,v) - \epsilon S_{\mathrm{K3}}(u,v)\right\}\\ &\le \mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) - \mathbb{E}\hat{S}_{\mathrm{K3}}(u,v) \le -\epsilon \hat{S}_{\mathrm{K3}}(u,v)\right\}; \quad (*)\end{aligned}$$

$$\begin{aligned}\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \ge 1+\epsilon\right\} &= \mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) \ge S_{\mathrm{K3}}(u,v) + \epsilon S_{\mathrm{K3}}(u,v)\right\}\\ &\le \mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) - \mathbb{E}\hat{S}_{\mathrm{K3}}(u,v) \ge \epsilon \hat{S}_{\mathrm{K3}}(u,v) - \tfrac{2}{\ell}\right\}. \quad (**)\end{aligned}$$

Consider the function $f_{\mathrm{K3}}$ introduced in Equation (3.19a), as a function of the symbol pairs

$$(X_1^{(u)}, X_1^{(v)}), (X_2^{(u)}, X_2^{(v)}), \ldots, (X_\ell^{(u)}, X_\ell^{(v)}).$$

Replacing any of the symbol pairs by another pair from $\{\mathtt{A},\mathtt{G},\mathtt{T},\mathtt{C}\}^2$ changes the value of $f_{\mathrm{K3}}$ by at most $\frac{2}{\ell}$. Thus McDiarmid's inequality can be applied to $f_{\mathrm{K3}}$ by plugging $c_i = \frac{2}{\ell}$ into Equation (3.26) to obtain bounds for the right-hand sides of Equations $(*)$ and $(**)$ shown by Equations (3.25a) and (3.25b). Equations (3.25c) and (3.25d) are obtained analogously. ■

McDiarmid's inequality is very useful in deriving concentration inequalities for complicated functions $f$ of random variables, but may bring in unattractive terms in the exponent, such as the $(-2/\ell)$ in Equations (3.25b) and (3.25c) due to the bias $\mathbb{E}f(\boldsymbol{\eta}) - f(\mathbb{E}\boldsymbol{\eta})$. The following lemma derives a bound using Hoeffding's inequality instead.

**Lemma 3.18.** *Let $\hat{S}_{\mathrm{K3}}$ be the estimator for $S_{\mathrm{K3}}$ defined by Definition 3.8. For all nodes $u, v$ of a phylogeny in Kimura's three parameter model, if $S_{\mathrm{JC}}(u,v) \ne 0$, then for every sequence length $\ell$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \le 1-\epsilon\right\} \le 6\exp\left(-\frac{1}{72}\ell S_{\mathrm{K3}}^2(u,v)\epsilon^2\right); \tag{3.27a}$$

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \ge 1+\epsilon\right\} \le 6\exp\left(-\frac{1}{72}\ell S_{\mathrm{K3}}^2(u,v)\epsilon^2\right). \tag{3.27b}$$

PROOF. If $S_{\mathrm{K3}}(u,v) > 0$, then the left-hand side of Equation (3.27a) can be rewritten as

$$\mathbb{P}\left\{\frac{\hat{S}_{\mathrm{K3}}(u,v)}{S_{\mathrm{K3}}(u,v)} \le 1-\epsilon\right\} = \mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) - S_{\mathrm{K3}}(u,v) \le -\epsilon S_{\mathrm{K3}}(u,v)\right\}. \qquad (*)$$

. By Equations (3.19a) and (3.20),

$$\hat{S}_{\mathrm{K3}}(u,v) = f(P,Q,R) = \left(1-\tfrac{2}{\ell}(P+Q)\right)\left(1-\tfrac{2}{\ell}(P+R)\right)\left(1-\tfrac{2}{\ell}(Q+R)\right).$$

The function $f$ is a Lipschitz function, since $0 \le P,Q,R \le \ell$ and thus

$$\begin{aligned} |f(P,Q,R) - f(P',Q',R')| &\le 4\left(\frac{|P-P'|}{\ell} + \frac{|Q-Q'|}{\ell} + \frac{|R-R'|}{\ell}\right) \\ &\le \frac{12}{\ell}\max\left\{|P-P'|, |Q-Q'|, |R-R'|\right\}. \end{aligned}$$

Since $S_{\mathrm{K3}}(u,v,w) = f(\mathbb{E}P, \mathbb{E}Q, \mathbb{E}R)$, the right-hand side of Equation $(*)$ can be bounded from above as

$$\begin{aligned} \mathbb{P}&\left\{\hat{S}_{\mathrm{K3}}(u,v) - S_{\mathrm{K3}}(u,v) \le -\epsilon S_{\mathrm{K3}}(u,v)\right\} \\ &\le \mathbb{P}\left\{|P-\mathbb{E}P| \ge \frac{\epsilon S_{\mathrm{K3}}(u,v)}{12}\right\} \\ &\quad + \mathbb{P}\left\{|Q-\mathbb{E}Q| \ge \frac{\epsilon S_{\mathrm{K3}}(u,v)}{12}\right\} \qquad (**) \\ &\quad + \mathbb{P}\left\{|R-\mathbb{E}R| \ge \frac{\epsilon S_{\mathrm{K3}}(u,v)}{12}\right\}. \end{aligned}$$

Applying Hoeffding's inequality (Theorem 3.12) to the random variables $P$, $Q$, and $R$, the right-hand side of Equation $(**)$ can be further bounded from above by

$$\mathbb{P}\left\{\hat{S}_{\mathrm{K3}}(u,v) - S_{\mathrm{K3}}(u,v) \le -\epsilon S_{\mathrm{K3}}(u,v)\right\} \le 6\exp\left(-\frac{1}{72}\ell S_{\mathrm{K3}}^2(u,v)\epsilon^2\right).$$

Together with Equation $(*)$ this proves Equation (3.27a) if $S_{\mathrm{K3}}(uv) > 0$. Equation (3.27b) and the bounds in the case when $S_{\mathrm{K3}}(u,v)$ is negative are proven similarly. ∎

### 3.4.3 Paralinear distance

In order to define empirical paralinear distance, we first introduce a method to estimate the mutation matrices from a sample.

**Definition 3.9.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. Let $u, v$ be two nodes associated with random taxon sequences $X^{(u)}$ and $X^{(v)}$. Let*

$$\ell = \left|X^{(u)}\right| = \left|X^{(v)}\right|.$$

*Define*

$$N_{ij} = \sum_{k=1}^{\ell} \mathbb{I}\left\{X_k^{(u)} = i, X_k^{(v)} = j\right\} \qquad i, j \in \mathcal{A};$$
$$N_i = \sum_{j \in \mathcal{A}} N_{ij} \qquad i \in \mathcal{A}.$$

*The $m \times m$* empirical mutation matrix $\hat{\mathbf{M}}_{uv}$ *is defined by its entries as*

$$\hat{\mathbf{M}}_{uv}[i, j] = \begin{cases} \frac{N_{ij}}{N_i} & \text{if } N_i \neq 0; \\ \mathbb{I}\{i = j\} & \text{if } N_i = 0. \end{cases} \tag{3.28}$$

It is also possible to use a Laplace-style estimator for the entries of $\mathbf{M}_{uv}$ by defining, for example

$$\tilde{\mathbf{M}}_{vv}[i, j] = \frac{N_{ij} + 1/\ell}{N_i + m/\ell}.$$

However, this estimator converges only with a speed of $O\left(\ell^{-1}\right)$, as opposed to the exponential convergence of the entries in $\hat{\mathbf{M}}_{uv}$ shown by the next lemma.

**Lemma 3.19.** *Let $u$ and $v$ be two nodes of a phylogeny in the i. i. d. Markov model. For arbitrary symbols $i, j \in \mathcal{A}$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{\hat{\mathbf{M}}_{uv}[i, j] - \mathbf{M}_{uv}[i, j] \geq \epsilon\right\} \leq \exp\left(-(1 - e^{-2})\ell\pi_i^{(u)}\epsilon^2\right); \tag{3.29a}$$

$$\mathbb{P}\left\{\hat{\mathbf{M}}_{uv}[i, j] - \mathbf{M}_{uv}[i, j] \leq -\epsilon\right\} \leq \exp\left(-(1 - e^{-2})\ell\pi_i^{(u)}\epsilon^2\right). \tag{3.29b}$$

PROOF. As a shorthand notation, define

$$\begin{aligned} p =& \mathbf{M}_{uv}[i,j]; \\ \hat{p} =& \hat{\mathbf{M}}_{uv}[i,j]; \\ q =& \pi_i^{(u)}. \end{aligned}$$

We first prove Equation (3.29a). Let us assume that $\epsilon < 1-p = 1-\mathbf{M}_{uv}[i,j]$. Otherwise the equation is trivial because $\hat{\mathbf{M}}_{uv}[i,j]$ is never larger than one. The random variable $N_{ij}$ has a binomial distribution with parameters $N_i$ and $p$. Thus for every $k = 0, 1, \ldots, \ell$, Hoeffding's inequality (Theorem 3.12) implies that

$$\mathbb{P}\Big\{N_{ij} \geq k(p+\epsilon) \;\Big|\; N_i = k\Big\} \leq e^{-2k\epsilon^2}. \tag{$*$}$$

The inequality holds vacuously even for $k = 0$. Since the random variable $N_i$ has a binomial distribution with parameters $\ell$ and $q$, by Equation $(*)$,

$$\begin{aligned} \mathbb{P}\Big\{\hat{p} \geq p+\epsilon\Big\} &= \sum_{k=0}^{\ell} \mathbb{P}\Big\{N_{ij} \geq k(p+\epsilon) \;\Big|\; N_i = k\Big\}\mathbb{P}\Big\{N_i = k\Big\} \\ &\leq \sum_{k=0}^{\ell} \binom{\ell}{k} q^k (1-q)^{\ell-k} e^{-2\epsilon^2} = \Big(1 - q + qe^{-2\epsilon^2}\Big)^{\ell}. \end{aligned} \tag{$**$}$$

Define

$$\phi(x) = -\ln\Big(1 - q + qe^{-2x}\Big).$$

Since the function $\phi$ is concave and $\phi(0) = 0$, for every $x < x'$,

$$\phi(x) \geq x\frac{\phi(x')}{x'}.$$

In particular,

$$\phi(\epsilon^2) \geq \epsilon^2 \frac{\phi\Big((1-p)^2\Big)}{(1-p)^2}, \tag{$***$}$$

since $\epsilon < 1 - p$. Therefore,

$$\begin{aligned}
\mathbb{P}\Big\{\hat{p} \geq p + \epsilon\Big\} &= \exp\Big(-\ell\phi(\epsilon^2)\Big) && \text{by Eq. } (**) \\
&\leq \exp\left(\ell\epsilon^2 \frac{\ln(1 - q - qe^{-2(1-p)^2})}{(1-p)^2}\right) && \text{by Eq. } (***) \\
&\leq \exp\left(-\ell q\epsilon^2 \frac{1 - e^{-2(1-p)^2}}{(1-p)^2}\right) && x \leq -\ln(1-x) \\
&\leq \exp\Big(-(1 - e^{-2})\ell q\epsilon^2\Big) && \min_{x\in[0,1]} \frac{1 - e^{-2x^2}}{x^2} = 1 - e^{-2}
\end{aligned}$$

corresponding to Equation (3.29a). The proof of Equation (3.29b) is analogous. ∎

**Definition 3.10.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. Let $u, v$ be two nodes. The empirical similarity $\hat{S}_\mathrm{L}$ is defined using the empirical mutation matrices of Definition 3.9 as*

$$\hat{S}_\mathrm{L}(u, v) = \sqrt{\Big(\det \hat{\mathbf{M}}_{uv}\Big)\Big(\det \hat{\mathbf{M}}_{vu}\Big)}. \tag{3.30a}$$

*The empirical distance $\hat{D}_\mathrm{L}$ between the two nodes is defined as*

$$\hat{D}_\mathrm{L}(u, v) = \begin{cases} -\ln \hat{S}_\mathrm{L}(u, v) & \text{if } \hat{S}_\mathrm{L}(u, v) \neq 0; \\ \infty & \text{if } \hat{S}_\mathrm{L}(u, v) = 0. \end{cases} \tag{3.30b}$$

By Lemma 3.19, the entries of $\hat{\mathbf{M}}_{uv}$ converge quickly to those of $\mathbf{M}_{uv}$. The speed of convergence for row $i$ is primarily determined by $(1 - e^{-2})\pi_i^{(u)} \approx 0.9\pi_i^{(u)}$. One can suspect that the determinants of the mutation matrices and therefore the empirical similarity from Equation (3.30a) also converge quickly. In order to establish the upper bounds we need an auxiliary lemma regarding the difference between determinants of stochastic matrices.

**Lemma 3.20.** *Let $\mathbf{A}$ and $\mathbf{B}$ be two $m \times m$ stochastic matrices and*

$$\mathcal{L}_\infty(\mathbf{A}, \mathbf{B}) = \max\Big\{|\mathbf{A}[i, j] - \mathbf{B}[i, j]| \colon i, j = 1, \ldots, m\Big\}.$$

*Then*

$$|\det \mathbf{A} - \det \mathbf{B}| \leq m(m-1)\mathcal{L}_\infty(\mathbf{A}, \mathbf{B}) . \tag{3.31}$$

PROOF. Define $\mathbf{A}_{kj}$ as the the matrix obtained by deleting row $k$ and column $j$ in $\mathbf{A}$. We bound $|\det \mathbf{A}_{kj}|$ from above in the following manner. Let

$$\mathbf{\Lambda} = \operatorname{diag}\Big(1 - \mathbf{A}[1, j], \ldots, 1 - \mathbf{A}[k-1, j], 1 - \mathbf{A}[k+1, j], \ldots, 1 - \mathbf{A}[m, j]\Big);$$
$$\mathbf{A}'_{kj} = \mathbf{\Lambda}^{-1}\mathbf{A}_{kj};$$

i.e., $\mathbf{A}'_{kj}$ is the matrix obtained from $\mathbf{A}_{kj}$ by dividing rows $i = 1, \ldots, k-1$ with $(1 - \mathbf{A}[i, j])$ and rows $i = k, \ldots, m-1$ with $(1 - \mathbf{A}[i+1, j])$. The matrix $\mathbf{A}'_{kj}$ is a stochastic matrix, hence for every $k' \neq k$,

$$|\det \mathbf{A}_{kj}| = \Big|(\det \mathbf{\Lambda})(\det \mathbf{A}'_{kj})\Big| \leq \det \mathbf{\Lambda} \leq 1 - \mathbf{A}[k', j].$$

Let $\mathbf{A}$ and $\mathbf{B}$ differ only in row $k$; i.e., $\mathbf{A}[i, j] = \mathbf{B}[i, j]$ for every $i \neq k$ and every $j$. By expanding the determinants by row $k$, and choosing an arbitrary $k' \neq k$,

$$\begin{aligned}
|\det \mathbf{A} - \det \mathbf{B}| &= \left|\sum_{j=1}^{m}(-1)^{k+j}(\mathbf{A}[k, j] - \mathbf{B}[k, j]) \det \mathbf{A}_{kj}\right| \\
&\leq \sum_{j=1}^{m}|\mathbf{A}[k, j] - \mathbf{B}[k, j]|(1 - \mathbf{A}[k', j]) \\
&\leq \mathcal{L}_\infty(\mathbf{A}, \mathbf{B}) \sum_{j=1}^{m}(1 - \mathbf{A}[k', j]) = (m-1)\mathcal{L}_\infty(\mathbf{A}, \mathbf{B}) . \qquad (*)
\end{aligned}$$

Let $m' \leq m$ be the number of different rows between $\mathbf{A}$ and $\mathbf{B}$. Let $\mathbf{A}^{(0)} = \mathbf{A}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m')} = \mathbf{B}$ be a series of matrices such that each one of them differs by only one row from the previous one. By Equation $(*)$,

$$|\det \mathbf{A} - \det \mathbf{B}| \leq \sum_{k=1}^{m'}|\det \mathbf{A}^{(i)} - \det \mathbf{A}^{(i-1)}| \leq m'(m-1)\mathcal{L}_\infty(\mathbf{A}, \mathbf{B}) ,$$

implying Equation (3.31). ■

**Lemma 3.21.** *Let $u$ and $v$ be two nodes of an evolutionary tree in the i. i. d.*

*Markov model. Define*

$$\pi_{\min}^{(u)} = \min_{i\in\mathcal{A}} \pi_i^{(u)}. \tag{3.32}$$

*For arbitrary $\epsilon > 0$,*

$$\mathbb{P}\Bigl\{\Bigl|\det\hat{\mathbf{M}}_{uv} - \det\mathbf{M}_{uv}\Bigr| \geq \epsilon\Bigr\} \leq 2m^2\exp\Bigl(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}^{(u)}\epsilon^2\Bigr). \tag{3.33}$$

Proof. By Lemmas 3.19 and 3.20,

$$\begin{aligned}
\mathbb{P}&\Bigl\{\Bigl|\det\hat{\mathbf{M}}_{uv} - \det\mathbf{M}_{uv}\Bigr| \geq \epsilon\Bigr\} \\
&\leq \mathbb{P}\Bigl\{\exists i,j\colon \Bigl|\det\hat{\mathbf{M}}_{uv}[i,j] - \det\mathbf{M}_{uv}[i,j]\Bigr| > \frac{\epsilon}{m(m-1)}\Bigr\} \\
&\leq 2m^2\exp\Bigl(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}^{(u)}\epsilon^2\Bigr),
\end{aligned}$$

concluding the proof. ■

**Theorem 3.22.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. Let*

$$\pi_{\min} = \min_{u\in V} \pi_{\min}^{(u)} = \min_{u\in V}\min_{i\in\mathcal{A}} \pi_i^{(u)}; \tag{3.34}$$

$$\pi_{\text{span}} = \min_{u,v\in V} \frac{\prod_{i\in\mathcal{A}} \pi_i^{(u)}}{\prod_{i\in\mathcal{A}} \pi_i^{(v)}}. \tag{3.35}$$

*For all nodes $u, v \in V$, sample length $\ell$ and $\epsilon > 0$,*

$$\mathbb{P}\Bigl\{\Bigl|\frac{\hat{S}_{\mathrm{L}}(u,v)}{S_{\mathrm{L}}(u,v)} - 1\Bigr| \geq \epsilon\Bigr\} \leq 4m^2\exp\Bigl(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}\pi_{\text{span}}S_{\mathrm{L}}^2(u,v)\epsilon^2\Bigr). \tag{3.36a}$$

*Moreover, if $\mathcal{P}$ is in the time-reversible model, then*

$$\mathbb{P}\Bigl\{\Bigl|\frac{\hat{S}_{\mathrm{L}}(u,v)}{S_{\mathrm{L}}(u,v)} - 1\Bigr| \geq \epsilon\Bigr\} \leq 4m^2\exp\Bigl(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}S_{\mathrm{L}}^2(u,v)\epsilon^2\Bigr). \tag{3.36b}$$

Proof. By Lemma 3.21,

$$\mathbb{P}\left\{\left|\frac{\det \hat{\mathbf{M}}_{uv}}{\det \mathbf{M}_{uv}} - 1\right| \geq \epsilon\right\} \leq 2m^2 \exp\left(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}\left(\det \mathbf{M}_{uv}\right)^2\epsilon^2\right). \quad (*)$$

By Equations (3.7b) and (3.8c),

$$\left(\det \mathbf{M}_{uv}\right)^2 = S_{\mathrm{L}}^2(u,v)\frac{\prod_{i=1}^m \pi_i^{(v)}}{\prod_{i=1}^m \pi_i^{(u)}};$$
$$\left(\det \mathbf{M}_{vu}\right)^2 = S_{\mathrm{L}}^2(u,v)\frac{\prod_{i=1}^m \pi_i^{(u)}}{\prod_{i=1}^m \pi_i^{(v)}}.$$

Consequently, by Equation $(*)$,

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{\hat{S}_{\mathrm{L}}(u,v)}{S_{\mathrm{L}}(u,v)} - 1\right| \geq \epsilon\right\} \leq &\mathbb{P}\left\{\left|\frac{\det \hat{\mathbf{M}}_{uv}}{\det \mathbf{M}_{uv}} - 1\right| \geq \epsilon\right\} \\ &+ \mathbb{P}\left\{\left|\frac{\det \hat{\mathbf{M}}_{vu}}{\det \mathbf{M}_{vu}} - 1\right| \geq \epsilon\right\} \\ \leq &4m^2 \exp\left(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}\pi_{\mathrm{span}}S_{\mathrm{L}}^2(u,v)\epsilon^2\right),\end{aligned}$$

proving the first half of the Theorem. If the mutation process is time-reversible, then the base frequencies are stationary, and thus $\det \mathbf{M}_{uv} = \det \mathbf{M}_{vu}$. Hence

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{\hat{S}_{\mathrm{L}}(u,v)}{S_{\mathrm{L}}(u,v)} - 1\right| \geq \epsilon\right\} \leq &\mathbb{P}\left\{\left|\frac{\det \hat{\mathbf{M}}_{uv}}{\det \mathbf{M}_{uv}} - 1\right| \geq \epsilon\right\} \\ &+ \mathbb{P}\left\{\left|\frac{\det \hat{\mathbf{M}}_{vu}}{\det \mathbf{M}_{vu}} - 1\right| \geq \epsilon\right\} \\ \leq &4m^2 \exp\left(-\frac{1-e^{-2}}{m^2(m-1)^2}\ell\pi_{\min}S_{\mathrm{L}}^2(u,v)\epsilon^2\right). \quad \blacksquare\end{aligned}$$

# Chapter 4

# Algorithms

## 4.1 Efficient topology recovery

Joseph Felsenstein lists 168 phylogeny software packages on his web page (Felsenstein 2000). Some packages implement more than one evolutionary tree reconstruction algorithm, and there are many algorithms that have not yet been implemented at all. It is beyond the scope of this dissertation to discuss in detail each algorithm invented so far, but we do present an overview of major approaches to phylogeny reconstruction. We stress efficiency issues in our review, both in terms of using computational resources and in terms of using biological resources, i.e., available molecular sequence data. Efficiency becomes especially consequential when one aspires to reconstruct large evolutionary trees. As a trivial example, if an algorithm's running time is exponential in the number of taxa, then that algorithm is impractical in reconstructing the topology of a tree with several hundred or thousand leaves. Recall that our goal is topology reconstruction, i.e., the recovery of the topology from sample sequences that an evolutionary tree generates within the probabilistic framework introduced in §1.4. Let $\mathcal{P} = (V, E, \mathbb{P})$ be an evolutionary tree. Let $L \subseteq V$ be the set of observable nodes, including every leaf. Ordinarily $L$ comprises exactly the leaves of $\mathcal{P}$. A topology reconstruction algorithm (see §1.4.2) outputs a hypothetical topology $\Psi^*$ based on a sample $\langle X^{(u)} \colon u \in L \rangle$ drawn according to the probability distribution $\mathbb{P}$. In what follows we define the computational and statistical efficiency of a topology reconstruction algorithm.

**Computational efficiency** The topological equivalence relation $\underset{L}{\sim}$ defines a finite number of equivalence classes over unrooted trees that include the vertices in $L$. We can thus envisage using a topology reconstruction algorithm that examines all possibilities in the hypothesis class $\mathcal{C}$ and picks one based on the sample. Unfortunately, such an approach is not feasible even in the case of moderately large phylogenies because the number of topological equivalence classes is too large. Edwards and Cavalli-Sforza (1963) state (concurred also by Harding (1971), Felsenstein (1978b), and Rohlf (1983)) that if $L$ is the set of leaves and $|L| = n$, then the number of different topology classes equals

$$(2n-5)!! = 1 \cdot 3 \cdots (2n-3)(2n-5). \tag{4.1}$$

At $n = 30$, for example, there are more than $8 \cdot 10^{36}$ possibilities to choose from, at $n = 40$, there are more than $10^{55}$ possibilities, and thus exhaustive search is doomed to fail. A topology reconstruction algorithm is *computationally efficient* if it produces its output in a time that is polynomial in the sample size, i.e., the number of characters in the sample. If the sample sequences have length at most $\ell$, and $|L| = n$, then the running time of a computationally efficient algorithm is bounded by a polynomial in $\ell$ and $n$.

**Statistical efficiency** Computational efficiency offers no hint of how successful the algorithm can be in recovering the topology. Recall that a topology reconstruction algorithm is successful if its output $\Psi^*$ is topologically equivalent to $\Psi(\mathcal{P})$ over the set of observed nodes $L$. The results of §3.4 show that empirical similarities and distances converge rapidly to their true values in the i. i. d. Markov model. A topology reconstruction algorithm building the tree from empirical distances should thus recover the topology with increasing success as larger samples become available. Higher success on longer sample sequences in more general sequence generation models is only possible if more information is gained from a longer sample. As the most extreme counterexample, if the random taxon sequence distribution is completely arbitrary, then the sample conveys no information about the topology, regardless of sample size.

Consistency is the formalization of the idea that a good topology reconstruction algorithm should recover the topology in the i. i. d. Markov model if an infinite amount of data is available, since then the random sequence distribution can be exactly calculated. We construct an infinite sample in the following manner. Let $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots$ be a series of random vectors distributed independently and identically to the random labels $\langle \xi^{(u)} \colon u \in L\rangle$ and define the sample sequence $\{\mathbf{X}_\ell \colon \ell = 1, 2, \ldots\}$ by

$$\mathbf{X}_\ell = \langle \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_\ell \rangle.$$

Observe that $\mathbf{X}_\ell$ has identical distribution to $\langle X^{(u)} \colon u \in L\rangle$ given that the sequences have length $\ell$. For a topology reconstruction algorithm $\mathcal{F}$, the probability $\mathbb{P}\Big\{\mathcal{F}\Big(\mathbf{X}_\ell\Big) \underset{L}{\sim} \Psi(\mathcal{P})\Big\}$ is its success probability on sequences of length $\ell$. It is a self-explanatory requirement that the success probability should converge to certainty as $\ell$ goes to infinity, i.e., convergence in probability, but consistency is in fact defined as almost sure convergence to the true topology, which is a stronger attribute. The algorithm $\mathcal{F}$ is *consistent*, if and only if

$$\lim_{\ell\to\infty} \mathcal{F}\Big(\mathbf{X}_\ell\Big) \underset{L}{\sim} \Psi(\mathcal{P}) \tag{4.2}$$

with probability one. The algorithm $\mathcal{F}$ is consistent on the hypothesis class $\mathcal{C}$ if it is consistent for all phylogenies $\mathcal{P} \in \mathcal{C}$.

Equation (4.2) expresses that the algorithm $\mathcal{F}$ recovers the topology if infinite amount of data is available. From a practical aspect, we are more concerned about the success on bounded sample lengths. Let $\mathbf{X}^{(L)} = \Big\langle X^{(u)} \colon u \in L\Big\rangle$, and let $\Big|\mathbf{X}^{(L)}\Big| = \min_{u\in L} \Big|X^{(u)}\Big|$. Given an *error probability* $0 < \delta < 1$, the *sample complexity* of $\mathcal{F}$ on a phylogeny $\mathcal{P}$ is defined as

$$\ell(\delta) = \min_{\ell}\Bigg\{\ell \colon \mathbb{P}\Big\{\mathcal{F}(\mathbf{X}^{(L)}) \underset{L}{\sim} \Psi(\mathcal{P}) \;\Big|\; |\mathbf{X}^{(L)}| \geq \ell\Big\} \geq 1 - \delta\Bigg\}, \tag{4.3}$$

i.e., the shortest sample length for which $\mathcal{F}$ recovers the topology with probability at least $(1 - \delta)$. If there is no such finite length, then $\ell(\delta) = \infty$.

The algorithm $\mathcal{F}$ is *statistically efficient* if the sample complexity is polynomial in the number of observed nodes $n = |L|$ and the error probability $\delta$, i.e., if $\ell(\delta)$ is bounded by a polynomial in $n$ and $\delta^{-1}$. Note that statistical efficiency is meaningful in the case of any phylogeny and not only in the i. i. d. Markov model.

Despite the abundance of available reconstruction algorithms (Swofford *et al.* 1996; Felsenstein 1988), most algorithms fail to be either computationally or statistically efficient. In reality, numerous theoretical results on topology reconstruction consist of showing the inconsistency, or lack of computational and statistical efficiency of certain algorithms. Very attractive ideas often produce NP-hard problems preventing the creation of computationally efficient algorithms. Statistical efficiency is even more rarely attained, a phenomenon to which we return in §4.4.6. We review the efficiency issues in conjunction with the three major families of algorithms: maximum likelihood, character-based, and distance-based methods. For simplicity, we concentrate on the case of reconstructing an evolutionary tree in the general Markov model from sequences associated with the leaves.

## 4.2 Maximum likelihood

One of the earliest proposals (Edwards and Cavalli-Sforza 1963) for evolutionary tree reconstruction is the use of maximum likelihood methods (Felsenstein 1973, 1981, 1983). Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny with leaf set $L \subset V$. Based on the random taxon sequence distribution $\mathbb{P}$, the likelihood of $\mathcal{P}$ generating a sample $\mathbf{x}$ is well-defined as

$$P(\mathbf{x}; \mathcal{P}) = \mathbb{P}\Big\{\langle X^{(u)}\colon u \in L\rangle = \mathbf{x}\Big\}.$$

The maximum likelihood algorithm selects the phylogeny $\mathcal{P}^*$ from the hypothesis class $\mathcal{C}$ for a given sample $\mathbf{x}$ that maximizes the likelihood, i.e., it computes the function

$$\mathcal{F}_{\mathrm{ML}}(\mathbf{x}) = \operatorname*{arg\,max}_{\mathcal{P}^* \in \mathcal{C}} P(\mathbf{x}; \mathcal{P}^*). \tag{4.4}$$

Chang (1996) proves that $\Psi(\mathcal{F}_{\mathrm{ML}})$ is a consistent topology estimate in the i. i. d. Markov model if the determinants of the edge mutation matrices differ from 0, and $\pm 1$. Unfortunately, the maximum in Equation (4.4) is hard to compute exactly, even in the i. i. d. Markov model, offering little alternative to exhaustive search examining all topologies. Moreover, the maximum is not even necessarily unique (Steel 1994a; Tuffley and Steel 1997). The maximum likelihood method is thus not effective computationally. Heuristics such as DNAML of Felsenstein (1993), FastDNAml of Olsen *et al.* (1994), and TrExML of Wolf *et al.* (2000) offer no guarantees of statistical efficiency. Even the heuristics are computationally expensive, and are rarely used for trees with more than 30–40 leaves. Parallel implementations (Matsuda *et al.*

1994; Trelles *et al.* 1998) can be used for trees with up to a few hundred leaves, but may run for several hours or even days on contemporary supercomputers. While maximum likelihood is not computationally efficient in reconstructing large trees, it can be used for reconstructing small subtrees, which are then combined into one large tree. Such an approach is employed among others in the Quartet Puzzling of Strimmer and von Haeseler (1996) and the Disc Covering Method of Huson *et al.* (1999).

## 4.3 Character-based methods

Character-based methods are founded in some sense on an entirely opposite philosophy to that of maximum likelihood, in that they select a topology without any reference to sequence probabilities. The principles of character-based methods can be traced back to the ideas of Willi Hennig (1950, 1966). The main theme of character-based methods is the simultaneous derivation of the topology and the unobserved sequences from the observed sequences. The underlying concepts are better justified if we bear in mind that the methods were originally devised in the context of using morphological characters such as "number of vertebrae", "has wings", etc., for which a probabilistic mutation model is inadequate. However, character-based methods are nowadays often used with molecular data, and possibly constitute the most popular approach to phylogenetic analysis, despite theoretical and practical drawbacks. Character-based methods are commonly further categorized into the classes of *compatibility methods* and *parsimony methods*.

### 4.3.1 Compatibility methods

Compatibility as a basis for phylogenetic analysis was proposed by Le Quesne (1969, 1972, 1974) (see also Estabrook 1972). The idea is that good characters for evolutionary tree reconstruction are binary indicator characters for traits that are acquired at most once, such as, for example, "has spinal chord". Such binary characters can also be based on molecular sequences, and can be indicators for highly specific subsequences in genes. The character sequences associated with the nodes consist of the indicators for the traits, so that each character position corresponds to exactly one feature. The use of such uniquely evolved features leads to the idea of *perfect phylogeny*. The perfect phylogeny for a sample is a rooted tree, where each sequence position is assigned an edge on which the character in that particular position changes from 0 to 1. Formally, let $L$ be a set of terminal taxa with associated sequences taken from $\{0,1\}^\ell$. A perfect phylogeny for a sample $\langle \mathbf{x}^{(u)} \colon u \in L \rangle$ is a rooted tree $\mathcal{T} = (V, E)$ over the leaf set $L$ with a bijection $f \colon E \mapsto \{1, 2, \ldots, \ell\}$ that has the following property. For each node $u \in V$, if the path from the root to $u$ is $u_0, e_1, \ldots, e_l, u_l = u$, then the characters of the sequence $\mathbf{x}^{(u)} = x_1^{(u)} \cdots x_\ell^{(u)}$ are determined as

$$x_k^{(u)} = \begin{cases} 1 & \text{if } k \in \{f(e_1), \ldots, f(e_l)\}; \\ 0 & \text{otherwise.} \end{cases}$$

The problem of perfect phylogeny (see, e.g., Warnow 1994) is that of determining whether a perfect phylogeny exists for a given sample, and determining it if it does. A solution to the problem was first proposed by Estabrook *et al.* (1975). Gusfield (1991, 1997) gives fast algorithms for the problem that run in $O(n\ell)$ time when $|L| = n$. Compatibility methods seek to select a maximal subset of character positions on which a perfect phylogeny can be built. The selection of such a maximal subset is NP-hard as proven by Day and Sankoff (1986).

The generalization of the perfect phylogeny problem to non-binary characters requires that the edges of the derived tree are labeled with sequence positions and character transitions by a mapping $f \colon E \mapsto \{1, \ldots, \ell\} \times \mathcal{A}^2$ so that on edge $e$, if $f(e) = \langle k, i, j \rangle$, then the parent sequence on that edge changes in position $k$ from symbol $i$ to $j$. As before, the path from the root to a node determines the associated sequence, and for each position $k$ and symbol $j \in \mathcal{A}$, there is at most one edge with $f(e) = \langle k, i, j \rangle$. Dress and Steel (1993) and Kannan and Warnow (1994) present computationally efficient algorithms for the generalized perfect phylogeny problem when $|\mathcal{A}| = 3, 4$. Furthermore, Agarwala and Fernández-Baca (1994) devised a polynomial time algorithm for an arbitrary fixed alphabet, which was further improved by Kannan and Warnow (1997). The problem is NP-hard in general, i.e., if the alphabet size is part of the input (Bodlaender *et al.* 1992; Steel 1992).

### 4.3.2 Parsimony methods

Conceptually, parsimony methods extend the idea of compatibility. The extension is the following. Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny with leaf set $L \subseteq V$. Character-based methods, including compatibility and parsimony, return a topology $\Psi^* = (V^*, E^*)$ with $L \subset V^*$, and a set of sequences $\mathbf{x} = \left\langle \mathbf{x}^{(u)} \colon u \in V^* \right\rangle$. Thus, there is a sequence $\mathbf{x}^{(u)}$ associated with each node $u \in V^*$, based on the sample, such that for every leaf $u \in L$, the sequence $\mathbf{x}^{(u)}$ is the sample sequence for $u$. Specifically, in the case of perfect phylogeny, on each edge $uv \in E^*$, $\mathbf{x}^{(u)}$ and $\mathbf{x}^{(v)}$ differ in exactly one position. Parsimony methods generalize this notion by introducing a sequence pair-weighting function $d\colon \mathcal{S} \times \mathcal{S} \mapsto [0, +\infty]$, and seeking the best choice of $(\Psi^*, \mathbf{x})$ that minimizes the sum of the weights on the edges

$$d(\Psi^*, \mathbf{x}) = \sum_{uv \in E^*} d\left(\mathbf{x}^{(u)}, \mathbf{x}^{(v)}\right). \tag{4.5}$$

Define the *Hamming distance* $\mathcal{H}$ between two sequences of equal length as the number of positions in which they differ

$$\mathcal{H}(s_1 s_2 \cdots s_\ell, t_1 t_2 \cdots t_\ell) = \sum_{k=1}^{\ell} \mathbb{I}\{s_k \neq t_k\}.$$

If it exists, the perfect phylogeny for a sample minimizes the sum for the weighting function

$$d(s,t) = \begin{cases} 0 & \text{if } |s| = |t| \text{ and } \mathcal{H}(s,t) = 1; \\ \infty & \text{otherwise.} \end{cases}$$

Of course this weighting function is not particularly interesting since the minimum is either 0 or $\infty$. A popular choice in parsimony methods is *Fitch parsimony*, which employs the Hamming distance as the weighting function

$$d(s,t) = \begin{cases} \mathcal{H}(s,t) & \text{if } |s| = |t|; \\ \infty & \text{otherwise.} \end{cases}$$

Consequently, the selection by Fitch parsimony minimizes the total number of character changes along the edges. As such, it can be considered as the most succinct way of representing the differences in the sample by a topology, which explains the name “parsimony”. The idea of using maximum parsimony as the basis for topology reconstruction was already described in the seminal paper of Edwards and Cavalli-Sforza (1963). The introduction of the term parsimony is attributed to Camin and Sokal (1965) by Felsenstein

(1988).

Generally, $d(s,t)$ is defined as the sum of weighted substitutions between $s$ and $t$. In other words, it is defined via a function $g\colon \mathcal{A}\times\mathcal{A}\mapsto[0,+\infty]$ so that

$$d(s_1s_2\cdots s_\ell, t_1t_2\cdots t_\ell)=\sum_{k=1}^{\ell} g(s_k,t_k). \tag{4.6}$$

Fitch parsimony uses $g(i,j)=\mathbb{I}\{i\neq j\}$. Other character weighting functions in the parsimony literature include Wagner parsimony and Dollo parsimony. In Wagner parsimony (Kluge and Farris 1969; Farris 1970) $g$ is additive and symmetric, i.e.,

$$\forall i,j\colon g(i,j)=g(j,i) \qquad \forall i,i',j\colon g(i,j)+g(j,i')=g(i,i').$$

In Dollo parsimony (Farris 1977) the alphabet is ordered and

$$\forall i>j\colon g(i,j)=\infty.$$

More complicated choices of weighting functions include ones based on PAM matrices (Dayhoff *et al.* 1978) and restriction site weights (Albert *et al.* 1992).

The minimization of the sum in Equation (4.5) raises two algorithmic problems.

> *Minimum mutation problem.* Given a topology $\Psi^*=(V^*,E^*)$ with leaf set $L\subset V^*$ and an assignment $\left\langle \mathbf{x}^{(u)}\colon u\in L\right\rangle$ of length $\ell$ sequences to the leaves, determine the sequence assignment $\mathbf{x}=\left\langle \mathbf{x}^{(u)}\colon u\in V^*\right\rangle$ that minimizes the penalty $d(\Psi^*,\mathbf{x})$.

> *Parsimony optimization problem.* Given a leaf set $L$ and an assignment $\left\langle \mathbf{x}^{(u)}\colon u\in L\right\rangle$ of length $\ell$ sequences to the leaves, determine the topology $\Psi^*$ and assignment $\mathbf{x}$ that minimizes the penalty $d(\Psi^*,\mathbf{x})$.

The first algorithm for the minimum mutation problem with Fitch parsimony was given by Fitch (1971) and Hartigan (1973). The Fitch-Hartigan algorithm runs in $O\left(|V^*|\ell\right)$ time. In the case where the weighting function $d$ is defined as the sum of weighted substitutions as in Equation (4.6), Sankoff (1975) and Sankoff and Rousseau (1975) describe fast algorithms running in $O\left(|V^*|\ell\right)$ time.

Despite the tractability of the minimum mutation problem, the parsimony optimization problem is NP-hard. In its simplest form, notably Fitch parsimony with a binary alphabet, the problem is equivalent to the unweighted Steiner tree problem on the $\ell$-dimensional binary hypercube. The Steiner tree problem is that of selecting the minimum weight subtree in an edge-weighted graph that spans a specific subset of nodes. Parsimony optimization is therefore equivalent to the Steiner tree problem by making nodes correspond to sample sequences, with the specific feature that the underlying graph (i.e., the $\ell$-dimensional hypercube) is not part of the problem instance description. The unweighted Steiner tree problem on binary hypercubes is NP-hard (Foulds and Graham 1982; Graham and Foulds 1982) and so is its weighted version (Gusfield 1984) in general. Day (1983a) proves that the problem is NP-hard for Wagner parsimony, and Day *et al.* (1986) prove that it is NP-hard for other popular weighting functions, including Dollo parsimony. There are results indicating that even the approximation problem is difficult (e.g., Fernández-Baca and Lagergren 1998). Therefore computational efficiency for algorithms delivering exact solutions to the parsimony optimization problem cannot be expected. Hendy and Penny (1982) describe two branch-and-bound algorithms for the problem, which they recommend to use for up to 16 leaves. Purdom *et al.* (2000) discuss recent improvements of branch-and-bound algorithms for parsimony.

Heuristics usually start by deriving an initial topology $\Psi_0^*$ with assigned sequences $\mathbf{x}$, and then search the space of possible topologies by swapping edges and subtrees, recomputing the optimal sequence assignment $\mathbf{x}$. The initial topology is often obtained by a greedy algorithm that adds leaves consecutively, described by Farris (1970) and first used by Kluge and Farris (1969). Fast heuristic search methods have been incorporated into major phylogeny packages such as Hennig86 (Farris 1988) and PAUP (Swofford 1990). Some recent ideas on accelerating the heuristic search have been discussed by Goloboff (1996) and Moilanen (1999).

Statistical efficiency of the maximum parsimony heuristics has been not established, although advocates of parsimony methods did report high success rates on some simulated data. Siddall (1998), for example, carried out extensive simulation experiments on four-leaf trees in the Jukes-Cantor model and found that heuristic parsimony implemented in PAUP performed better in certain cases than maximum likelihood and Neighbor-Joining (Saitou and Nei 1987) on sample lengths 100, 500, and 1000. Hillis *et al.* (1994), Hillis (1996), Rice and Warnow (1997), and Huson *et al.* (1998) also reported the

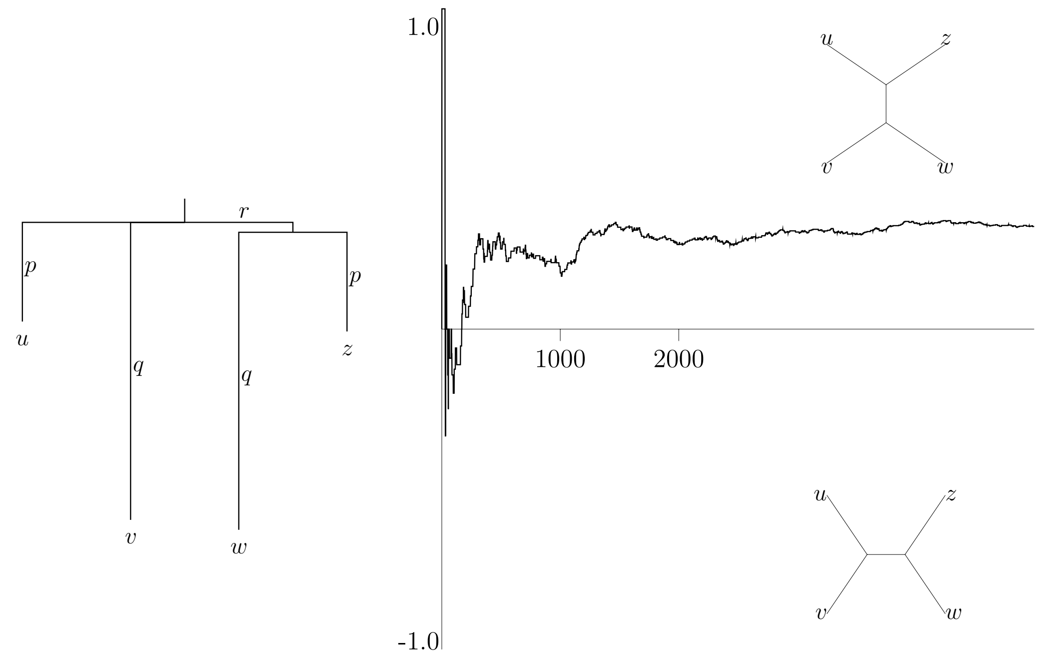
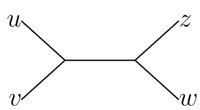


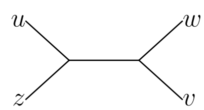

FIGURE 4.1: *An example showing the inconsistency of parsimony. The evolutionary tree on the left-hand side has four leaves, $u$, $v$, $w$, and $z$. The mutation model is the Jukes-Cantor model on a binary alphabet. The edge mutation probabilities are $p = 0.1$, $q = 0.3$ and $r = 0.01$. The graph on the right-hand side shows the ratio of labelings in which a wrong topology $\Psi_1$, shown on the top, is preferred over the true topology $\Psi_0$, shown on the bottom, for sample lengths up to 5000. The ratio is defined as follows. Let $k_0(i)$ be the number of labelings in which Fitch parsimony assigns a lower penalty score to the true topology $\Psi_0$ than to the topology $\Psi_1$ in the first $i$ labelings. Similarly, let $k_1(i)$ be the number of labelings in which $\Psi_1$ is preferred over $\Psi_0$. The graph plots $(k_1(i) - k_0(i))/(k_1(i) + k_0(i))$. We intend to illustrate in this random PostScript figure that with high probability, $k_1(i) > k_0(i)$ for finite $i$, and $\lim_{i\to\infty}(k_1(i) - k_0(i)) > 0$, i.e., that Fitch parsimony selects the wrong topology with high probability on finite samples and is inconsistent.*

success of heuristic parsimony for short sample lengths in simulated experiments on larger trees. Nevertheless, statistical efficiency and even consistency of parsimony methods are not to be expected in general. Cavender (1978) and Felsenstein (1978a) pointed out that maximum parsimony leads to a statistically inconsistent prediction of the topology on certain trees with four leaves in the i. i. d. Markov model. Hendy and Penny (1989) reported that the method is inconsistent also in the case of some trees with five leaves and constant substitution rates. The inconsistency result is sometimes referred to as “long branch attraction,” and is essentially due to the fact that edges with high mutation probabilities can make otherwise remote leaves seem closely related. Figure 4.1 shows a simple example with the following structure. Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the Jukes-Cantor model with a binary alphabet, for which

$$\begin{aligned} V &= \{x, x', u, v, w, z\}, \\ E &= \{xx', xu, xv, x'w, x'z\}, \end{aligned}$$
$$p_{xu} = p_{x'z} = p, \qquad p_{xv} = p_{x'w} = q, \qquad p_{xx'} = r.$$

Consider the following topologies.

$\Psi_0$: 


$\Psi_1$: 


The correct topology is $\Psi_0$, while $\Psi_1$ is one of the wrong topologies. Felsenstein (1978a) discovered that for certain edge mutation probabilities, Fitch parsimony prefers the topology $\Psi_1$ over $\Psi_0$, with high probability on finite sample lengths and with asymptotical certainty as the sample length grows toward infinity. In the random labelings, for which the topology $\Psi_0$ is preferred over $\Psi_1$,

$$\xi^{(v)} = \xi^{(u)} \qquad \xi^{(w)} = 1 - \xi^{(u)} \qquad \xi^{(z)} = 1 - \xi^{(u)}.$$

In labelings where $\Psi_1$ is preferred over $\Psi_0$,

$$\xi^{(v)} = 1 - \xi^{(u)} \qquad \xi^{(w)} = \xi^{(u)} \qquad \xi^{(z)} = 1 - \xi^{(u)}.$$

Denote the first event by $A_0$, the second by $A_1$. For example, if $r = 1/2$ and $q = 1 - p$, then

$$\mathbb{P}A_0 = 2p^2(1-p)^2, \qquad \mathbb{P}A_1 = \frac{1}{2}\Big(p^4 + (1-p)^4 + 2p^2(1-p)^2\Big).$$

Thus maximum parsimony prefers the wrong topology $\Psi_1$ if

$$p^4 + (1-p)^4 + 2p^2(1-p)^2 > 4p^2(1-p)^2$$
$$(1-2p)^4 > 0,$$

i.e., if $p \neq 1/2$. In general,

$$\begin{aligned}\mathbb{P}A_0 &= (1-r)\Big(2p(1-p)q(1-q)\Big) + r\Big(p^2q^2 + (1-p)^2(1-q)^2\Big)\\ \mathbb{P}A_1 &= (1-r)\Big(p^2(1-q)^2 + (1-p)^2q^2\Big) + r\Big(2p(1-p)q(1-q)\Big).\end{aligned}$$

Thus $\mathbb{P}A_1 > \mathbb{P}A_0$, if

$$1 - 2r > \frac{\Big((1-2p)+(1-2q)\Big)^2 - \Big((1-2p)-(1-2q)\Big)^2}{\Big((1-2p)+(1-2q)\Big)^2 + \Big((1-2p)-(1-2q)\Big)^2}.$$

In such a case Fitch parsimony is inconsistent, and in i. i. d. random labelings generating a sample, $A_1$ occurs more frequently than $A_0$ with high probability. An example is depicted in Figure 4.1. The statistical inadequacy of maximum parsimony is especially exposed by the results of Tuffley and Steel (1997) showing that Fitch parsimony is equivalent to maximum likelihood in the general Markov model. They credit Penny *et al.* (1994) with the result in the special case of a binary alphabet. These results, in turn, also prove that maximum likelihood is inconsistent in the general Markov model, as one would expect given that the number of parameters estimated by maximum likelihood is unbounded.

## 4.4 Distance-based methods

Compared to character-based and maximum likelihood methods, the family of distance-based methods comprises a vast variety of algorithmic approaches. The idea underlying these approaches is fairly simple. Since evolutionary distances determine the evolutionary tree topology, estimated distances should offer a sufficient basis for topology recovery. For a more formal discussion, let $D$ be a distance metric over a class $\mathcal{C}$ of evolutionary trees, and let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in $\mathcal{C}$. For a set $L \subseteq V$ of observed

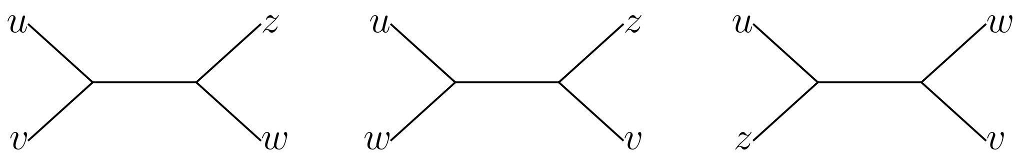



FIGURE 4.2: *Three possible topologies on four leaves. The quartet topologies are denoted by $uv|wz$, $uv|wz$, and $uv|wz$, respectively.*

nodes, the $|L| \times |L|$ *distance matrix* $\mathbf{\Delta}$ consists of the distances between the observed nodes

$$\mathbf{\Delta} = \Big[D(u,v) \colon u, v \in L\Big]. \tag{4.7}$$

Distance-based algorithms construct a hypothetical phylogeny $\mathcal{P}^*$ or topology $\Psi^*$ after a preprocessing step in which $\mathbf{\Delta}$ is estimated from the sample. The sample preprocessing produces an estimated distance matrix

$$\hat{\mathbf{\Delta}} = \Big[\hat{D}(u,v) \colon u, v \in L\Big], \tag{4.8}$$

where $\hat{D}(u,v)$ is the estimated distance between nodes $u$ and $v$ calculated from the sample sequences $X^{(u)}$ and $X^{(v)}$. Disregarding the origin of the estimated distance matrix $\hat{\mathbf{\Delta}}$, distance-based algorithms can be interpreted as clustering methods (Hartigan 1975; Barthélemy and Guénoche 1991) for "objects" in $L$, and in fact some methods arise from problems unrelated to molecular sequence data, for example in cognitive psychology (Sattath and Tversky 1977; Cunningham 1978; de Soete *et al.* 1987). Our discussion builds largely on studying how well $\hat{\mathbf{\Delta}}$ can estimate the matrix $\mathbf{\Delta}$ based on an analysis of the random taxon sequence distributions.

## 4.4.1 The four-point condition

Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny with leaf set $L$. Let $D$ be a distance metric on $\mathcal{P}$. By additivity, if the path between two arbitrary leaves $u, v \in L$ of $\Psi(\mathcal{P})$ is $u_0 = u, e_1, \ldots, e_l, u_l = v$, then $D(u,v) = \sum_{k=1}^{l} D(u_{k-1}, u_k)$. Equipping each edge of $\Psi(\mathcal{P})$ with a weight equal to the distance between the edge endpoints, $D(u,v)$ is the sum of the edge weights on the path between $u$ and $v$ in $\Psi(\mathcal{P})$.

In this respect, the distance matrix $\mathbf{\Delta}$ in Equation (4.7) is produced by an edge-weighted unrooted tree. Such an interpretation leads to the question of relationships between weighted trees and distance matrices. Stating the problem more explicitly, how can we tell whether a matrix corresponds to a weighted unrooted tree?

**Definition 4.1.** *Let $\mathbf{\Delta}$ be an $|L| \times |L|$ matrix whose columns and rows are indexed by elements of a set $L$ such that each entry $\mathbf{\Delta}[u, v] \in [0, \infty]$. The matrix $\mathbf{\Delta}$ is a* tree metric *over $L$ if and only if there exists an unrooted tree $\mathcal{T} = (V, E)$ with edge weights $d\colon E \mapsto [0, \infty]$ such that $L \subset V$ comprises the leaves, and for all $u, v \in E$, $\mathbf{\Delta}[u, v]$ equals the sum of edge weights on the path between $u$ and $v$. If such a tree $\mathcal{T}$ exists, then $\mathcal{T}$ is said to* fit *$\mathbf{\Delta}$. The edge weights $d$ are also referred to as* edge lengths.

The problem is thus to characterize tree metrics in general. The solution was first given independently by Smolenskiĭ (1962) and Hakimi and Yau (1964). They prove that $\mathbf{\Delta}$ is a tree metric if and only if the following properties hold.

Identity: for every $u \in L$, $\mathbf{\Delta}[u, u] = 0$.

Symmetry: for all $u, v \in L$, $\mathbf{\Delta}[u, v] = \mathbf{\Delta}[v, u]$.

Three-point condition: for every triple of different elements $u, v, w \in L$, $\mathbf{\Delta}[u, v] + \mathbf{\Delta}[v, w] \geq \mathbf{\Delta}[u, w]$.

Furthermore they prove that if $\mathbf{\Delta}$ has only positive finite entries outside the diagonal, then there is a unique tree (in the sense of topological equivalence over $L$) that fits $\mathbf{\Delta}$. By further studying tree metrics, another fundamental characteristic, known as the four-point condition, has been discovered, which is described as follows. A *quartet* is a set of four leaves in a tree. By Equation (4.1) there are three different quartet topologies. Figure 4.2 shows the three possibilities. For the quartet $\{u, v, w, z\}$ the three possible topologies are denoted by $uv|wz$, $uw|vz$ and $uz|vw$. When the matrix $\mathbf{\Delta}$ contains evolutionary distances, as in Equation (4.7), then by additivity of $D$, the quartet topology is $uv|wz$ if

$$\mathbf{\Delta}[u, v] + \mathbf{\Delta}[w, z] < \mathbf{\Delta}[u, w] + \mathbf{\Delta}[v, z] = \mathbf{\Delta}[u, z] + \mathbf{\Delta}[v, w]. \tag{4.9}$$

Furthermore, for any set of four leaves $\{u, v, w, z\}$,

$$\mathbf{\Delta}[u,v] + \mathbf{\Delta}[w,z] \le \max\Big\{\mathbf{\Delta}[u,w] + \mathbf{\Delta}[v,z], \mathbf{\Delta}[u,z] + \mathbf{\Delta}[v,w]\Big\}. \tag{4.10}$$

Equation (4.10) is the *four-point condition* for the quartet $\{u, v, w, z\}$.

**Theorem 4.1.** *A matrix $\mathbf{\Delta}$ for a set $L$ is a tree metric if and only if Equation* (4.10) *holds for every set $\{u, v, w, z\}$. If $\mathbf{\Delta}$ is a tree metric and the four-point condition of Equation* (4.10) *has strict inequality for every quartet, then there is a unique tree (in the sense of topological equivalence over $L$) that fits $\mathbf{\Delta}$.*

Theorem 4.1 was independently discovered by Zaretskiĭ (1965), Buneman (1971), and Patrinos and Hakimi (1972). If Equation (4.10) has strict inequality for every quartet, then all the quartet topologies are determined by Equation (4.9) and the topology can be recovered from the quartet topologies (Colonius and Schulze 1981; Bandelt and Dress 1986). Phrased differently, if $\mathbf{\Delta}$ is a tree metric with finite positive entries, then there is a unique tree $\mathcal{T}$ in which Equation (4.9) holds for every topological minor $uv|wz$ on four leaves. When $\mathbf{\Delta}$ is not a tree metric, however, it is NP-hard to determine the tree $\mathcal{T}$ that has the maximum number of topological minors on quartets satisfying Equation (4.9) (Steel 1992).

Many distance-based algorithms work by deducing the topologies of a set of quartets from the estimated distance matrix $\hat{\mathbf{\Delta}}$, and by employing a combinatorial method that derives a hypothetical topology $\Psi^*$ based on the quartet topologies. Typically, for a set $\{u, v, w, z\}$ of leaves, the topology $uv|wz$ is deduced if

$$\begin{aligned} \hat{\mathbf{\Delta}}[u,v] + \hat{\mathbf{\Delta}}[w,z] &< \hat{\mathbf{\Delta}}[u,w] + \hat{\mathbf{\Delta}}[v,z] \\ \hat{\mathbf{\Delta}}[u,v] + \hat{\mathbf{\Delta}}[w,z] &< \hat{\mathbf{\Delta}}[u,z] + \hat{\mathbf{\Delta}}[v,w]. \end{aligned} \tag{4.11}$$

Equation (4.11) is the *relaxed four-point condition.* Unlike in Equation (4.9) the equality of the right-hand sides is not required. The long list of distance-based algorithms using a combinatorial approach in conjunction with quartet topologies includes Buneman's (1971) algorithm, ADDTREE of Sattath and Tversky (1977), the $Q^*$ method of Berry and Gascuel (1997), Quartet Puzzling of Strimmer and von Haeseler (1996), the refined Buneman method (Berry and Bryant 1999), and the Short Quartet methods (Erdős *et al.* 1997, 1998, 1999a, 1999b).

### 4.4.2 The LogDet metric

Theorem 4.1 gives a sufficient and necessary condition for a matrix $\mathbf{\Delta}$ to be a tree metric. While an evolutionary distance metric over a phylogeny $\mathcal{P}$ gives rise to a tree metric, the converse is not always true, i.e., not every tree metric corresponds to an evolutionary distance. Specifically, we describe the LogDet metric (Steel 1994b; Lockhart *et al.* 1994) as being the most important example.

**Theorem 4.2.** (Steel 1994b.) *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model with leaf set $L \subset V$. Recall from Lemma 3.7 that $\mathbf{J}_{uv}$ denotes the joint probability matrix for arbitrary nodes $u, v \in V$. Define the $|L \times |L|$ matrix $\mathbf{\Delta}_{\mathrm{LD}}$ with rows and columns indexed by the leaves as*

$$\mathbf{\Delta}_{\mathrm{LD}}[u,v] = \begin{cases} -\ln|\det \mathbf{J}_{uv}| & \text{if } \det \mathbf{J}_{uv} \neq 0; \\ \infty & \text{if } \det \mathbf{J}_{uv} = 0. \end{cases} \tag{4.12}$$

*The matrix $\mathbf{\Delta}_{\mathrm{LD}}$ is a tree metric known as the* LogDet metric*.*

Proof. We show that $\mathbf{\Delta}_{\mathrm{LD}}$ defines a tree metric over $\Psi(\mathcal{P}) = (V', E')$ by explicitly calculating the edge weights $d$ required by Definition 4.1. By Lemma 3.7, $\mathbf{\Delta}_{\mathrm{LD}}$ can be written in terms of the paralinear distance $D_{\mathrm{L}}$ as

$$\begin{aligned} \mathbf{\Delta}_{\mathrm{LD}}[u,v] &= -\ln|\det \mathbf{J}_{uv}| \\ &= -\ln \frac{|\det \mathbf{J}_{uv}|}{\left(\prod_{i\in\mathcal{A}} \pi_i^{(u)}\right)\left(\prod_{i\in\mathcal{A}} \pi_i^{(v)}\right)} - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u)} - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(v)} \\ &= D_{\mathrm{L}}(u,v) - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u)} - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(v)}. \end{aligned}$$

Define the edge weights $d\colon E' \mapsto [0, \infty]$ as follows. For every edge $uv \in E'$,

$$d(uv) = \begin{cases} D_{\mathrm{L}}(u,v) - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u)} & \text{if } u \text{ is a leaf;} \\ D_{\mathrm{L}}(u,v) & \text{otherwise.} \end{cases}$$

Obviously, $d(uv) \geq D_{\mathrm{L}}(u,v) \geq 0$ for every edge $uv$. Let $u'$ and $v'$ be two arbitrary leaves of $\Psi(\mathcal{P})$. Let

$$u' = u_0, e_1, u_1, e_2, \ldots, e_l, u_l = v'$$

be the path between them. By Theorem 3.6, $D_{\mathrm{L}}$ is additive along this path and thus

$$\begin{aligned}\sum_{k=1}^{l} d(e_k) =& \Big(D_{\mathrm{L}}(u', u_1) - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u')}\Big) + D_{\mathrm{L}}(u_1, u_2) + \cdots \\ &+ D_{\mathrm{L}}(u_{l-2}, u_{l-1}) + \Big(D_{\mathrm{L}}(u_{l-1}, v') - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(v')}\Big) \\ =& D_{\mathrm{L}}(u', v') - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(u')} - \frac{1}{2}\sum_{i\in\mathcal{A}} \ln \pi_i^{(v')} = \mathbf{\Delta}_{\mathrm{LD}}[u'v'].\end{aligned}$$

Consequently, $\Psi(\mathcal{P})$ with edge weights $d$ satisfies Definition 4.1 and thus $\mathbf{\Delta}_{\mathrm{LD}}$ is a tree metric. ∎

While $\mathbf{\Delta}_{\mathrm{LD}}$ can be viewed in light of Equation (4.12) as a function over distributions on sequence pairs, it is not an evolutionary distance according to our definition. For example, it does not satisfy Condition (O) of Fact 3.1, i.e., if $\mathbb{P}\Big\{\xi^{(u)} = \xi^{(v)}\Big\} = 1$, then $\mathbf{\Delta}_{\mathrm{LD}}[u, v]$ is not zero. Furthermore, it does not satisfy additivity (Condition (A)) since for three nodes $u, v, w$ on a path,

$$\Big(-\ln|\det \mathbf{J}_{uv}|\Big) + \Big(-\ln|\det \mathbf{J}_{vw}|\Big) \neq -\ln|\det \mathbf{J}_{uw}|$$

in general (also pointed out by e.g., Gu and Li 1996). Nevertheless, $\mathbf{\Delta}_{\mathrm{LD}}$ can be estimated from the sample sequences associated with the leaves, and the resulting matrix can be the input of a distance-based algorithm.

**Definition 4.2.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model with leaf set $L \subset V$. Let $u$ and $v$ be two nodes associated with random taxon sequences $X^{(u)}$ and $X^{(v)}$. Let $\ell = \Big|X^{(u)}\Big| = \Big|X^{(v)}\Big|$. The $m \times m$* empirical joint probability matrix *$\hat{\mathbf{J}}_{uv}$ is defined by its entries as*

$$\hat{\mathbf{J}}_{uv}[i, j] = \frac{1}{\ell}\sum_{k=1}^{\ell} \mathbb{I}\Big\{X_k^{(u)} = i, X_k^{(v)} = j\Big\}. \tag{4.13}$$

Definition 4.2 suggests a convenient way to estimate $\mathbf{\Delta}_{\mathrm{LD}}$ as $\hat{\mathbf{\Delta}}_{\mathrm{LD}}[u, v] = -\ln|\det \hat{\mathbf{J}}_{uv}|$. In the following we study the convergence of $\hat{\mathbf{J}}_{uv}$ to $\mathbf{J}_{uv}$.

**Lemma 4.3.** *Let $\hat{\mathbf{J}}_{uv}$ be the empirical joint probability distribution matrix of Definition 4.2. Then*

$$\mathbb{E}\det\hat{\mathbf{J}}_{uv} = \Big(1-\frac{1}{\ell}\Big)\Big(1-\frac{2}{\ell}\Big)\cdots\Big(1-\frac{m-1}{\ell}\Big)\det\mathbf{J}_{uv}. \tag{4.14}$$

Proof. By definition of the determinant,

$$\det\hat{\mathbf{J}}_{uv} = \sum_{j_1,\dots,j_m}{}' (-1)^{\kappa(j_1,\dots,j_m)}\prod_{i=1}^{m}\hat{\mathbf{J}}_{uv}[i,j_i],$$

where $\sum'$ denotes the sum over permutations and $\kappa(\cdot)$ equals $\pm 1$ depending on whether the number of switched pairs in the permutation is odd or even. Since the vector $\Big\langle \ell\hat{\mathbf{J}}_{uv}[i,j]\colon i,j=1,\dots,m\Big\rangle$ is multinomially distributed, Equation (3.23g) applies and thus

$$\begin{aligned}\mathbb{E}\det\hat{\mathbf{J}}_{uv} &= \sum_{j_1,\dots,j_m}{}' (-1)^{\kappa(j_1,\dots,j_m)}\prod_{i=1}^{m}\hat{\mathbf{J}}_{uv}[i,j_i]\\ &= \sum_{j_1,\dots,j_m}{}' (-1)^{\kappa(j_1,\dots,j_m)}\frac{\ell(\ell-1)\cdots(\ell-m+1)}{\ell^m}\prod_{i=1}^{m}\mathbf{J}_{uv}[i,j_i]\\ &=\Big(1-\frac{1}{\ell}\Big)\Big(1-\frac{2}{\ell}\Big)\cdots\Big(1-\frac{m-1}{\ell}\Big)\det\mathbf{J}_{uv},\end{aligned}$$

which is tantamount to Equation (4.14). ∎

**Definition 4.3.** *Let $\mathcal{P}=(V,E,\mathbb{P})$ be an evolutionary tree in the i. i. d. Markov model with leaf set $L$. The $|L|\times|L|$* empirical LogDet metric *is defined by its entries using the empirical joint probability matrix as*

$$\hat{\mathbf{\Delta}}_{\mathrm{LD}}[uv] = \begin{cases} -\ln|\det\hat{\mathbf{J}}_{uv}| & \text{if } \det\hat{\mathbf{J}}_{uv}\neq 0;\\ \infty & \text{if } \det\hat{\mathbf{J}}_{uv}=0.\end{cases}$$

*The $|L|\times|L|$* bias-corrected LogDet metric *is defined by its entries as*

$$\tilde{\mathbf{\Delta}}_{\mathrm{LD}}[u,v] = \hat{\mathbf{\Delta}}_{\mathrm{LD}}[u,v] + \sum_{k=1}^{m-1}\ln\Big(1-\frac{k}{\ell}\Big).$$

**Theorem 4.4.** *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model. For all leaves $u$ and $v$, and sample length $\ell$, the following hold. Define $\gamma_{\ell,m} = (1-\frac{1}{\ell})(1-\frac{2}{\ell})\cdots(1-\frac{m-1}{\ell})$,*

$$\hat{d}_{uv} = \det \hat{\mathbf{J}}_{uv}, \tag{4.15a}$$

$$\tilde{d}_{uv} = \frac{\det \hat{\mathbf{J}}_{uv}}{\gamma_{\ell,m}}. \tag{4.15b}$$

*For every $\epsilon > 0$,*

$$\mathbb{P}\left\{\frac{\tilde{d}_{uv}}{\det \mathbf{J}_{uv}} \leq 1-\epsilon\right\} \leq \exp\left(-\frac{\gamma_{\ell,m}^2 (m-1)^{2(m-1)}}{2} \ell \det{}^2 \mathbf{J}_{uv} \epsilon^2\right); \tag{4.16a}$$

$$\mathbb{P}\left\{\frac{\tilde{d}_{uv}}{\det \mathbf{J}_{uv}} \geq 1+\epsilon\right\} \leq \exp\left(-\frac{\gamma_{\ell,m}^2 (m-1)^{2(m-1)}}{2} \ell \det{}^2 \mathbf{J}_{uv} \epsilon^2\right); \tag{4.16b}$$

$$\mathbb{P}\left\{\left|\frac{\hat{d}_{uv}}{\det \mathbf{J}_{uv}} - 1\right| \geq \epsilon\right\} \quad \leq \quad 2m^2 \exp\left(-\frac{(m-1)^{2(m-1)}}{2m^4} \ell \det{}^2 \mathbf{J}_{uv} \epsilon^2\right). \tag{4.17}$$

We need the next lemma for the proof of Theorem 4.4.

**Lemma 4.5.** *Let $\mathbf{J}_1$ and $\mathbf{J}_2$ be two $m \times m$ matrices with non-negative entries such that $\sum_{i=1}^m \sum_{j=1}^m \mathbf{J}_k[i,j] = 1$ for $k = 1, 2$. Then*

$$\left|\det \mathbf{J}_1 - \det \mathbf{J}_2\right| \leq \frac{2m^2}{(m-1)^{m-1}} \mathcal{L}_\infty(\mathbf{J}_1, \mathbf{J}_2), \tag{4.18a}$$

*with $\mathcal{L}_\infty(\mathbf{J}_1, \mathbf{J}_2) = \max_{i,j}\left|\mathbf{J}_1[i,j] - \mathbf{J}_2[i,j]\right|$. Furthermore, if $\mathbf{J}_1$ and $\mathbf{J}_2$ differ in two entries only, i.e., if there exist $\epsilon > 0$, $i_1, j_1, i_2, j_2 \in \{1, \ldots, m\}$ such that*

$$\mathbf{J}_2[i,j] = \begin{cases} \mathbf{J}_1[i,j] - \epsilon & \text{if } i = i_1 \text{ and } j = j_1; \\ \mathbf{J}_1[i,j] + \epsilon & \text{if } i = i_2 \text{ and } j = j_2; \\ \mathbf{J}_1[i,j] & \text{otherwise,} \end{cases}$$

*then*

$$\left|\det \mathbf{J}_1 - \det \mathbf{J}_2\right| \leq 2\epsilon(m-1)^{-(m-1)}. \tag{4.18b}$$

We leave the proof of the lemma to the end of this chapter.

PROOF OF THEOREM 4.4. We prove first Equations (4.16a) and (4.16b) by using McDiarmid's inequality (Theorem 3.17). Define the i. i. d. random vector variables $\boldsymbol{\eta}_k = \left\langle X_k^{(u)}, X_k^{(v)} \right\rangle$ for $k = 1, \dots, \ell$. Then $\tilde{d}_{uv} = f(\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_\ell)$ where the function $f$ is defined by Equation (4.13) and (4.15b). By Equations (4.15b) and (4.18b), $f(\mathbf{x}_1, \dots, \mathbf{x}_\ell)$ changes by at most

$$\frac{2}{\ell\gamma_{\ell,m}}(m-1)^{-(m-1)}$$

if one of the $\mathbf{x}_k$ values is altered. Hence Equations (4.16a) and (4.16b) follow from McDiarmid's inequality applied to the function $f$.

By Equation (4.18a), if

$$\max_{i,j}\left|\hat{\mathbf{J}}_{uv}[i,j] - \mathbf{J}_{uv}[i,j]\right| < \frac{\epsilon|\det \mathbf{J}_{uv}|}{2m^2(m-1)^{-(m-1)}},$$

then $\left|\det \hat{\mathbf{J}}_{uv} - \det \mathbf{J}_{uv}\right| < \epsilon|\det \mathbf{J}_{uv}|$. Thus,

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{\hat{d}_{uv}}{\det \mathbf{J}_{uv}} - 1\right| \geq \epsilon\right\} &= \mathbb{P}\left\{\left|\det \hat{\mathbf{J}}_{uv} - \det \mathbf{J}_{uv}\right| \geq \epsilon|\det \mathbf{J}_{uv}|\right\} \\
&\leq \sum_{i,j}\mathbb{P}\left\{\left|\hat{\mathbf{J}}_{uv}[i,j] - \mathbf{J}_{uv}[i,j]\right| \geq \frac{\epsilon|\det \mathbf{J}_{uv}|}{2m^2(m-1)^{-(m-1)}}\right\} \\
&\leq 2m^2 \exp\left(-\frac{(m-1)^{2(m-1)}}{2m^4}\ell \det^2 \mathbf{J}_{uv}\epsilon^2\right),
\end{aligned}$$

proving Equation (4.17). ■

### 4.4.3 Numerical taxonomy

By Definition 4.1, an unrooted tree $\mathcal{T} = (V, E)$ with leaf set $L \subset V$ and edge weights $d\colon E \mapsto [0, \infty]$ gives rise to a tree metric $\boldsymbol{\Delta}_{\mathcal{T},d}$. The rows and columns of $\boldsymbol{\Delta}_{\mathcal{T},d}$ are indexed with the leaves. Each entry is the sum of edge weights on the path between the leaf of the row index and the leaf of the column index. The problem of numerical taxonomy is that of finding an optimal $\boldsymbol{\Delta}_{\mathcal{T},d}$ that is "closest" to an estimated input matrix $\hat{\boldsymbol{\Delta}}$. Evident choices for measuring the "closeness" of two matrices are the $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_\infty$ distances, defined as

$$\mathcal{L}_\alpha(\mathbf{\Delta}_1, \mathbf{\Delta}_2) = \begin{cases} \left(\sum_{i,j} \left|\mathbf{\Delta}_1[i,j] - \mathbf{\Delta}_2[i,j]\right|^\alpha\right)^{1/\alpha} & \text{if } \alpha \neq \infty; \\ \max_{i,j} \left|\mathbf{\Delta}_1[i,d] - \mathbf{\Delta}_2[i,j]\right| & \text{if } \alpha = \infty. \end{cases}$$

For a formal definition, let

$$\mathcal{F}\colon [0,\infty]^{|L|^2} \times [0,\infty]^{|L|^2} \mapsto [0,\infty]$$

be a lack-of-fit function on pairs of $|L| \times |L|$ matrices. The *numerical taxonomy problem* is the minimization of $\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*)$ for an input matrix $\hat{\mathbf{\Delta}}$ where $\mathbf{\Delta}^*$ is a tree metric sought. This problem was stated explicitly by Cavalli-Sforza and Edwards (1967a). The numerical taxonomy problem is the cornerstone of a whole school of phylogeny reconstruction, known as phenetics (Sokal and Sneath 1963; Sneath and Sokal 1973). Cavalli-Sforza and Edwards (1967a) seek a tree metric $\mathbf{\Delta}^*$ that minimizes the $\mathcal{L}_2$ distance between the matrices. In particular, they seek to minimize

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \sum_{i,j} \left(\hat{\mathbf{\Delta}}[i,j] - \mathbf{\Delta}^*[i,j]\right)^2.$$

The $\mathcal{L}_2$ problem also frequently arises in mathematical psychology (Cunningham 1978; de Soete 1983; Hubert and Arabie 1995; Smith 1998). Another classic approach in molecular systematics is that of Fitch and Margoliash (1967). They seek to minimize the "percent standard deviation" by using

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \sum_{i,j} \left(\frac{\hat{\mathbf{\Delta}}[i,j] - \mathbf{\Delta}^*[i,j]}{\hat{\mathbf{\Delta}}[i,j]}\right)^2.$$

Farris (1972) is usually credited with first proposing the use of the $\mathcal{L}_1$ metric for measuring the lack-of-fit. As our final example, Waterman *et al.* (1977) propose the minimization of the function

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \sum_{i,j} \left(\frac{\hat{\mathbf{\Delta}}[i,j] - \mathbf{\Delta}^*[i,j]}{(\hat{\mathbf{\Delta}}[i,j])^2}\right)^2.$$

The topology reconstruction method of Fitch and Margoliash uses exhaustive search among possible topologies and essentially consists of a procedure calculating the edge weights, i.e., returning a topology $\Psi^*$ that minimizes

$$\min_{d} \mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\Psi^*,d}).$$

Bulmer (1991) generalizes the approach by introducing an algorithm to select optimal edge weights for a given topology when $\mathcal{F}$ is a generalized least-squares function defined as

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \sum_{i,j,k,l} f_{i,j,k,l}\Big(\hat{\mathbf{\Delta}}[i,j] - \mathbf{\Delta}^*[i,j]\Big)\Big(\hat{\mathbf{\Delta}}[k,l] - \mathbf{\Delta}^*[k,l]\Big).$$

Bulmer also uses exhaustive search. These exhaustive search methods are feasible only for trees with up to 10–20 leaves. In fact, when $\mathcal{F}$ is the $\mathcal{L}_1$ or $\mathcal{L}_2$ metric, then the optimization problem is NP-hard (Křivánek and Morávek 1986; Day 1987). The problem is also NP-hard for the $\mathcal{L}_\infty$ metric, i.e., if

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \max_{i,j}\Big|\hat{\mathbf{\Delta}}[i,j] - \mathbf{\Delta}^*[i,j]\Big|,$$

shown by Agarwala *et al.* (1999). Agarwala *et al.* also show that the approximation version is difficult in that finding a tree metric $\mathbf{\Delta}^*$ with

$$\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*\right) < \frac{9}{8} \min_{\mathcal{T},d} \mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right)$$

is NP-hard. On the other hand, they also describe an algorithm, called Single Pivot, which finds a tree metric in $O\left(|L|^3\right)$ time such that

$$\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*\right) < 3 \min_{\mathcal{T},d} \mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right).$$

Cohen and Farach (1997) has introduced the closely related Double Pivot algorithm, which runs in $O\left(|L|^4\right)$ time.

### 4.4.4 Minimum evolution

Numerical taxonomy is concerned with finding the tree that best fits a given estimated distance matrix $\hat{\mathbf{\Delta}}$ according to some lack-of-fit measure. Another approach to phylogeny reconstruction is that of *minimum evolution* defined as follows. For an unrooted tree $\mathcal{T} = (V, E)$ with edge weights $d\colon E \mapsto [0, \infty]$ define the sum of edge weights score as

$$\mathrm{ME}(\mathcal{T}, d) = \sum_{e \in E} d(e).$$

$\mathrm{ME}(\mathcal{T}, d)$ is also referred to as *tree length*. Let $\mathcal{F}$ be a lack-of-fit function over $|L| \times |L|$ matrices, typically chosen to be the $\mathcal{L}_2$ distance. Minimum evolution methods aim at finding the topology $\Psi^*$ for a given input matrix $\hat{\mathbf{\Delta}}$ that minimizes

$$\mathrm{ME}(\Psi^*, d^*) \quad \text{with} \quad d^* = \arg\min_d \mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\Psi^*, d^*}). \tag{4.19}$$

Notice the common theme of the optimality criteria in both minimum evolution and numerical taxonomy: the edge weights are fitted for a given topology, and the fitting is evaluated by using a penalty function. The edge weights are fitted by minimizing a lack-of-fit function on the arising tree metric. The penalty function is the same lack-of-fit function in the case of numerical taxonomy. In the case of minimum evolution, the penalty function is the sum of edge weights. The principle of minimum evolution can be traced back to Kidd and Sgaramella-Zonta (1972) and has been extensively studied by Rzhetsky and Nei (1992b, 1992a, 1993). In particular, Rzhetsky and Nei (1993) prove that if the distance estimates are unbiased, i.e., if $\mathbb{E}\left[\hat{\mathbf{\Delta}}\right]$ equals the true tree metric, and $\mathcal{F}$ is the $\mathcal{L}_2$ distance, then the expected value of $\mathrm{ME}(\Psi^*, d^*)$ is minimal among possible topologies. In addition, they also describe an $O\left(|L|^3\right)$ time algorithm to solve the least-squares optimization for calculating the edge weights $d^*$. Gascuel (1997b) describes an $O\left(|L|^2\right)$ algorithm for the optimization crediting Vach and Degens (1991) with introducing the first algorithm to have the same asymptotical running time. Bryant and Waddell (2000) also describe an $O\left(|L|^2\right)$ algorithm for the same problem, along with an $O\left(|L|^3\right)$ time algorithm for weighted least-squares fitting of the edge weights as proposed by Fitch and Margoliash (1967), and an $O\left(|L|^4\right)$ algorithm for generalized least-squares edge weight fitting as proposed by Bulmer (1991).

As with other general optimization criteria, exhaustive search among phylogenies to minimize $\mathrm{ME}(\cdot)$ is feasible only for trees with up to 10–20 leaves. Its NP-hardness has not been proven, although Day (1983a) proves that if the lack-of-fitness is a simple function forcing the tree metric to have entries at least as large as the input matrix, i.e.,

$$\mathcal{F}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}^*) = \begin{cases} 0 & \text{if } \hat{\mathbf{\Delta}}[i,j] \leq \mathbf{\Delta}^*[i,j] \text{ for all } i,j; \\ \infty & \text{otherwise,} \end{cases}$$

then the minimization of ME is NP-hard.

Saitou and Nei (1987) introduced their Neighbor-Joining algorithm as a heuristic method to approximate the optimal tree albeit without pertaining theoretical guarantees. Studier and Keppler (1988) described a version of Neighbor-Joining running in $O\left(|L|^3\right)$ time, which was subsequently proven to be equivalent to the original method (Gascuel 1994). Neighbor-Joining today is the most popular distance-based algorithm. Its statistical efficiency has been studied in many simulated experiments by Saitou, Nei, and their colleagues (e.g., Sourdis and Nei 1988; Saitou and Imanishi 1989; Jin and Nei 1991; Nei *et al.* 1998), generally finding that it is one of the most successful topology reconstruction algorithms available. Other minimum evolution related heuristics include the algorithm of Rzhetsky and Nei (1992a), the stepwise search algorithm of Kumar (1996), Gascuel's BioNJ and UNJ (1997a, 1997b, 2000). and the Weighbor algorithm of Bruno *et al.* (2000).

### 4.4.5 Statistical efficiency of distance-based algorithms

Despite the large number of evolutionary tree reconstruction algorithms produced in the last thirty-some years, theoretical investigations concerning the algorithms' efficiency are fairly recent. A seminal paper in this direction is that of Farach and Kannan at the STOC conference of 1996. However, unlike in our framework, they interpret the problem of evolutionary tree reconstruction as that of estimating the distribution of random taxon sequences.

**Definition 4.4.** *Let $\mathcal{P}_0 = (V_0, E_0, \mathbb{P}_0)$ and $\mathcal{P}_1 = (V_1, E_1, \mathbb{P}_1)$ be two evolutionary trees in the i. i. d. Markov model on the same set of leaves*

$$L \subseteq V_0 \cap V_1.$$

*Let $|L| = n$. For every vector $\mathbf{x} \in \mathcal{A}^{|L|}$, define $p_0(\mathbf{x})$ as the probability that a random leaf labeling according to the distribution $\mathbb{P}_0$ produces $\mathbf{x}$. Define $p_1(\mathbf{x})$ similarly, as the probability that a random leaf labeling according to the distribution $\mathbb{P}_1$ produces $\mathbf{x}$. The* variational distance *$\mathcal{V}$ between $\mathcal{P}_0$ and $\mathcal{P}_1$ is* *defined as*

$$\mathcal{V}(\mathcal{P}_0, \mathcal{P}_1) = \sum_{\mathbf{x} \in \mathcal{A}^{|L|}} \Big| p_0(\mathbf{x}) - p_1(\mathbf{x}) \Big|.$$

Let $\mathcal{C}$ be a subclass of the i. i. d. Markov model and let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the class. The random taxon distribution learning problem is that of deriving an evolutionary tree $\mathcal{P}^*$ from sample sequences associated with the leaves of $\mathcal{P}$ such that $\mathcal{V}(\mathcal{P}, \mathcal{P}^*) \leq \epsilon$ for a given $\epsilon$. Farach and Kannan (1999) prove that if $\mathcal{C}$ is the class of phylogenies in the Jukes-Cantor model over a binary alphabet, then for sample lengths $\ell \gg \ln n$,

$$\mathcal{V}(\mathcal{P}, \mathcal{P}^*) \leq O\left( \frac{n\sqrt{\log n}}{\min_{u,v \in V} \Big| S_{JC}(u,v) \Big| \sqrt{\ell}} \right)$$

with $1 - o\,(1)$ probability where $\mathcal{P}^*$ is the tree produced by the Single Pivot algorithm of Agarwala *et al.* (1999). Ambainis *et al.* (1997) have extended this result to to the i. i. d. Markov model with constant substitution rates and time-reversibility by showing that the variational distance is bounded by

$$\mathcal{V}(\mathcal{P}, \mathcal{P}^*) \leq O\left( \frac{n\sqrt{\log n}}{\min_{u,v \in V} \Big| S_{\mathrm{L}}(u,v) \Big| \sqrt{\ell}} \right)$$

with high probability when the Single Pivot algorithm is used with the paralinear distance. The sample length bounds of Farach and Kannan (1999) and Ambainis *et al.* (1997) depend on the smallest similarity in the tree. An important theoretical contribution has been made by Cryan *et al.* (1998) describing a computationally efficient algorithm for the i. i. d. Markov model over a binary alphabet, which for all $\epsilon, \delta > 0$, builds an evolutionary tree $\mathcal{P}^*$ such that

$$\mathcal{V}(\mathcal{P}, \mathcal{P}^*) < \epsilon$$

with probability at least $(1 - \delta)$ from samples of a length polynomial in $n$, $(1/\epsilon)$, and $(1/\delta)$.

The study of sample complexity in conjunction with topology recovery was initiated by Steel *et al.* (1996) followed by a series of articles (Erdős *et al.* 1999a; Erdős *et al.* 1999b; Erdős *et al.* 1997). Analyzing the success conditions of certain distance-based algorithms, Erdős *et al.* (1999b) prove the following theorem in particular.

**Theorem 4.6.** (Erdős *et al.* 1999b.) *Let $\mathbf{\Delta}_{\mathcal{T},d}$ be an $|L| \times |L|$ tree metric for a topology $\mathcal{T} = (V, E)$ over the leaf set $L$ with edge weights $d$. Let $d_{\min} = \min_{e \in E} d(e)$. For an arbitrary $|L| \times |L|$ matrix calculated in the preprocessing step of a distance-based algorithm, the following statements hold.*

- *(a) A hypothetical exact algorithm for minimizing $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \cdot\right)$ returns $\mathcal{T}$ if $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right) < d_{\min}/4$.*
- *(b) The Single Pivot and Double Pivot algorithms (Agarwala* et al. *1999; Cohen and Farach 1997) return $\mathcal{T}$ if $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right) < d_{\min}/8$.*
- *(c) If $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right) < d_{\min}/2$, then $\hat{\mathbf{\Delta}}$ is a tree metric and $\mathcal{T}$ fits it, and thus algorithms using the relaxed four-point condition of Equation* (4.11) *to deduce quartet topologies and combining them into a tree (e.g., Buneman 1971) return $\mathcal{T}$.*

A similar result on success conditions for Neighbor-Joining and related algorithms is proven by Atteson (1997).

**Theorem 4.7.** (Atteson 1997.) *Let $\mathbf{\Delta}_{\mathcal{T},d}$ be an $|L| \times |L|$ tree metric for a topology $\mathcal{T} = (V, E)$ over the leaf set $L$ with edge weights $d$. Let $d_{\min} = \min_{e \in E} d(e)$. For an arbitrary $|L| \times |L|$ matrix calculated in the preprocessing step of a distance-based algorithm, if $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}}, \mathbf{\Delta}_{\mathcal{T},d}\right) < d_{\min}/2$, then the following statements hold.*

- *(a) ADDTREE (Sattath and Tversky 1977) returns $\mathcal{T}$.*
- *(b) Neighbor-Joining (Saitou and Nei 1987) returns $\mathcal{T}$.*
- *(c) BioNJ and UNJ (Gascuel 1997a; Gascuel 1997b) return $\mathcal{T}$.*

Erdős *et al.* (1999b) and Atteson (1997) analyze the convergence rate of the empirical Jukes-Cantor distance to obtain sample complexity bounds based on Theorems 4.6 and 4.7. Erdős *et al.* (1999b) also analyze the convergence rate of the LogDet metric estimation, although in a less exact way than Theorem 4.4 and only for the empirical LogDet metric. Theorems 4.6, 4.7, and our results on the convergence of empirical distances in §3.4, produce the following general theorem on the sample complexity of a number of distance-based algorithms.

| | ADDTREE, Neighbor-Joining, BioNJ,UNJ, Buneman | Single Pivot, Double Pivot |
|---|---|---|
| **JC** | $\beta_0 = 2\left(\frac{m}{m-1}\right)^2$ | $\beta_0 = 32\left(\frac{m}{m-1}\right)^2$ |
| **K3P** | $\beta_0 = 288$ | $\beta_0 = 4608$ |
| **Paralinear** | $\beta_0 = 4\frac{m^2(m-1)^2}{1-e^{-2}}\pi_{\min}\pi_{\text{span}}$ | $\beta_0 = 64\frac{m^2(m-1)^2}{1-e^{-2}}\pi_{\min}\pi_{\text{span}}$ |
| **Paralinear**, t.r.[1] | $\beta_0 = 4\frac{m^2(m-1)^2}{1-e^{-2}}$ | $\beta_0 = 64\frac{m^2(m-1)^2}{1-e^{-2}}$ |
| **LogDet**, empirical | $\beta_0 = \frac{8}{\gamma_{\ell,m}^2(m-1)^{2(m-1)}}$ | $\beta_0 = \frac{128}{\gamma_{\ell,m}^2(m-1)^{2(m-1)}}$ |
| **LogDet**, b.c.[2] | $\beta_0 = \frac{8m^4}{(m-1)^{2(m-1)}}$ | $\beta_0 = \frac{128m^4}{(m-1)^{2(m-1)}}$ |

| | |
|---|---|
| **JC** | $\beta_1 = 1$ |
| **K3P** | $\beta_1 = 6$ |
| **Paralinear** | $\beta_1 = 2m^2$ |
| **Paralinear**, t.r.[1] | $\beta_1 = 2m^2$ |
| **LogDet**, empirical | $\beta_1 = 1$ |
| **LogDet**, b.c.[2] | $\beta_1 = m^2$ |

[1] t.r. = time-reversible; [2] b.c. = bias-corrected

FIGURE 4.3: *Constants in the sample complexity bound of Equation* (4.20) *stating that for all the distances and algorithms considered,* $\ell(\delta) \leq \beta_0(2\ln n + \ln\frac{1}{\delta} + \ln\beta_1)/(S_{\min}^2 S_1^2)$. *Different constants are given for the paralinear distance in the i. i. d. Markov model and for the paralinear distance in the i. i. d. Markov model with time-reversibility. For the LogDet metric, different constants are shown depending on whether the empirical or the bias-corrected estimates of Definition 4.3 are used.*

**Theorem 4.8.** *Let $\mathcal{P}$ be a phylogeny in the i. i. d. Markov model with topology $\Psi = (V, E)$ and leaf set $L$ of size $|L| = n$. Let $\mathbf{\Delta}$ be a tree metric generated by $\Psi$ with edge weights $d\colon E \mapsto [0, \infty]$. When $\mathbf{\Delta}$ is the LogDet metric, or is derived from the Jukes-Cantor, Kimura's three parameter, or Lake's paralinear distance, then the following sample complexity bounds hold for the distance-based algorithms listed in Theorems 4.6 and 4.7. Let*

$$S_1 = 1 - \max_{e \in E} \exp\Big(-d(e)\Big)$$
$$S_{\min} = \min_{u,v \in L} \exp\Big(-\mathbf{\Delta}[u,v]\Big).$$

*For every confidence level $\delta > 0$, the sample complexities of the algorithms are bounded from above as*

$$\ell(\delta) \le \beta_0 \frac{2 \ln n + \ln \frac{1}{\delta} + \ln \beta_1}{S_{\min}^2 S_1^2}, \tag{4.20}$$

*with the constants $\beta_0, \beta_1$ shown in Figure 4.3.*

Proof. The proof relies on Lemma 3.11 for the Jukes-Cantor distance, on Lemma 3.18 for Kimura's three parameter distance, on Theorem 3.22 for the paralinear distance, and on Theorem 4.4 for the LogDet metric to calculate sample lengths ensuring the conditions of Theorems 4.6 and 4.7. Define the $|L| \times |L|$ matrices $\mathbf{S}$ and $\hat{\mathbf{S}}$ by $\mathbf{S}[u,v] = \exp(-\mathbf{\Delta}[u,v])$ and $\hat{\mathbf{S}}[u,v] = \exp(-\hat{\mathbf{\Delta}}[u,v])$ where $\hat{\mathbf{\Delta}}$ is the estimator for $\mathbf{\Delta}$ calculated in the preprocessing step of a distance-based algorithm. In order to simplify the discussion, we summarize the results concerning the convergence rates of empirical distances and the LogDet metric estimates by stating that for every estimator there exist constants $a, b > 0$ such that for all $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \frac{\hat{\mathbf{S}}[u,v]}{\mathbf{S}[u,v]} \le 1 - \epsilon \right\} &\le a \exp\Big(-b\ell(\mathbf{S}[u,v])^2 \epsilon^2\Big); \\ \mathbb{P}\left\{ \frac{\hat{\mathbf{S}}[u,v]}{\mathbf{S}[u,v]} \ge 1 + \epsilon \right\} &\le a \exp\Big(-b\ell(\mathbf{S}[u,v])^2 \epsilon^2\Big). \end{aligned} \tag{$*$}$$

Let us consider the Neighbor-Joining method. By Theorem 4.7, if each estimated entry $\hat{\mathbf{\Delta}}[u,v]$ is within $(\min_e d(e)/2)$ error from the entry $\mathbf{\Delta}[u,v]$, then Neighbor-Joining recovers the topology correctly. Obviously, for every $u, v \in L$, if $\min_e d(e) > 0$,

$$\mathbb{P}\left\{\left|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\right| \geq \frac{\min_e d(e)}{2}\right\} = \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \geq \frac{\min_e d(e)}{2}\right\} + \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \leq -\frac{\min_e d(e)}{2}\right\}.$$

By Equation $(*)$,

$$\begin{aligned} \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \geq \frac{\min_e d(e)}{2}\right\} & \\ &\leq a \exp\left(-b\ell\Big(\mathbf{S}[u,v]\Big)^2\Big(1 - \sqrt{1 - S_1}\Big)^2\right) \\ &\leq a \exp\left(-\frac{b}{4}\ell\Big(\mathbf{S}[u,v]\Big)^2 S_1^2\right). \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \leq -\frac{\min_e d(e)}{2}\right\} & \\ &\leq a \exp\left(-b\ell\Big(\mathbf{S}[u,v]\Big)^2\Big((1 - S_1)^{-1/2} - 1\Big)^2\right) \\ &\leq a \exp\left(-\frac{b}{4}\ell\Big(\mathbf{S}[u,v]\Big)^2 S_1^2\right). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}\left\{\left|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\right| \geq \frac{\min_e d(e)}{2}\right\} &\leq 2a \exp\left(-\frac{b}{4}\ell\Big(\mathbf{S}[u,v]\Big)^2 S_1^2\right) \\ &\leq 2a \exp\left(-\frac{b}{4}\ell S_{\min}^2 S_1^2\right). \end{aligned}$$

Hence

$$\mathbb{P}\left\{\max_{u,v}\left|\hat{\mathbf{\Delta}}[u,v]-\mathbf{\Delta}[u,v]\right|\geq\frac{\min_e d(e)}{2}\right\}$$
$$\leq\sum_{u,v}\mathbb{P}\left\{\left|\hat{\mathbf{\Delta}}[u,v]-\mathbf{\Delta}[u,v]\right|\geq\frac{\min_e d(e)}{2}\right\}\leq n^2 a\exp\left(-\frac{b}{4}\ell S_{\min}^2 S_1^2\right).$$

Therefore, if

$$\ell\geq 4\frac{2\ln n+\ln(1/\delta)+\ln a}{bS_{\min}^2 S_1^2}, \tag{$**$}$$

then

$$\mathbb{P}\left\{\max_{u,v}\left|\hat{\mathbf{\Delta}}[u,v]-\mathbf{\Delta}[u,v]\right|\geq\frac{\min_e d(e)}{2}\right\}\leq\frac{1}{\delta};$$

i.e., Neighbor-Joining is successful with at least $(1-\delta)$ probability. Furthermore, if $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}},\mathbf{\Delta}\right)<\min_e d(e)/2$, then ADDTREE, BioNJ, UNJ and Buneman's algorithm also succeed. Sample complexity for the Single Pivot and Double Pivot algorithms is bounded analogously, replacing the sample bound of Equation $(**)$ with

$$\ell\geq 64\frac{2\ln n+\ln(1/\delta)+\ln a}{bS_{\min}^2 S_1^2},$$

in order to attain $\mathcal{L}_\infty\left(\hat{\mathbf{\Delta}},\mathbf{\Delta}\right)<\min_e d(e)/8$ with probability at least $(1-\delta)$. The constants in Figure 4.3 are obtained by substituting the values for $a$ and $b$ with ones in our convergence rate results. ■

**Definition 4.5.** *Let $\mathcal{P}$ be a phylogeny in the i. i. d. Markov model with topology $\Psi=(V,E)$ and leaf set $L\subset V$. Let $\mathbf{\Delta}$ be a tree metric generated by $\Psi$ with positive finite edge lengths $d\colon E\mapsto(0,\infty)$. For every $\ell>0$, let $\hat{\mathbf{\Delta}}^{(\ell)}$ be an estimator for $\mathbf{\Delta}$ calculated from sample sequences of length $\ell$, associated with the leaves. For all $\ell>0$, $u,v\in L$, define*

$$S_{uv}=\exp\left(-\mathbf{\Delta}[u,v]\right);$$
$$\hat{S}_{uv}^{(\ell)}=\exp\left(-\hat{\mathbf{\Delta}}^{(\ell)}[u,v]\right).$$

*The estimator* $\left\{\hat{\mathbf{\Delta}}^{(\ell)}\right\}$ *is* $(a,b)$-regular *if and only if there exist* $a, b > 0$*, such that for every sample length* $\ell > 0$*, leaves* $u, v \in L$*, and* $\epsilon > 0$*,*

$$\mathbb{P}\left\{\frac{\hat{S}_{uv}^{(\ell)}}{S_{uv}} \leq 1 - \epsilon\right\} \leq a \exp\left(-b\ell S_{uv}^2 \epsilon^2\right);$$
$$\mathbb{P}\left\{\frac{\hat{S}_{uv}^{(\ell)}}{S_{uv}} \geq 1 + \epsilon\right\} \leq a \exp\left(-b\ell S_{uv}^2 \epsilon^2\right).$$

Remark. We omit the upper index $(\ell)$ on the sample length if it is obvious from the context. □

With Definition 4.5, Theorem 4.8 can be roughly stated as a result on the sample complexity of distance-based algorithms using $(a,b)$-regular tree metric estimators.

### 4.4.6 Sample complexity and tree radius

The sample complexity bounds of Theorem 4.8 are finite only if the tree metric has positive finite entries. For evolutionary distance metrics this corresponds to positive finite distances between tree nodes, which in turn implies that for the corresponding similarity $S$, $0 < |S(u,v)| < 1$ for all tree nodes $u, v$. Equivalently, there must exist $S_0, S_1 \in (0,1)$ such that for every edge $uv$,

$$S_0 \leq \Big|S(u,v)\Big| \leq 1 - S_1. \tag{4.21}$$

These bounds enter into our sample complexity results in Equation (4.20) if we recognize that $S_{\min} \geq S_0^k$ where $k$ is the maximum path length in the tree. The boundedness of distances and similarities on the edges is both necessary and meaningful. If $D(u,v) = \infty$ on an edge $uv$, then for every node pair $u', v'$ for which the path between $u'$ and $v'$ in $\Psi(\mathcal{P})$ includes the edge $uv$, $D(u',v') = \infty$. In that case the edge $uv$ splits the nodes into two sets $U$ and $V$ such that the path between every $u' \in U$ and every $v' \in V$ includes $uv$. The evolutionary relationships between nodes of $U$ and $V$ cannot then be deduced by any distance-based algorithm (see example in Figure 4.4). On the other hand, $D(u,v) = 0$ on an edge $uv$ also produces unresolvable relationships, as Figure 4.4 illustrates. Moreover, $D(u,v) = 0$ corresponds to $|\det \mathbf{M}_{uv}| = 1$ in the i. i. d. Markov model for all the evolutionary distance functions consid-

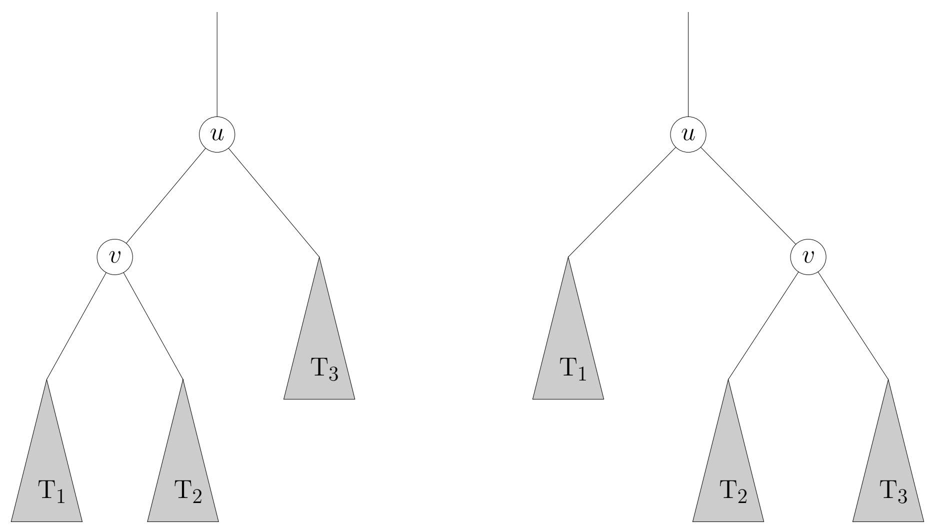



FIGURE 4.4: *If the distance $D(u, v) = 0$ on an edge $uv$, then the evolutionary relationships in the tree cannot be deduced unambiguously since the two topologies shown lead to the same pairwise distances between nodes. The ambiguity also arises if $D(u, v) = \infty$ allowing, among others, the same two topologies.*

ered, corresponding to the case in which $\mathbf{M}_{uv}$ is a permutation matrix, which is nonsense for DNA or protein sequences in a molecular biology context. Chang (1996) proves that when $\mathcal{P} = (V, E, \mathbb{P})$ is an evolutionary tree in the i. i. d. Markov model, the topology $\Psi(\mathcal{P})$ is determined by $\mathbb{P}$ if and only if for every edge $uv \in E$, $\det \mathbf{M}_{uv} \neq 0, \pm 1$.

Let us assume thus that the similarities are bounded as in Equation (4.21). Due to the multiplicativity of similarities, the similarity between two nodes may be exponentially small in the path length between them. In particular, if the tree is unbalanced, then $S_{\min}$ in Equation (4.20) becomes exponentially small in the number of leaves $n$. If there exist two leaves $u, v$ with path length $k$ between them in $\Psi(\mathcal{P})$, then $S_{\min} \geq S_0^k$. The lower bound is realized when all similarities equal $S_0$ between endpoints of the edges along the path. A distance-based algorithm that depends on the accurate estimation of such small similarities requires exponential sample lengths to recover the topology. In this light observe that Theorem 4.8 does not prove the statistical efficiency of any distance-based algorithm. For statistical efficiency it is important to make sure that only closely related nodes are used with large similarities, or equivalently, with small distances. The limits of that effort are captured via the following definition.

**Definition 4.6.** *The* outer radius $\varrho_{\text{out}}(\mathcal{P})$ *is defined as the smallest number such that any two leaves in* $\Psi(\mathcal{P})$ *are connected with a tree path containing at most* $2\varrho_{\text{out}}(\mathcal{P})$ *edges. The* inner radius $\varrho_{\text{in}}(\mathcal{P})$ *is defined as the smallest number such that for every edge* $e \in \Psi(\mathcal{P})$ *there is a path from each endpoint to a leaf with at most* $\varrho_{\text{in}}(\mathcal{P})$ *edges that does not go through* $e$.

Remark. The value $2\varrho_{\text{out}}(\mathcal{P})$ is commonly referred to as the diameter of the tree, but we choose this unusual way of presentation to emphasize the parallels and differences between inner and outer radii.

It is trivial that $2\varrho_{\text{out}}(\mathcal{P}) \geq \varrho_{\text{in}}(\mathcal{P})$. Moreover, while $\varrho_{\text{out}}(\mathcal{P})$ can be as large as $(n/2)$ for an evolutionary tree with $n$ leaves, $\varrho_{\text{in}}(\mathcal{P})$ is always logarithmic in $n$.

**Fact 4.9.** *The inner radius of an evolutionary tree* $\mathcal{P}$ *with* $n$ *leaves is bounded from above as*

$$\varrho_{\text{in}}(\mathcal{P}) \leq 1 + \lfloor \log_2(n-1) \rfloor.$$

Proof. The lemma is a result of the fact that the minimal topology for a fixed $\varrho_{\text{in}}$ is the one consisting of a single edge connected to the root of a full binary tree with $\varrho_{\text{in}}$ levels. ■

In fact, the inner and outer radii differ in their magnitudes even in average cases as analyzed by Erdős *et al.* (1999a). The distributions considered are uniform distributions of topologies and the Yule-Harding distribution (Harding 1971; Brown 1994). The Yule-Harding distribution arises in the following random tree construction mechanism. Let $L = \{u_1, u_2, \ldots, u_n\}$ be a set of leaves. Generate a random permutation $\langle u_{i_1}, u_{i_2}, \ldots, u_{i_n} \rangle$ of the leaves and connect $u_{i_1}$ to $u_{i_2}$ with a single edge. The remaining leaves are added to the tree in the order they appear in the permutation. At step $k$, leaf $u_{i_k}$ is connected to the tree built thus far by selecting a leaf uniformly from $u_{i_1}, \ldots, u_{i_{k-1}}$, and connecting $u_{i_k}$ to the incident edge.

**Theorem 4.10.** (Erdős *et al.* 1999a; Erdős *et al.* 1999b) *For an evolutionary tree* $\mathcal{P}$ *with* $n$ *leaves and random rooted topology selected with uniform probabilities,*

$$\varrho_{\text{in}}(\mathcal{P}) \leq (2 + o(1)) \log_2 \log_2(2n)$$

*with probability* $1 - o(1)$*, and for every* $\epsilon > 0$*,*

$$\varrho_{\text{out}}(\mathcal{P}) \geq \epsilon\sqrt{n}$$

*with probability* $1 - O(\epsilon^2)$*.*

*For an evolutionary tree* $\mathcal{P}$ *with* $n$ *leaves and a random topology selected under the Yule-Harding distribution,*

$$\varrho_{\text{in}}(\mathcal{P}) \leq (1 + o(1)) \log_2 \log_2(n)$$

*with probability* $1 - o(1)$*, and*

$$\varrho_{\text{out}}(\mathcal{P}) = \Omega(\log n)$$

*with probability* $1 - o(1)$*.*

Using Definition 4.6, the sample complexity bounds of Equation (4.20) can be rewritten as

$$\ell(\delta) \leq \beta_0 \frac{2\ln n + \ln(1/\delta) + \ln \beta_1}{S_0^{4\varrho_{\text{out}}(\mathcal{P})} S_1^2}.$$

The bound is exponential in the worst case. Moreover, it is exponential for almost all trees under the uniform topology distribution. Noting the lack of provable statistical efficiency of existing distance-based algorithms, Erdős *et al.* propose a family of statistically efficient algorithms, known as Short Quartet Methods (Erdős *et al.* 1998; Erdős *et al.* 1997; Erdős *et al.* 1999a; Erdős *et al.* 1999b). They show that sample complexity of their algorithms are bounded by

$$\ell(\delta) \leq \beta_0 \frac{2\ln n + \ln(1/\delta) + \ln \beta_1}{S_0^{4\varrho_{\text{in}}(\mathcal{P})+6} S_1^2}, \tag{4.22}$$

with some constants $\beta_0, \beta_1$ in the Jukes-Cantor model, and in the i. i. d. Markov model with the empirical LogDet metric. Using their proofs for these particular cases with our results on empirical distances, it can be shown that the same bounds are valid for Kimura's three parameter distance, Lake's paralinear distance and our bias-corrected LogDet metric.

The statistical efficiency of the Short Quartet Methods is matched with experimental success in simulations over caterpillar trees, in which each inner node has at least one leaf child, and thus $\varrho_{\text{out}} = n$ and $\varrho_{\text{in}} = 3$ (Erdős *et al.* 1997). On more balanced, biologically motivated trees, however, they perform poorly (Huson *et al.* 1999). In the next chapter we describe a family of statistically efficient algorithms with sample complexity similar to Equation (4.22). The theoretical efficiency is matched with high success rates in simulated experiments on theoretically interesting as well as biologically motivated large trees with high mutation probabilities.

## 4.A Technical proofs

Proof of Lemma 4.5. Assume that all entries of $\mathbf{J}_1$ and $\mathbf{J}_2$ are finite, since otherwise the lemma is trivial. We prove Equation (4.18b) first. Define the matrix $\mathbf{J}$ as follows.

$$\mathbf{J}[i,j] = \begin{cases} \mathbf{J}_1[i,j] + \epsilon & \text{if } i = i_2 \text{ and } j = j_2; \\ \mathbf{J}_1[i,j] & \text{otherwise.} \end{cases}$$

We claim that

$$|\det \mathbf{J} - \det \mathbf{J}_1| \leq \epsilon (m-1)^{-(m-1)} \tag{$*$}$$

$$|\det \mathbf{J} - \det \mathbf{J}_2| \leq \epsilon (m-1)^{-(m-1)}. \tag{$**$}$$

Expanding $\det \mathbf{J}$ and $\det \mathbf{J}_1$ by row $i_2$,

$$|\det \mathbf{J} - \det \mathbf{J}_1| \leq \epsilon |\det \mathbf{J}'|, \tag{$***$}$$

where $\mathbf{J}'$ is the matrix obtained from $\mathbf{J}_1$ by deleting row $i_2$ and column $j_2$. The matrix $\mathbf{J}'$ has only non-negative entries, and

$$\sum_{i,j} \mathbf{J}'[i,j] \leq 1.$$

Since the arithmetic mean bounds the geometric mean,

$$\begin{aligned}
|\det \mathbf{J}'| &\le \operatorname{perm} \mathbf{J}' \\
&\le \prod_{i=1}^{m-1} \left( \sum_{j=1}^{m-1} \mathbf{J}'[i,j] \right) \le \left( \frac{1}{m-1} \sum_{i=1}^{m-1} \left( \sum_{j=1}^{m-1} \mathbf{J}'[i,j] \right) \right)^{m-1} \\
&\le (m-1)^{-(m-1)}.
\end{aligned}$$

Hence Equation $(*)$ holds by Equation $(***)$. Equation $(**)$ is proven in the same manner.

By Equations $(*)$ and $(**)$,

$$\begin{aligned}
|\det \mathbf{J}_1 - \det \mathbf{J}_2| &\le |\det \mathbf{J} - \det \mathbf{J}_1| + |\det \mathbf{J} - \det \mathbf{J}_2| \\
&\le 2\epsilon (m-1)^{-(m-1)},
\end{aligned}$$

proving Equation (4.18b).

We derive Equation (4.18a) by repeated applications of Equation (4.18b). Define the series of matrices $\mathbf{J}_1 = \mathbf{J}'_0, \mathbf{J}'_1, \ldots$ recursively by the following algorithm. The matrix $\mathbf{J}'_{k+1}$ is derived from the matrix $\mathbf{J}'_k$ unless $\mathbf{J}'_k = \mathbf{J}_2$ using the procedure below.

1. Select an entry $[i,j]$ for which $\Big|\mathbf{J}'_k[i,j] - \mathbf{J}_2[i,j]\Big|$ is maximal.

2. If $\mathbf{J}'_k[i,j] > \mathbf{J}_2[i,j]$, then select an entry $[i',j']$ with $\mathbf{J}'_k[i',j'] < \mathbf{J}_2[i',j']$. Otherwise select an entry $[i',j']$ with $\mathbf{J}'_k[i',j'] > \mathbf{J}_2[i',j']$.

3. Let $\mathbf{J}'_{k+1} \leftarrow \mathbf{J}'_k$ and set

$$\begin{aligned}
\mathbf{J}'_{k+1}[i,j] &\leftarrow \mathbf{J}_2[i,j] \\
\mathbf{J}'_{k+1}[i',j'] &\leftarrow \mathbf{J}'_k[i',j'] + \Big(\mathbf{J}'_k[i,j] - \mathbf{J}_2[i,j]\Big).
\end{aligned}$$

Obviously,

$$\sum_{i,j} \mathbf{J}'_k[i,j] = \sum_{i,j} \mathbf{J}_1[i,j] = 1 = \sum_{i,j} \mathbf{J}_2[i,j].$$

Therefore, in each iteration step $k$, $\mathbf{J}'_k$ must differ from $\mathbf{J}_2$ in at least two entries. By symmetry, we may assume that $\mathbf{J}'_k[i,j] > \mathbf{J}_2[i,j]$ in the $k$-th iteration at Step (1). Then there must exist an entry $[i',j']$ for which

$\mathbf{J}'_k[i',j'] < \mathbf{J}_2[i',j']$ since the entries add up to one. Thus each iteration step can be carried out. Furthermore, the number of entries in which $\mathbf{J}'_k$ and $\mathbf{J}_2$ differ decreases by at least one in each iteration. Consequently the algorithm finishes in at most $m^2$ steps. (In fact, it finishes in at most $(m^2-1)$ steps since in the last iteration at least two entries are set equal to the corresponding entries in $\mathbf{J}_2$.) Since in each iteration step an entry with the maximum $\Big|\mathbf{J}'_k[i,j] - \mathbf{J}_2[i,j]\Big|$ is selected,

$$\max_{i,j}\Big|\mathbf{J}'_k[i,j] - \mathbf{J}_2[i,j]\Big| \le \max_{i,j}\Big|\mathbf{J}'_{k-1}[i,j] - \mathbf{J}_2[i,j]\Big|$$

for all $k > 0$. Hence by Equation (4.18b),

$$\begin{aligned}
|\det \mathbf{J}_1 - \det \mathbf{J}_2| &\le \sum_{k\ge 0} |\det \mathbf{J}'_{k+1} - \det \mathbf{J}'_k| \\
&\le \sum_{k\ge 0}\Big(2(m-1)^{-(m-1)} \max_{i,j}\Big|\mathbf{J}'_k[i,j] - \mathbf{J}_2[i,j]\Big|\Big) \\
&\le 2m^2(m-1)^{-(m-1)} \max_{i,j}\Big|\mathbf{J}_1[i,j] - \mathbf{J}_2[i,j]\Big|,
\end{aligned}$$

which is tantamount to Equation (4.18a). ■

# Chapter 5

# Harmonic Greedy Triplets

## 5.1 Introduction

Chapter 4 reviewed existing topology reconstruction algorithms with special emphasis on efficiency. In particular we pointed out that computational difficulties are frequently encountered in conjunction with algorithms based on optimization principles, such as maximum likelihood methods, character-based methods, and those distance-based methods operating on principles of numerical taxonomy or minimum evolution. Nei *et al.* (1998) and Gascuel (2000) argue along the same lines, also criticizing the lack of statistical efficiency of optimization algorithms based on performance in simulated experiments. This chapter introduces a family of algorithms designed with the goal of efficient topology recovery in mind. We avoid using penalty functions in the design process, and instead evaluate fundamental ideas of topology recovery by applying our results on the convergence speed of evolutionary distances. The algorithms use the simplest structures possible to reconstruct the topologies, namely, triplets of leaves. The rest of this section introduces basic notions and techniques for topology recovery from triplets. From §5.2 on, we focus on our novel algorithms.

### 5.1.1 Triplets

Let $\mathcal{T}$ be an unrooted tree, typically the topology of a phylogeny. A *triplet* $uvw$ consists of three distinct leaves $u$, $v$, and $w$ of $\mathcal{T}$. There is an inner node $o$ at which the pairwise paths between the leaves intersect, as shown in Figure 5.1. The node $o$ is called the *center* of $uvw$, and the triplet $uvw$ *defines*

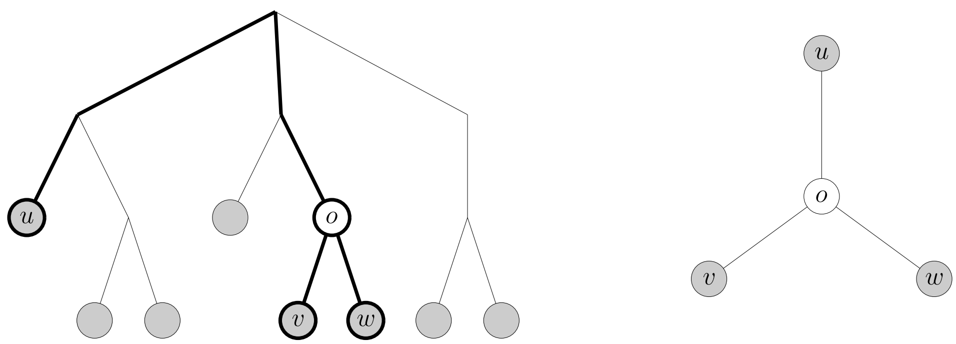



FIGURE 5.1: *The triplet uvw with its center o.*

the node $o$. Note that a star is formed by the edges on the paths between $o$ and the three leaves in $uvw$. Triplets of an evolutionary tree are the triplets of its topology. Chang (1996) proves that triplets are the simplest substructures identifying an evolutionary tree in the i. i. d. Markov model as stated by the next theorem.

**Theorem 5.1.** (CHANG 1996) *Let $\mathcal{P} = (V, E, \mathbb{P})$ be a phylogeny in the i. i. d. Markov model with leaf set $L \subset V$. Assume that each edge mutation matrix $\mathbf{M}_e$ belongs to a matrix class $\mathcal{M}$ which is reconstructible from rows (see §2.5.6), and* $\det \mathbf{M}_e \neq 0, \pm 1$. *Define the* triplet label distribution *for each triplet uvw as the joint distribution of $\left\langle \xi^{(u)}, \xi^{(v)}, \xi^{(w)} \right\rangle$. The set of triplet label distributions determine the topology $\Psi(\mathcal{P})$ and the distribution $\mathbb{P}$. The set of pairwise joint distributions $\left\{ \mathbb{P}_{uv} \colon u, v \in L \right\}$ does not necessarily determine $\mathbb{P}$.*

Early results of Smolenskiĭ (1962), Hakimi and Yau (1964), and Farris (1972) show that triplets can serve as basic building blocks to recover a tree $\mathcal{T}$ that fits a given tree metric $\mathbf{\Delta}$ with edge weights $d$. Let $\mathcal{T} = (V, E)$ be an unrooted tree with edge weights $d \colon E \mapsto [0, \infty]$ and denote the set of leaves by $L$. Let $\mathbf{\Delta}$ be the corresponding tree metric, i.e., the $|L| \times |L|$ matrix whose rows and columns are indexed by the leaves in such a way that for all $u, v \in L$, $\mathbf{\Delta}[u, v]$ is the sum of edge weights on the path between $u$ and $v$. For all nodes $u, v \in V$ define $\Delta_{uv}$ as the sum of edge weights between $u$ and $v$. Obviously, $\Delta_{uv} = \Delta_{vu}$. If $u$ and $v$ are leaves, then $\Delta_{uv} = \mathbf{\Delta}[u, v]$. By definition, for every triplet $uvw$ with center $o$,

$$\begin{aligned} \Delta_{uv} &= \mathbf{\Delta}[u, v] = \Delta_{uo} + \Delta_{vo}; \\ \Delta_{uw} &= \mathbf{\Delta}[u, w] = \Delta_{uo} + \Delta_{wo}; \\ \Delta_{vw} &= \mathbf{\Delta}[v, w] = \Delta_{vo} + \Delta_{wo}. \end{aligned}$$

Hence if $\mathbf{\Delta}[u, v]$, $\mathbf{\Delta}[u, w]$, and $\mathbf{\Delta}[v, w]$ are finite, then

$$\Delta_{uo} = \Delta_{ou} = \frac{\mathbf{\Delta}[u, v] + \mathbf{\Delta}[u, w] - \mathbf{\Delta}[v, w]}{2}; \tag{5.1a}$$

$$\Delta_{vo} = \Delta_{ov} = \frac{\mathbf{\Delta}[u, v] + \mathbf{\Delta}[v, w] - \mathbf{\Delta}[u, w]}{2}; \tag{5.1b}$$

$$\Delta_{wo} = \Delta_{ow} = \frac{\mathbf{\Delta}[u, w] + \mathbf{\Delta}[v, w] - \mathbf{\Delta}[u, v]}{2}. \tag{5.1c}$$

In particular, if $\mathcal{T}$ is the topology of a phylogeny $\mathcal{P}$ with distance metric $D$ and the evolutionary distances in the tree are finite, then

$$D(u, o) = \frac{D(u, v) + D(u, w) - D(v, w)}{2}; \tag{5.2a}$$

$$D(v, o) = \frac{D(u, v) + D(v, w) - D(u, w)}{2}; \tag{5.2b}$$

$$D(w, o) = \frac{D(u, w) + D(v, w) - D(u, v)}{2}. \tag{5.2c}$$

For brevity's sake, we use the notation

$$\mathrm{TC}(\mathbf{\Delta}, u, vw) = \frac{\mathbf{\Delta}[u, v] + \mathbf{\Delta}[u, w] - \mathbf{\Delta}[v, w]}{2}. \tag{5.3}$$

Notice that $\mathrm{TC}(\mathbf{\Delta}, u, vw) = \mathrm{TC}(\mathbf{\Delta}, u, wv)$. Equation (5.1) can be written with this notation as

$$\begin{aligned}\Delta_{uo} = \Delta_{ou} &= \mathrm{TC}(\mathbf{\Delta}, u, vw);\\ \Delta_{vo} = \Delta_{ov} &= \mathrm{TC}(\mathbf{\Delta}, v, uw);\\ \Delta_{wo} = \Delta_{ow} &= \mathrm{TC}(\mathbf{\Delta}, w, uv).\end{aligned}$$

Equation (5.1) shows how to calculate $\Delta_{z'z}$ if exactly one of $z, z'$ is a leaf. If both of them are inner nodes, then by taking a triplet $uv'w'$ with center $z'$ and a triplet $uvw$ with center $z$,

$$\Delta_{z'z} = \Bigl|\mathrm{TC}(\mathbf{\Delta}, u, v'w') - \mathrm{TC}(\mathbf{\Delta}, u, vw)\Bigr|.$$

This equation suggests a simple algorithm for recovering trees from tree metrics, or building topologies from distances between leaves. The algorithm is explicitly described by Waterman *et al.* (1977), and its basic idea can be traced back to Smolenskiĭ (1962) and Farris (1970, 1972). Figure 5.2 shows a very similar algorithm, called NAIVE-FIT-TREE. Let $L = \{u_1, \ldots, u_n\}$ be a set of leaves with $n \geq 2$, and let $\mathbf{\Delta}$ be a tree metric over $L$ with positive finite entries outside the diagonal. The NAIVE-FIT-TREE algorithm outputs a tree $\mathcal{T} = (V, E)$ with $L \subset V$ and a $|V| \times |V|$ matrix $\mathbf{\Delta}^*$ with rows and columns indexed by the nodes of $\mathcal{T}$. For each leaf pair $u, v \in L$, $\mathbf{\Delta}^*[u, v] = \mathbf{\Delta}[u, v]$ and for every edge $w'w \in E$, the entry $\mathbf{\Delta}^*[w', w]$ equals the edge weight. For all nodes $u, v$, $\mathbf{\Delta}^*[u, v]$ is the sum of edge weights along the path between $u$ and $v$. The algorithm adds leaves and inner nodes iteratively by applying Equation (5.1). After initializing $\mathcal{T}$ with two leaves, in each step $k = 3, \ldots, n$, the algorithm adds one leaf and one new inner node. Lines N1–N2 initialize $\mathcal{T}$ as a tree consisting of two nodes. Line N6 selects the edge $vw$ on which the center of the triplet $u_i u_j u_k$ falls. The center is on the path between $u_i$ and $u_j$ in $\mathcal{T}$, and it may be the node $v$ or a new inner node on edge $vw$. Line N7 tests whether it is a new node, and if so, it exits the loop over $i$ and $j$. Lines N9–N10 add the leaf $u_k$ and a new inner node $z$ to $\mathcal{T}$. Line N11 sets the weights of the newly created edges.

**Algorithm** Naive-Fit-Tree

**Input:** leaf set $L = \{u_1, \ldots, u_n\}$ with $n \geq 2$,
tree metric $\mathbf{\Delta}$ over $L$

**Output:** unrooted tree $\mathcal{T} = (V, E)$, matrix $\mathbf{\Delta}^*$.

N1 Set $V \leftarrow \{u_1, u_2\}$, $E \leftarrow \{u_1 u_2\}$.

N2 Set $\mathbf{\Delta}^*[u_1, u_2], \mathbf{\Delta}^*[u_2, u_1] \leftarrow \mathbf{\Delta}[u_1, u_2]$

N3 **for** each $k = 3, \ldots, n$ **do**

N4     **for** each $i, j \colon 1 \leq i < j < n$ **do**

N5         Set $d \leftarrow \mathrm{TC}(\mathbf{\Delta}, u_i, u_j u_k)$.

N6         Let $vw$ be the edge on the path between $u_i$ and $u_j$ such that $\mathbf{\Delta}^*[u_i, v] \leq d$ and $\mathbf{\Delta}^*[u_i, w] > d$.

N7         **if** $\mathbf{\Delta}^*[u_i, v] < d$ **then** exit from the **for** loop.

N8     Create new inner node $z$.

N9     Set $V \leftarrow V \cup \{z, u_k\}$.

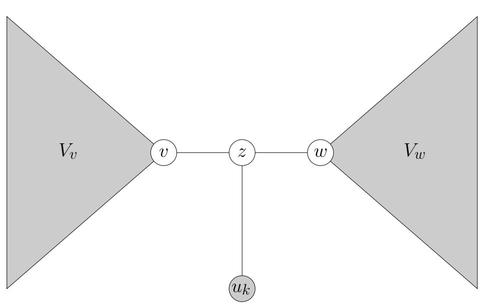
N10     Set $E \leftarrow (E - \{vw\}) \cup \{vz, wz, u_k z\}$.

N11     Set $\mathbf{\Delta}^*[v, z], \mathbf{\Delta}^*[z, v] \leftarrow d - \mathbf{\Delta}^*[u_i, v]$.
    Set $\mathbf{\Delta}^*[w, z], \mathbf{\Delta}^*[z, w] \leftarrow \mathbf{\Delta}^*[u_i, w] - d$.
    Set $\mathbf{\Delta}^*[u_k, z], \mathbf{\Delta}^*[z, u_k] \leftarrow \mathrm{TC}(\mathbf{\Delta}, u_k, u_i u_j)$.

N12     Update $\mathbf{\Delta}^*$ in rows and columns corresponding to $z$.

N13 Output $\mathcal{T}$ and $\mathbf{\Delta}^*$.

Figure 5.2: *The* Naive-Fit-Tree *algorithm that builds a tree $\mathcal{T}$ fitting a given tree metric $\mathbf{\Delta}$. The algorithm differs only in minor details from that of Waterman* et al. *(1977).*




The actions performed by Line N12 are the following. The nodes of $V$ are grouped as

$$V = V_v \cup V_w \cup \{z, u_k, v, w\}$$

so that for all nodes $v' \in V_v$, $w' \in V_w$, the path between $v'$ and $w'$ goes through $v$ and $w$.

For each node $v' \in V_v$, we set

$$\mathbf{\Delta}^*[v', z], \mathbf{\Delta}^*[z, v'] \leftarrow \mathbf{\Delta}^*[v', v] + \mathbf{\Delta}^*[v, z];$$
$$\mathbf{\Delta}^*[v', u_k], \mathbf{\Delta}^*[u_k, v'] \leftarrow \mathbf{\Delta}^*[v', z] + \mathbf{\Delta}^*[z, u_k].$$

For each node $w' \in V_w$, we set

$$\mathbf{\Delta}^*[w', z], \mathbf{\Delta}^*[z, w'] \leftarrow \mathbf{\Delta}^*[w', w] + \mathbf{\Delta}^*[u, z];$$
$$\mathbf{\Delta}^*[w', u_k], \mathbf{\Delta}^*[u_k, w'] \leftarrow \mathbf{\Delta}^*[w', z] + \mathbf{\Delta}^*[z, u_k].$$

An alternative solution to calculating all entries of $\mathbf{\Delta}^*$ is keeping track only of the edge weights and using an oracle-like procedure that computes other entries on request as the sum of edge weights on a path. This alternative solution is preferred by Waterman *et al.* (1977) but otherwise their algorithm is identical to NAIVE-FIT-TREE. It is not in our interest in the present study to prove the correctness of NAIVE-FIT-TREE; Waterman *et al.* (1977) provide the details.

### 5.1.2 Fitting a tree metric by using triplets

The NAIVE-FIT-TREE algorithm runs in $O(n^4)$ time since in each of the $(n-2)$ iterations of Line N3, the number of $(i, j)$ pairs inspected is $O(n^2)$, and for each pair, $O(n)$ edges are examined in Line N6. Waterman *et al.* (1977) do not discuss how to implement the algorithm more efficiently. Various possibilities exist to accelerate the algorithm. A simple way of reducing the time complexity is to restrict the set of $(i, j)$ pairs in each iteration to those with an arbitrarily fixed $i$. For example, with $i = 1$ across all iterations, the running time is $O(n^3)$. We choose a different course to attain $O(n^2)$ time complexity for an "offspring" of NAIVE-FIT-TREE. Other algorithms with $O(n^2)$ time complexity are discussed among others by Hein (1989), Culberson and Rudnicki (1989), Bandelt (1990), and Gusfield (1997). Our implementation stores $\mathcal{T}$ as a vector `T` in which each entry represents a node. The entries $\mathtt{T}[1], \ldots, \mathtt{T}[n]$ correspond to the leaves $u_1, \ldots, u_n$, respectively. The entries $\mathtt{T}[n+1], \ldots, \mathtt{T}[2n-2]$ correspond to the inner nodes. For simplicity we assume that $u_k = k$ for every leaf and the node set of $\mathcal{T}$ is $V = \{1, \ldots, 2n-2\}$. In this way for every node $u$, the corresponding entry is $\mathtt{T}[u]$. In order to

**Procedure** ADD-ON-EDGE

**Input:** tree structure T, non-root node $z$, new leaf $w$, new inner node $i$.

**Output:** none (updates T).

A1 Set $z' \leftarrow \mathtt{T}[z]$.parent.

A2 Set $\mathtt{T}[z]$.parent $\leftarrow i$ and $\mathtt{T}[w]$.parent $\leftarrow i$.

A3 Set $\mathtt{T}[w].\mathrm{left}, \mathtt{T}[w].\mathrm{right} \leftarrow \mathtt{null}$.

A4 **if** $z$ is left child $\triangleright$ *(i.e,* $\mathtt{T}[z'].\mathrm{left} = z$*)*
**then** set $\mathtt{T}[z']$.left $\leftarrow i$, $\mathtt{T}[i]$.left $\leftarrow z$, $\mathtt{T}[i]$.right $\leftarrow w$.
**else** set $\mathtt{T}[z']$.right $\leftarrow i$, $\mathtt{T}[i]$.right $\leftarrow z$, $\mathtt{T}[i]$.left $\leftarrow w$.

A5 Set $\mathtt{T}[i].\mathrm{added}, \mathtt{T}[z].\mathrm{added} \leftarrow \mathtt{true}$.

FIGURE 5.3: *The* ADD-ON-EDGE *procedure, which adds a new inner node $i$ on the edge between $z$ and its parent, and connects the leaf $w$ to $i$.*

follow which nodes are added, we maintain the attributes

$$\mathtt{T}[i].\mathrm{added} \in \{\mathtt{true}, \mathtt{false}\}$$

for $i = 1, 2, \ldots, 2n-2$ in the algorithm.

In order to test quickly whether an edge or inner node is on the path between two leaves, we store $\mathcal{T}$ rooted at $u_1$. For each entry $\mathtt{T}[i]$, the attributes

$$\mathtt{T}[i].\mathrm{parent}, \qquad \mathtt{T}[i].\mathrm{left}, \qquad \mathtt{T}[i].\mathrm{right}$$

define the parent-child relationships in $\mathcal{T}$ rooted at $u_1$. Notice that every edge $z'z$ of $\mathcal{T}$ can be specified by $z$ with $z' = \mathtt{T}[z]$.parent. The addition of a new inner node and a leaf is performed by the ADD-ON-EDGE procedure described in Figure 5.3. Let us introduce the notations

$$u \swarrow^{z}, \qquad {}^{z}\searrow v, \qquad \overset{w}{\underset{z}{\uparrow}},$$

meaning that in the rooted tree represented by T, node $u$ is in the left subtree of $z$, node $v$ is in the right subtree of $z$, and $w$ is not in the subtree rooted

**Procedure** SET-LENGTH

**Input:** tree structure `T`, non-root node $z$, new leaf $w$, new inner node $i$, edge lengths $\Delta_i^*$, $\Delta_z^*$, $\Delta_w^*$.

**Output:** none (updates `T`).

Set `T`$[z]$.length $\leftarrow \Delta_z^*$, `T`$[w]$.length $\leftarrow \Delta_w^*$, and `T`$[i]$.length $\leftarrow \Delta_i^*$.

FIGURE 5.4: *The* SET-LENGTH *procedure, which simply sets the new edge lengths after* ADD-ON-EDGE$(\mathtt{T}, z, w, i)$ *is called.*

at $z$. These relationships are preserved by the ADD-ON-EDGE procedure.

We omit tracking all entries of $\mathbf{\Delta}^*$ and store only the edge weights in an attribute

$$\mathtt{T}[i].\text{length}$$

that contains the weight of the edge between $i$ and its parent. The edge weights are updated by the SET-LENGTH procedure described in Figure 5.4, which is called every time a new inner node and a leaf are added by the ADD-ON-EDGE procedure.

Finally, we keep track of defining triplets for inner nodes used by the algorithm via an attribute `T`$[i]$.def comprising the fields

$$\mathtt{T}[i].\text{def.left}, \quad \mathtt{T}[i].\text{def.right}, \quad \mathtt{T}[i].\text{def.up}$$

for each inner node $i$. The fields of the attribute specify the directions in `T` so that if $u = \mathtt{T}[i].\text{def.left}$, $v = \mathtt{T}[i].\text{def.right}$, and $w = \mathtt{T}[i].\text{def.up}$, then

$$u \swarrow^{i}, \qquad {}^{i}\searrow v, \qquad \begin{array}{c} w \\ \uparrow \\ i \end{array}.$$

The proper upkeep of the `T`$[i]$.def attributes is ensured by the SET-DEFTRIP procedure described in Figure 5.5, which is called every time a new inner node and a leaf are added by the ADD-ON-EDGE PROCEDURE.

The FIT-TREE algorithm is described in Figure 5.7. Its running time is $O(n^2)$ and it uses $O(n)$ additional space to the one occupied by the tree metric $\mathbf{\Delta}$. Line T1 initializes the tree comprising two leaves. Each iteration on $w$ adds the new leaf $w$ and a new inner node $i$. The loop of Lines T3

**Procedure** SET-DEFTRIP

**Input:** tree structure `T`, non-root node $z$, triplet $uvw$, new inner node $i$.

**Output:** none (updates `T`).

D1 Set `T`$[i]$.def.up $\leftarrow u$.

D2 **if** $z$ is left child

D3 **then** set `T`$[i]$.def.left $\leftarrow v$, `T`$[i]$.def.right $\leftarrow w$;

D4 **else** set `T`$[i]$.def.right $\leftarrow w$, `T`$[i]$.def.left $\leftarrow v$.

FIGURE 5.5: *The* SET-DEFTRIP *procedure. After* ADD-ON-EDGE$(\mathtt{T}, z, w, i)$ *is called to add the new inner node $i$ on the edge between $z$ and its parent, this procedure sets the* .def *attribute for the node $i$. It is assumed that $u$ is a leaf of the unrooted tree represented by* `T` *with* $\overset{u}{\underset{z}{\uparrow}}$, *and $v$ is a leaf such that if $z$ is the left child of its parent, then* ${}_{v}\swarrow^{z}$, *otherwise* ${}^{z}\searrow_{v}$.

**Procedure** INIT-TREE

**Input:** tree structure `T`, tree metric $\mathbf{\Delta}$, leaves $u$, $v$

**Output:** none (updates `T`)

I1 **for** each $i = 1, \ldots, 2n-2$ **do** set `T`$[i]$.added $\leftarrow$ `false`.

I2 Set `T`$[u]$.parent $\leftarrow$ `null`, `T`$[u]$.left $\leftarrow v$, `T`$[u]$.right $\leftarrow$ `null`.

I3 Set `T`$[v]$.parent $\leftarrow u$, `T`$[v]$.left $\leftarrow$ `null`, `T`$[v]$.right $\leftarrow$ `null`.

I4 Set `T`$[v]$.length $\leftarrow \mathbf{\Delta}[u, v]$.

I5 Set `T`$[u]$.added, `T`$[v]$.added $\leftarrow$ `true`.

FIGURE 5.6: *The* INIT-TREE *procedure initializes* `T` *to contain two leaves $u$ and $v$ and sets the edge length between them to the corresponding entry in the tree metric* $\mathbf{\Delta}$.

**Algorithm** FIT-TREE
**Input:** tree metric $\boldsymbol{\Delta}$.
**Output:** tree structure `T`.

T1 INIT-TREE($\mathtt{T}, \boldsymbol{\Delta}, 1, 2$). ▷ *(`T` consists of nodes $\{1,2\}$.)*
T2 **for** each $w = 3, \ldots, n$ **do** ▷ *(iteration on the remaining leaves)*
T3     **for** each $z = 2, \ldots, 2n-2$ with $\mathtt{T}[z]$.added **do**
T4         Set $z' \leftarrow \mathtt{T}[z]$.parent.
T5         **if** $z'$ is the root ▷ *(i.e, $\mathtt{T}[z']$.parent = `null`)*
T6         **then** set $u \leftarrow z'$;
T7         **else** set $u \leftarrow \mathtt{T}[z']$.def.up.
T8         **if** $z$ is a leaf ▷ *(i.e, $\mathtt{T}[z]$.left = `null`)*
T9         **then** set $v \leftarrow z$, $d' \leftarrow 0$;
T10         **else** $v \leftarrow \mathtt{T}[z]$.def.left, $w' \leftarrow \mathtt{T}[z]$.def.right, $d' \leftarrow \mathrm{TC}(\boldsymbol{\Delta}, u, vw')$.
T11         Set $d \leftarrow \mathrm{TC}(\boldsymbol{\Delta}, u, vw)$.
T12         **if** $d < d'$ and $d' - d < \mathtt{T}[z]$.length
T13         **then** exit for loop. ▷ *($uvw$ defines new node on $z'z$)*
T14     Set $i \leftarrow n + w - 2$. ▷ *(index for new inner node)*
T15     ADD-ON-EDGE($\mathtt{T}, z, w, i$).
T16     SET-LENGTH($\mathtt{T}, zw, i, \mathtt{T}[z].\mathrm{length} - (d' - d), d' - d, \mathrm{TC}(\boldsymbol{\Delta}, w, uv)$).
T17     SET-DEFTRIP($\mathtt{T}, z, uvw, i$).
T18 Output `T`.

FIGURE 5.7: *The* FIT-TREE *algorithm builds a weighted tree represented in the tree data structure* `T` *from a tree metric* $\boldsymbol{\Delta}$.

finds an edge $z'z$ on which $i$ can be added. The notation is illustrated in Figure 5.8, which also illustrates why the test of Line T12 correctly identifies new inner nodes. Lines T14–T17 update `T`.

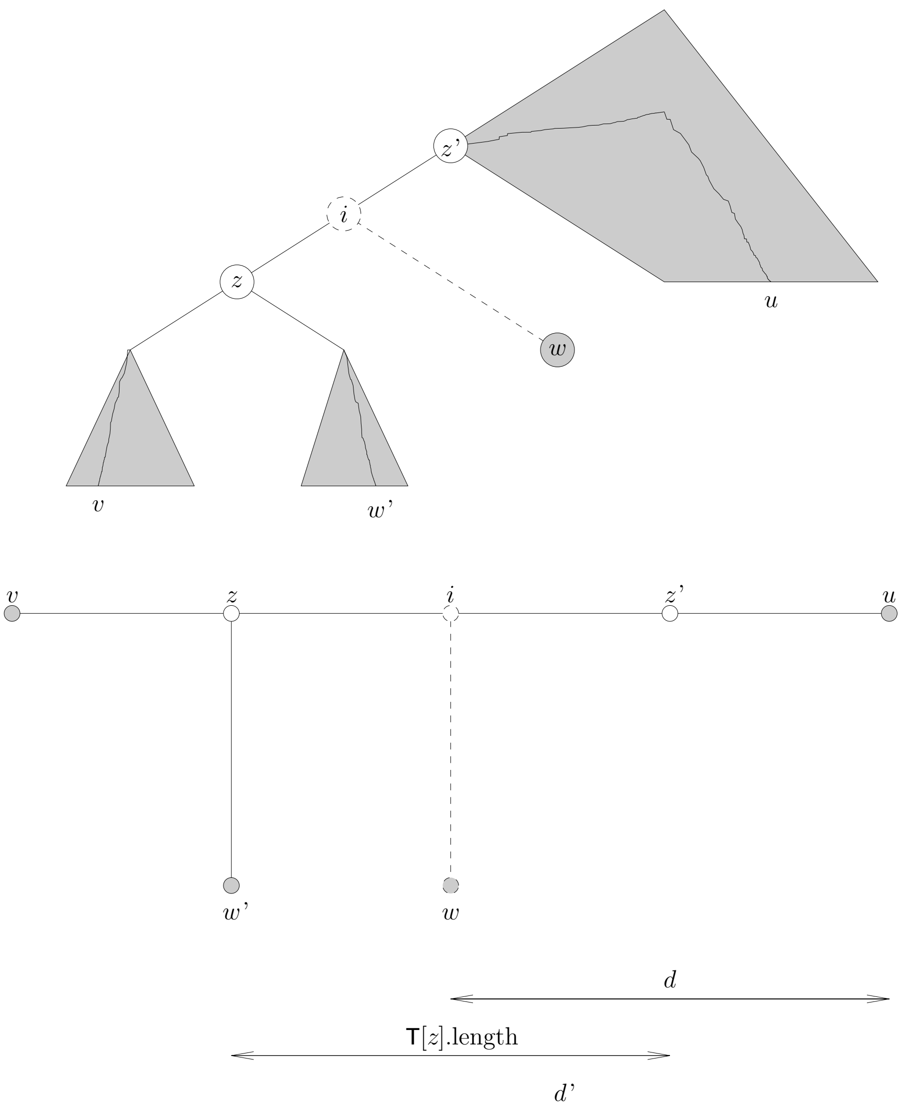



FIGURE 5.8: *Notation for Lines T4–T12 of the* FIT-TREE *procedure. The upper picture illustrates the selection of the leaves* $u$, $v$, $w$,*and* $w'$. *The lower picture shows the relationships between* $d$, $d'$, *and* $\mathtt{T}[z]$.length *in the test of Line T12.*

# 5.2 The Basic-HGT algorithm

## 5.2.1 Outline of Basic-HGT

In the rest of this chapter we describe a family of novel distance based algorithms, based on a principle, which we call that of "harmonic greedy triplets." We reported the application of this principle in (Csűrös and Kao 1999). The principle originates from the ideas leading to the FIT-TREE algorithm with two sets of differences. First, since we seek to recover the topology from an estimated tree metric, some precautions are taken in conjunction with determining triplet centers from estimated values using Equation (5.1). Secondly, the pool of candidate triplets for adding a new node on an edge is larger than the one used in the FIT-TREE algorithm. In order to fix the notation for the ensuing discussion, we first describe the specification for the Harmonic Greedy Triplets algorithms.

**Input.** The input to the algorithms is an estimated tree metric $\hat{\mathbf{\Delta}}$ over the leaf set $L = \{1, 2, \ldots, n\}$ with $n \geq 3$. The matrix $\hat{\mathbf{\Delta}}$ is an estimator for a tree metric $\mathbf{\Delta}$ generated by an evolutionary tree topology $\Psi = (V, E)$ with positive finite edge weights $d \colon E \mapsto (0, \infty)$.

**Output.** The output of the algorithms is a tree data structure `T` described in §5.1 representing an unrooted tree $\Psi^* = (V^*, E^*)$ with edge weights $d^* \colon E^* \mapsto (0, \infty)$.

**Success condition.** The algorithms are successful if $\Psi^*$ and $\Psi$ are topologically equivalent over the leaf set $L$, i.e., if

$$\Psi^* \underset{L}{\sim} \Psi.$$

As in §5.1.1, let $\Delta_{uv}$ denote the sum of edge weights on the path between arbitrary nodes $u, v \in V$. Similarly to Definition 4.5, define

$$S_{uv} = \exp\Bigl(-\Delta_{uv}\Bigr) \tag{5.4a}$$

for all nodes $u, v \in V$, and

$$\hat{S}_{uv} = \exp\Bigl(-\hat{\mathbf{\Delta}}[u,v]\Bigr) \tag{5.4b}$$

for all leaves $u, v \in L$. Our discussions related to the accuracy of $\hat{\mathbf{\Delta}}$ assume that it is an $(a,b)$-regular estimator for $\mathbf{\Delta}$ calculated from sample sequences of length $\ell$ (see Definition 4.5), i.e, that there exist $a, b > 0$ such that for all $\ell > 0$, leaves $u, v \in L$, and $\epsilon > 0$,

$$\mathbb{P}\biggl\{\frac{\hat{S}_{uv}}{S_{uv}} \leq 1-\epsilon\biggr\} \leq a\exp\Bigl(-b\ell S_{uv}^2\epsilon^2\Bigr); \tag{5.5a}$$

$$\mathbb{P}\biggl\{\frac{\hat{S}_{uv}}{S_{uv}} \geq 1+\epsilon\biggr\} \leq a\exp\Bigl(-b\ell S_{uv}^2\epsilon^2\Bigr). \tag{5.5b}$$

The formula

$$\mathrm{TC}(\mathbf{\Delta}, u, vw) = \frac{\mathbf{\Delta}[u,v] + \mathbf{\Delta}[u,w] - \mathbf{\Delta}[v,w]}{2}$$

is employed in the Fit-Tree algorithm to compare triplet centers. The Harmonic Greedy Triplets algorithms must use an estimated tree metric, and employ the formula $\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw)$. Consequently, we are interested in the difference between $\mathrm{TC}(\mathbf{\Delta}, u, vw)$ and $\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw)$. The next lemma provides a bound on that difference.

**Definition 5.1.** *For every triplet $uvw$, define the* average triplet size

$$\begin{aligned} S_{uvw} &= \frac{3}{e^{\mathbf{\Delta}[u,v]} + e^{\mathbf{\Delta}[u,w]} + e^{\mathbf{\Delta}[v,w]}} \\ &= \frac{3}{S_{uv}^{-1} + S_{uw}^{-1} + S_{vw}^{-1}}. \end{aligned} \tag{5.6}$$

**Lemma 5.2.** *If $\hat{\mathbf{\Delta}}$ is $(a,b)$-regular for some $a, b > 0$, then for every triplet $uvw$ and $0 < \epsilon < 1$,*

$$\mathbb{P}\biggl\{\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw) \geq \frac{-\ln(1-\epsilon)}{2}\biggr\} \leq 3a\exp\biggl(-\frac{b}{9}\ell S_{uvw}^2\epsilon^2\biggr). \tag{5.7}$$

Proof. See at end of chapter. ■

Remark. A similar bound can be obtained in an easier way without explaining the appearance of the harmonic average. Since

$$\Big|\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw)\Big| \\ \leq \frac{\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| + \Big|\hat{\mathbf{\Delta}}[u,w] - \mathbf{\Delta}[u,w]\Big| + \Big|\hat{\mathbf{\Delta}}[v,w] - \mathbf{\Delta}[v,w]\Big|}{2},$$

$$\begin{aligned} &\mathbb{P}\Big\{\Big|\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw)\Big| \geq \frac{-\ln(1-\epsilon)}{2}\Big\} \\ &\qquad \leq \mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| \geq \frac{-\ln(1-\epsilon)}{3}\Big\} \qquad (*) \\ &\qquad + \mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[u,w] - \mathbf{\Delta}[u,w]\Big| \geq \frac{-\ln(1-\epsilon)}{3}\Big\} \\ &\qquad + \mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[v,w] - \mathbf{\Delta}[v,w]\Big| \geq \frac{-\ln(1-\epsilon)}{3}\Big\}. \end{aligned}$$

Since $S_{uv} > 0$ for $u \neq v$,

$$S_{\min}(u,v,w) = \min\Big\{S_{uv}, S_{uw}, S_{vw}\Big\} \geq \frac{S_{uvw}}{3}.$$

Subsequently, using Equations $(*)$ and (5.5),

$$\begin{aligned} &\mathbb{P}\Big\{\Big|\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw)\Big| \geq \frac{-\ln(1-\epsilon)}{2}\Big\} \\ &\qquad \leq 6a \exp\Big(-\frac{b}{9}\ell\Big(S_{\min}(u,v,w)\Big)^2 \epsilon^2\Big) \qquad (5.8) \\ &\qquad \leq 6a \exp\Big(-\frac{b}{81}\ell S_{uvw}^2 \epsilon^2\Big) \quad \square \end{aligned}$$

The use of harmonic average in the average triplet size offers a number of inequalities involving similarities between triplet members. The next lemma shows a few of them, which we use in the analysis of the Harmonic Greedy Triplets algorithms.

**Lemma 5.3.** *Let $o$ be the center of triplet $uvw$. If $S_{uo} \leq S_{vo} \leq S_{wo}$, then $S_{uv} \leq S_{uw} \leq S_{vw}$, $S_{uw} \geq \frac{2}{3} S_{uvw}$, and $S_{vo}^2 \geq \frac{1}{3} S_{uvw}$.*

Proof. The lemma follows from the definition of $S_{uvw}$ and simple algebra. ■

Lemma 5.2 suggests that in order to minimize the error in calculating triplet centers, we should favor triplets with large average triplet scores. The *Harmonic Greedy Triplets principle* is that of employing a greedy selection based on the *estimated triplet score*

$$\begin{aligned}\hat{S}_{uvw} &= \frac{3}{e^{\hat{\mathbf{\Delta}}[u,v]} + e^{\hat{\mathbf{\Delta}}[u,w]} + e^{\hat{\mathbf{\Delta}}[v,w]}} \\ &= \frac{3}{\hat{S}_{uv}^{-1} + \hat{S}_{uw}^{-1} + \hat{S}_{vw}^{-1}}.\end{aligned}$$

During the course of building a hypothetical topology $\Psi^*$, the Harmonic Greedy Triplets algorithms maintain a set $\mathcal{R}$ of candidate triplet-edge pairs. Each element of $\mathcal{R}$ is a pair of $\langle uvw, e\rangle$ where $e$ is an edge in the partially built $\Psi^*$, and the triplet $uvw$ is such that $u, v \in V^*$, $w \notin V^*$, and the center of $uvw$ falls onto $e$. The algorithms have the following general outline.

1. Initialize $\Psi^*$ as a triplet $uvw$ with center $o$.
2. Initialize the candidate set $\mathcal{R}$.
3. **repeat**
4. Select a pair $\langle uvw, e\rangle$ from $\mathcal{R}$ with maximal $\hat{S}_{uvw}$.
5. Add a new inner node on $e$ connecting $w$ to $\Psi^*$.
6. Delete all pairs from $\mathcal{R}$ containing either $w$ or $e$.
7. Update $\mathcal{R}$ with respect to the newly created edges.
8. **until** all leaves are added.

The algorithms in the family differ mainly in their definitions of candidate pairs, i.e., in Lines 2, 6, and 7. The following definitions offer possible choices for candidate pairs.

**Definition 5.2.** *Let* T *be the tree data structure representing the hypothetical topology* $\Psi^*$ *during the execution of the algorithm. For each node* $u \in \Psi^*$*, define the set*

$$\mathrm{def}(u) = \begin{cases} \{u\} & \text{if } u \text{ is a leaf in } \Psi^*; \\ \Big\{\mathtt{T}[u].\mathrm{def.up}, \mathtt{T}[u].\mathrm{def.left}, \mathtt{T}[u].\mathrm{def.right}\Big\} & \text{if } u \text{ is an inner node.} \end{cases}$$

*A pair* $\langle uvw, z'z\rangle$ *is* relevant *if and only if the following hold.*

*(i)* $uvw$ *is triplet with* $u, v \in \Psi^*$, $w \notin \Psi^*$, *i.e,* $\mathtt{T}[u].\mathrm{added} = \mathtt{T}[v].\mathrm{added} = \mathtt{true}$, $\mathtt{T}[w].\mathrm{added} = \mathtt{false}$.

*(ii)* $z'z$ *is an edge in* $\Psi^*$ *with* $\mathtt{T}[z].\mathrm{parent} = z'$.

*(iii) The triplet* $uvw$ *shares leaves with the defining triplets for* $z$ *and* $z'$*, i.e,* $\mathrm{def}(z) \cap \{u, v\} \neq \emptyset$ *and* $\mathrm{def}(z') \cap \{u, v\} \neq \emptyset$.

*(iv) The edge* $z'z$ *lies on the path between* $u$ *and* $v$ *in* $\Psi^*$.

*A pair* $\langle uvw, z'z\rangle$ *is* strongly relevant *if and only if the following hold.*

*(i)* $\langle uvw, z'z\rangle$ *is a relevant pair.*

*(ii)* $u \in \mathrm{def}(z)$ *and* $v \in \mathrm{def}(z')$.

Calculating the relevant pairs for a given edge $e$ and new leaf $w \in \Psi^*$ is straightforward. Figure 5.9 shows how to calculate the set of strongly relevant pairs for fixed $e$ and $w$ in $O(1)$ time. The reason for restricting our attention to relevant or strongly relevant pairs is that we want to compare triplet centers only if the triplets share a leaf. If $z$ is an inner node of $\Psi^*$ with $\mathrm{def}(z) = \{u, v, w\}$, then we can test other triplets of the form $uv'w'$ and compare $\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw)$ to $\mathrm{TC}(\hat{\mathbf{\Delta}}, u, v'w')$ to recognize identical triplet centers. In order to compare the center of a triplet $uv'w'$ with the endpoints of the edge $z'z \in \Psi^*$, $\langle uv'w', z'z\rangle$ has to be a relevant pair.

### 5.2.2 Description of Basic-HGT

The HGT-Edge-Length procedure shown in Figure 5.10 calculates the edge lengths for inserting a new inner node on an edge $z'z$ as the center of a triplet $uvw$, where $\langle uvw, z'z\rangle$ is a relevant pair. The returned lengths also

| condition | | $u$ | $v$ |
|---|---|---|---|
| 0. | $z'$ is the root, $z$ is a leaf | $z'$ | $z$ |
| 1. | $z'$ is the root, $z$ is not leaf | $z'$ | $\mathtt{T}[z]$.def.left |
| | | $z'$ | $\mathtt{T}[z]$.def.right |
| 2. | $z'$ is not root, $z$ is a leaf | $\mathtt{T}[z']$.def.up | $z$ |
| 2a. | $\mathtt{T}[z']$.left $= z$ | $\mathtt{T}[z']$.def.left | $z$ |
| 2b. | $\mathtt{T}[z']$.right $= z$ | $\mathtt{T}[z']$.def.right | $z$ |
| 3. | $z'$ is not root, $z$ is not leaf | $\mathtt{T}[z']$.def.up | $\mathtt{T}[z]$.def.left |
| | | $\mathtt{T}[z']$.def.up | $\mathtt{T}[z]$.def.right |
| 3a. | $\mathtt{T}[z']$.left $= z$ | $\mathtt{T}[z']$.def.left | $\mathtt{T}[z]$.def.up |
| | | $\mathtt{T}[z']$.def.right | $\mathtt{T}[z]$.def.left |
| | | $\mathtt{T}[z']$.def.right | $\mathtt{T}[z]$.def.right |
| 3b. | $\mathtt{T}[z']$.right $= z$ | $\mathtt{T}[z']$.def.right | $\mathtt{T}[z]$.def.up |
| | | $\mathtt{T}[z']$.def.left | $\mathtt{T}[z]$.def.left |
| | | $\mathtt{T}[z']$.def.left | $\mathtt{T}[z]$.def.right |

FIGURE 5.9: *Rules for calculating the strongly relevant pairs $\langle uvw, z'z\rangle$ for an edge $z'z$ with $z = \mathtt{T}[z']$.parent and an arbitrarily fixed leaf $w \notin \Psi^*$. The table shows the possible choices for $u$ and $v$. For any edge, one of the cases 0, 1, 2, or 3 applies, with the appropriate subcase for 2 and 3, depending on whether $z$ is a left or a right child in* T.

tell whether the center $o$ of $uvw$ falls onto $z'z$. For example, if a negative $\Delta^*_{zo}$ is returned, then the center is in the subtree rooted at $z$ and does not fall onto $z'z$.

The BASIC-HGT algorithm employs a threshold $\Delta_{\min} \geq 0$ as an input parameter, which specifies the minimum distance between triplet centers to consider them separate. The HGT-SPLIT-EDGE procedure shown in Fig-

**Procedure** HGT-Edge-Length

**Input:** tree structure $\mathtt{T}$, estimated tree metric $\hat{\boldsymbol{\Delta}}$, non-root node $z$, triplet $uvw$.

**Output:** edge lengths $\Delta^*_{zo}$, $\Delta^*_{z'o}$

E1 Set $z' \leftarrow \mathtt{T}[z]$.parent.

E2 Set $x \in \{u, v\} \cap \mathrm{def}(z)$.
Set $x' \in \{u, v\} \cap \mathrm{def}(z')$.

E3 Set $\Delta^*_{xo} \leftarrow \mathrm{TC}(\hat{\boldsymbol{\Delta}}, x, \{u, v, w\} - \{x\})$.

E4 Set $\Delta^*_{x'o} \leftarrow \mathrm{TC}(\hat{\boldsymbol{\Delta}}, x', \{u, v, w\} - \{x'\})$.

E5 **if** $z'$ is the root **then** set $\Delta^*_{x'z'} \leftarrow 0$;
**else** set $\Delta^*_{x'z'} \leftarrow \mathrm{TC}(\hat{\boldsymbol{\Delta}}, x', \mathrm{def}(z') - \{x'\})$.

E6 **if** $z$ is a leaf **then** set $\Delta^*_{xz} \leftarrow 0$;
**else** set $\Delta^*_{xz} \leftarrow \mathrm{TC}(\hat{\boldsymbol{\Delta}}, x, \mathrm{def}(z) - \{x\})$.

E7 Set $d \leftarrow \Delta^*_{xo} - \Delta^*_{xz}$ and $d' \leftarrow \Delta^*_{x'o} - \Delta^*_{x'z'}$.

E8 **if** $z'$ is on the path between $z$ and $x$ **then** $d \leftarrow -d$.

E9 **if** $z$ is on the path between $z'$ and $x'$ **then** $d' \leftarrow -d'$.

E10 Set $\Delta^*_{zo} \leftarrow (d + \mathtt{T}[z].\mathrm{length} - d')/2$.
Set $\Delta^*_{z'o} \leftarrow (d' + \mathtt{T}[z].\mathrm{length} - d)/2$.

E11 **return** $\Delta^*_{zo}$ and $\Delta^*_{z'o}$.

Figure 5.10: *The* HGT-Edge-Length *procedure calculates the edge lengths for adding the center of the triplet $uvw$ on the edge between $z$ and its parent $z' = \mathtt{T}[z]$.parent. It is assumed that $\langle uvw, z'z\rangle$ is a relevant pair. The value $\Delta^*_{zo}$ is our estimation for $\Delta_{zo}$ for the center $o$ of $uvw$. The edge length $\Delta^*_{z'o}$ is the estimation for $\Delta_{z'o}$. The test of Line E8 consists of checking whether $z = x$ or $\mathtt{T}[z]$.def.up $= x$. The test of Line E9 consists of checking whether $z' = x'$, or $z$ is a left child and $\mathtt{T}[z']$.def.left $= x'$, or $z$ is a right child and $\mathtt{T}[z']$.def.right $= x'$.*

**Procedure** HGT-Split-Edge

**Input:** tree structure `T`, estimated tree metric $\hat{\mathbf{\Delta}}$, minimum distance parameter $\Delta_{\min}$, non-root node $z$, triplet $uvw$.

**Output:** `true` or `false`

$\triangleright$ *(whether the center of uvw falls onto the edge $z'z$)*

S1 Set $z' \leftarrow \mathtt{T}[z]$.parent.

S2 Set $x \in \{u,v\} \cap \operatorname{def}(z)$.
Set $x' \in \{u,v\} \cap \operatorname{def}(z')$.

S3 Set $\Delta^*_{xo} \leftarrow \mathrm{TC}(\hat{\mathbf{\Delta}}, x, \{u,v,w\} - \{x\})$.

S4 Set $\Delta^*_{x'o} \leftarrow \mathrm{TC}(\hat{\mathbf{\Delta}}, x', \{u,v,w\} - \{x'\})$.

S5 **if** $z'$ is the root **then** set $\Delta^*_{x'z'} \leftarrow 0$;
**else** set $\Delta^*_{x'z'} \leftarrow \mathrm{TC}(\hat{\mathbf{\Delta}}, x', \operatorname{def}(z') - \{x'\})$.

S6 **if** $z$ is a leaf **then** set $\Delta^*_{xz} \leftarrow 0$;
**else** set $\Delta^*_{xz} \leftarrow \mathrm{TC}(\hat{\mathbf{\Delta}}, x, \operatorname{def}(z) - \{x\})$.

S7 Set $d \leftarrow \Delta^*_{xo} - \Delta^*_{xz}$ and $d' \leftarrow \Delta^*_{x'o} - \Delta^*_{x'z'}$.

S8 **if** $|d| < \Delta_{\min}$ or $|d'| < \Delta_{\min}$ **then return** `false`.

S9 **if** $z'$ is on the path between $z$ and $x$ **then** $d_+ \leftarrow -d$ **else** $d_+ \leftarrow d$.
**if** $z$ is on the path between $z'$ and $x'$ **then** $d'_+ \leftarrow -d'$ **else** $d'_+ \leftarrow d'$.

S10 Set $\Delta^*_{zo} \leftarrow (d_+ + \mathtt{T}[z].\text{length} - d'_+)/2$.
Set $\Delta^*_{z'o} \leftarrow (d'_+ + \mathtt{T}[z].\text{length} - d_+)/2$.

S11 **if** $\Delta^*_{zo} \geq \mathtt{T}[z]$.length or $\Delta^*_{z'o} \geq \mathtt{T}[z]$.length **then return** `false`.

S12 **return** `true`.

Figure 5.11: *The* HGT-Split-Edge *procedure is analogous to the* HGT-Edge-Length *procedure. It determines whether the center of a triplet uvw falls onto the edge $z'z$ using a minimum distance parameter $\Delta_{\min}$. It is assumed that $\langle uvw, z'z\rangle$ is a relevant pair.*

ure 5.11 determines for a relevant pair $\langle uvw, z'z\rangle$ whether the center of $uvw$ falls onto the edge $z'z$ and whether it is at least $\Delta_{\min}$ away from both $z$ and $z'$. The procedure is almost identical to the HGT-Edge-Length pro-

**Procedure** HGT-INIT

**Input:** tree structure T, estimated tree metric $\hat{\mathbf{\Delta}}$, triplet $uvw$.

**Output:** none (updates T).

J1 INIT-TREE($\mathtt{T}, \hat{\mathbf{\Delta}}, uv$).

J2 Set $o \leftarrow n+1$. $\triangleright$ *(index for new inner node)*

J3 ADD-ON-EDGE($\mathtt{T}, v, w, o$).

J4 SET-LENGTH($\mathtt{T}, v, w, o, \mathrm{TC}(\hat{\mathbf{\Delta}}, v, uw), \mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw), \mathrm{TC}(\hat{\mathbf{\Delta}}, w, uv)$).

J5 SET-DEFTRIP($\mathtt{T}, v, uvw, o$).

FIGURE 5.12: *The* HGT-INIT *procedure initializes the tree structure* T *as the triplet* $uvw$ *with its center.*

cedure with the exception of the additional tests. A relevant pair $\langle uvw, z'z \rangle$ for which HGT-SPLIT-EDGE returns `true` is called a *splitting pair.*

The BASIC-HGT algorithm is detailed in Figure 5.13. Given $\Delta_{\min}$ and an estimated tree metric $\hat{\mathbf{\Delta}}$ derived from $n$ sample sequences the algorithm constructs a hypothetical topology $\Psi^*$ represented by the tree data structure T. The algorithm first constructs a star formed by a triplet and its center in Line B2. It then inserts a leaf and a corresponding inner node per iteration of the repeat at Line B5 until $\Psi^*$ has $n$ leaves. For $k = 3, \ldots, n$, let $\Psi_k^*$ be the version of $\Psi^*$ with $k$ leaves constructed during the run of the BASIC-HGT algorithm; i.e., $\Psi_3^*$ is constructed at Line B2, and $\Psi_k^*$ with $k \geq 4$ is constructed at Lines B9–B12 in the $(k-3)$-th iteration of the repeat. Note that $\Psi_n^*$ is output at Line B17.

**Lemma 5.4.** *For each* $k = 3, \ldots, n-1$*, the following statements hold at the start of the* $(k-2)$*-th iteration of the repeat at Line B5.*

*(i) For every edge* $z'z \in \Psi_k^*$, $\mathrm{def}(z') \cap \mathrm{def}(z) \neq \emptyset$.

*(ii) The set* $\mathcal{R}$ *consists of the splitting pairs for the edges in* $\Psi_k^*$.

PROOF. The statements are proved separately.

*Statement (i).* The proof is by induction on $k$. The base case follows from the fact that the statement holds for $\Psi_3^*$ constructed at Line B2. The induction step follows from the use of a relevant pair at Line B7.

**Algorithm** BASIC-HGT

**Input:** estimated tree metric $\hat{\mathbf{\Delta}}$, minimum distance parameter $\Delta_{\min}$.

**Output:** tree structure $\mathtt{T}$.

B1 Find a triplet $u_0v_0w_0$ with the maximum $\hat{S}_{u_0v_0w_0}$.

B2 HGT-INIT$(\mathtt{T}, \hat{\mathbf{\Delta}}, u_0v_0w_0)$.

B3 Set $o \leftarrow n+1$. $\triangleright$ *(index for first inner node)*

B4 Let $\mathcal{R}$ be the set of splitting pairs for $u_0o$, $ov_0$, $ow_0$.

B5 **repeat**

B6     **if** $\mathcal{R} = \emptyset$ **then fail** and **stop**.

B7     Find $\langle uvw, z'z \rangle \in \mathcal{R}$ with maximum $\hat{S}_{uvw}$.

B8     Set $o \leftarrow o+1$. $\triangleright$ *(index for new inner node)*

B9     ADD-ON-EDGE$(\mathtt{T}, z, w, o)$.

B10     Set $(\Delta^*_{zo}, \Delta^*_{z'o}) \leftarrow$ HGT-EDGE-LENGTH$(\mathtt{T}, \hat{\mathbf{\Delta}}, z, uvw)$.

B11     SET-LENGTH$(\mathtt{T}, z, w, o, \Delta^*_{zo}, \Delta^*_{z'o}, \mathrm{TC}(\hat{\mathbf{\Delta}}, w, uv))$.

B12     Set $\mathrm{def}(o) \leftarrow \{u, v, w\}$.

B13     Remove from $\mathcal{R}$ all pairs containing $z'z$ or $w$.

B14     Add to $\mathcal{R}$ the splitting pairs for the edges $z'o$, $oz$, $ow$.

B15     **for** each edge $x'x \neq z'o, oz, ow$ **do**
        add to $\mathcal{R}$ the splitting pairs of the form $u'ww'$.

B16 **until** all leaves are added.

B17 Output $\mathtt{T}$.

FIGURE 5.13: *The* BASIC-HGT *algorithm. Line B12 calls the* SET-DEFTRIP *procedure after determining whether* $\overset{o}{\underset{u}{\uparrow}}$ *or* $\overset{o}{\underset{v}{\uparrow}}$.

*Statement (ii).* The proof is by induction on $k$ based on the following facts. Let $x'x$ be an edge in $\Psi^*_k$ that also exists in $\Psi^*_{k+1}$, i.e, that is not selected at Line B7 at the $(k-2)$-th iteration of the repeat. Let $u'v'w'$ be a triplet such that $u'v' \in \Psi_k^*$ and $w' \notin \Psi_{k+1}^*$. Then at Line B13, $\langle u'v'w', x'x \rangle$ is a splitting pair in $\Psi_k^*$ if and only if it is also one in $\Psi_{k+1}^*$. Also, after a new leaf $w$ and new inner node $o$ are inserted on edge $z'z$ at Line B9, each edge $x'x \neq z'o, oz, ow$ in $\Psi_{k+1}^*$ may have new relevant pairs, which must be of the form $\langle u'ww', x'x \rangle$ with $w' \notin \Psi_{k+1}^*$. Such pairs are not relevant in $\Psi_k^*$ because two leaves ($w$ and $w'$) in the triplet are not yet added. ■

### 5.2.3 Time and space complexity

**Theorem 5.5.** *The* BASIC-HGT *algorithm runs in* $O(n^3)$ *time using* $O(n^2)$ *work space.*

PROOF. We analyze the time and space complexities separately as follows.

*Time complexity.* Line B1 takes $O(n^3)$ time. Line B4 takes $O(n)$ time to examine $3(n-3)$ pairs of edges and triplets. Since the repeat at line B5 iterates at most $n-3$ times, it suffices to implement $\mathcal{R}$ so that each iteration of the repeat takes $O(n^2)$ time. By the greedy policy at line B7, for each edge $z'z \in \Psi^*$ and each leaf $w \notin \Psi^*$, $\mathcal{R}$ needs to maintain only the splitting pair containing a triplet of the form $uvw$ with the maximum $\hat{S}_{uvw}$. We organize such pairs by implementing $\mathcal{R}$ as a two-dimensional array where the rows are indexed by $z'z$ and the columns by $z$. Then, Lines B6 and B7 take $O(n^2)$ time by traversing $\mathcal{R}$. Line B13 simply sets to null every entry in the row and column in $\mathcal{R}$ for $z'z$ and $w$, taking $O(n)$ time. Line B14 examines $O(n^2)$ triplets for each edge $z'o$, $oz$, $ow$ and updates $\mathcal{R}$ in $O(n^2)$ total time. At line B15, $w \notin \mathrm{def}(x') \cup \mathrm{def}(x)$. Consequently, $u' \in \mathrm{def}(x) \cap \mathrm{def}(x)$ for a relevant pair $\langle u'ww', x'x\rangle$. Thus, this line examines $O(n)$ triplets for each $x'x$ and updates $\mathcal{R}$ in $O(n^2)$ total time.

*Space complexity.* $\Psi^*$ stored in the tree data structure `T` takes $O(n)$ work space. The set $\mathcal{R}$ as a two-dimensional array takes up $O(n^2)$ space. The other variables needed by the algorithm occupy $O(1)$ space. Thus the total work space is as claimed. ■

### 5.2.4 Lemmas for bounding the sample size

Our statistical analysis involves a number of intermediate results culminating in Theorem 5.15 in §5.2.5 stating the statistical efficiency of the BASIC-HGT algorithm. The main idea is that the greedy selection favors triplets with large average size, and the centers of the triplets are estimated within a small error. Lemma 5.8 summarizes the analysis of the greedy selection. Our result concerning the error of triplet center estimation is stated in Lemma 5.9. Lemma 5.10 extends Lemmas 5.8 and 5.9 to the set of all triplets. Lemmas 5.11, 5.12, and 5.13 prove certain invariants of the algorithm.

The BASIC-HGT algorithm aims at recovering a topology $\Psi = (V, E)$ with positive edge weights $d\colon E \mapsto (0, \infty)$. By Equation 5.4a, for each edge $uv \in E$, $d(uv) = \Delta_{uv}$ and thus $S_{uv} = e^{-d(uv)}$. Since $d(uv) > 0$, $S_{uv} \in (0, 1)$. Moreover, since on every path $u_0, e_1, u_1, \ldots, e_l, u_l$ in $\Psi$, $\Delta_{u_0 u_l}$

equals the sum of edge weights along the path, $S_{u_0 u_l} = \prod_{k=1}^{l} S_{u_k u_{k-1}}$, and thus $S_{uv} \in (0,1)$ for all nodes $u, v \in V$ if $u \neq v$. Define

$$S_0 = \min\{S_{uv} \colon uv \in E\}; \tag{5.9a}$$

$$S_1 = 1 - \max\{S_{uv} \colon uv \in E\}. \tag{5.9b}$$

If nodes $w, w' \in V$ are connected via a path of length $l$, then

$$0 < S_0^l \leq S_{ww'} \leq (1 - S_1)^l < 1.$$

**Lemma 5.6.** *For any two nodes $u$ and $v$ of $\Psi$ with $S_{uv} < S$, there is a node $z$ on the path between $u$ and $v$ such that $SS_0^{1/2} \leq S_{uz} \leq SS_0^{-1/2}$.*

Proof. The lemma follows from the definition of $S_0$. ■

**Lemma 5.7.** *Let $\varrho_{\mathrm{in}}$ be the inner radius of $\Psi$. Every inner node $o \in V$ except the root has a defining triplet $uvw$ such that the path length between $u$ and $o$, $v$ and $o$, and $w$ and $o$ are not larger than $\varrho_{\mathrm{in}} + 1$, and thus*

$$S_{uvw} \geq S_0^{2(\varrho_{\mathrm{in}}+1)}.$$

*Every leaf of $\Psi$ is in such a triplet.*

Proof. The lemma is the direct consequence of the definition of inner radius. ■

The Basic-HGT algorithm strives to recover $\Psi$ by using triplets described in Lemma 5.7. Define

$$S_{\mathrm{lg}} = \frac{3\sqrt{2}}{2}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2 S_0^{2\varrho_{\mathrm{in}}+4} \qquad \left(\approx \frac{S_0^{2\varrho_{\mathrm{in}}+4}}{16}\right); \tag{5.10a}$$

$$S_{\mathrm{sm}} = \frac{S_{\mathrm{lg}}}{\sqrt{2}}; \tag{5.10b}$$

$$S_{\mathrm{md}} = \frac{S_{\mathrm{lg}} + S_{\mathrm{sm}}}{2}. \tag{5.10c}$$

A triplet $uvw$ is *large* if $S_{uvw} \geq S_{\mathrm{lg}}$; it is *small* if $S_{uvw} \leq S_{\mathrm{sm}}$. Note that by Lemma 5.7, each non-root inner node has a large defining triplet.

**Lemma 5.8.** *Assume that $\hat{\mathbf{\Delta}}$ is $(a,b)$-regular with $a,b>0$ calculated from sample sequences of length $\ell$. The first inequality holds for all large triplets $uvw$; the second inequality holds for all small triplets $uvw$.*

$$\mathbb{P}\Big\{\hat{S}_{uvw} \le S_{\mathrm{md}}\Big\} \le a\exp\bigg(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\bigg); \tag{5.11a}$$

$$\mathbb{P}\Big\{\hat{S}_{uvw} \ge S_{\mathrm{md}}\Big\} \le a\exp\bigg(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\bigg). \tag{5.11b}$$

Proof. See at the end of the chapter. ■

The input parameter $\Delta_{\min}$ in the Basic-HGT algorithm specifies the threshold on the edge lengths in $\Psi^*$. Since two triplet centers are considered separate by the algorithm if and only if $\Big|\mathrm{TC}(\hat{\mathbf{\Delta}},u,vw) - \mathrm{TC}(\hat{\mathbf{\Delta}},u,v'w')\Big| \ge \Delta_{\min}$, we impose

$$\Delta_{\min} \le \frac{\min_{e\in E} d(e)}{2} = \frac{-\ln(1-S_1)}{2}, \tag{5.12a}$$

and define

$$\vartheta = \frac{\Delta_{\min}}{\min_{e\in E} d(e)} = \frac{\Delta_{\min}}{-\ln(1-S_1)} \le \frac{1}{2}. \tag{5.12b}$$

The next lemma shows that a large triplet's center is estimated within a small error with high probability.

**Lemma 5.9.** *Assume that $\hat{\mathbf{\Delta}}$ is $(a,b)$-regular with $a,b>0$ calculated from sample sequences of length $\ell$. Let $uvw$ be a triplet that is not small, i.e.., $S_{uvw} > S_{\mathrm{sm}}$. Then*

$$\mathbb{P}\bigg\{\Big|\mathrm{TC}(\hat{\mathbf{\Delta}},u,vw) - \mathrm{TC}(\mathbf{\Delta},u,vw)\Big| \ge \frac{\Delta_{\min}}{2}\bigg\} \le 7a\exp\bigg(-b\frac{\vartheta^2}{162}\ell S_{\mathrm{lg}}^2 S_1^2\bigg). \tag{5.13}$$

Proof. See at the end of the chapter. ■

We next define and analyze two key events $\mathcal{E}_{\mathrm{g}}$ and $\mathcal{E}_{\mathrm{c}}$ as follows. The subscripts g and c denote the words "greedy" and "center", respectively.

- $\mathcal{E}_{\mathrm{g}}$ is the event that $\hat{S}_{uvw} > \hat{S}_{u'v'w'}$ for every large triplet $uvw$ and every small triplet $u'v'w'$.

- $\mathcal{E}_{\mathrm{c}}$ is the event that for every small triplet $uvw$ that is not small,

$$\left|\mathrm{TC}(\hat{\boldsymbol{\Delta}}, u, vw) - \mathrm{TC}(\boldsymbol{\Delta}, u, vw)\right| < \frac{\Delta_{\min}}{2};$$
$$\left|\mathrm{TC}(\hat{\boldsymbol{\Delta}}, v, uw) - \mathrm{TC}(\boldsymbol{\Delta}, v, uw)\right| < \frac{\Delta_{\min}}{2};$$
$$\left|\mathrm{TC}(\hat{\boldsymbol{\Delta}}, w, uv) - \mathrm{TC}(\boldsymbol{\Delta}, w, uv)\right| < \frac{\Delta_{\min}}{2}.$$

**Lemma 5.10.** *The probabilities of the complementary events to $\mathcal{E}_{\mathrm{g}}$ and $\mathcal{E}_{\mathrm{c}}$ are bounded from above as follows.*

$$\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{g}}\right\} \leq a\binom{n}{3}\exp\left(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\right); \tag{5.14a}$$
$$\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{c}}\right\} \leq 21a\binom{n}{3}\exp\left(-b\frac{\vartheta^2}{162}\ell S_{\mathrm{lg}}^2 S_1^2\right). \tag{5.14b}$$

Proof. Equation (5.14a). The event $\bar{\mathcal{E}}_{\mathrm{g}}$ implies that there is a large triplet $uvw$ with $\hat{S}_{uvw} \leq S_{\mathrm{md}}$ or that there is a small triplet $u'v'w'$ with $\hat{S}_{u'v'w'} \geq S_{\mathrm{md}}$. Thus, by Lemma 5.8,

$$\begin{aligned}\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{g}}\right\} &\leq \sum_{\text{large } uvw} \mathbb{P}\left\{\hat{S}_{uvw} \leq S_{\mathrm{md}}\right\} + \sum_{\text{small } u'v'w'} \mathbb{P}\left\{\hat{S}_{u'v'w'} \geq S_{\mathrm{md}}\right\} \\ &\leq a\binom{n}{3}\exp\left(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\right),\end{aligned}$$

which is tantamount to Equation (5.14a). In order to prove Equation (5.14b), we simply sum up Equation (5.13) three times for every triplet. ■

The Basic-HGT algorithm builds the series of hypothetical topologies $\Psi_3^* = (V_3^*, E_3^*), \ldots, \Psi_n^* = (V_n^*, E_n^*)$ such that $\Psi_n^*$ is the hypothetical reconstruction of the topology $\Psi = (V, E)$. Define $L_k \subset V_k^*$ as the leaf set of $\Psi_k^*$.

By the iterative addition of leaves and internal taxa,

$$
\begin{aligned}
L_3 \subset L_4 \subset \cdots \subset L_n, &\qquad \forall k > 3\colon \left|L_k - L_{k-1}\right| = 1;\\
V_3^* \subset V_4^* \subset \cdots \subset V_n^*, &\qquad \forall k > 3\colon \left|V_k^* - V_{k-1}^*\right| = 2;\\
\forall k > 3\colon \left|E_k^* - E_{k-1}^*\right| = 3, &\qquad \forall k > 3\colon \left|E_{k-1}^* - E_k^*\right| = 1.
\end{aligned}
$$

For every node $z \in V_k^*$, there is a corresponding node $f(z) \in V$, where the mapping $f$ is established by the defining triplets. If $z$ is a leaf, then $f(z) = z$; otherwise $f(z)$ equals the center of the triplet formed by $\operatorname{def}(z)$ in $\Psi$. Notice that the mapping does not change with $k$ in the sense that if $z \in V_{k'}^*, V_{k'+1}^*, \ldots, V_n^*$, then $f(z)$ is the same for all $k \geq k'$. In the following we omit the explicit referencing of this mapping for economy's sake.

Our proof for the sample length bounds of Basic-HGT essentially consists of showing that the following conditions for $k = 3, \ldots, n$ are implied by $\mathcal{E}_{\mathrm{g}}$ and $\mathcal{E}_{\mathrm{c}}$, and thus hold with high probability.

$\mathcal{X}_k$: The tree $\Psi_k^*$ is a topological minor of $\Psi$ over $L_k$, i.e., $\Psi_k^* \underset{L_k}{\pitchfork} \Psi$.

$\mathcal{Y}_k$: For every inner node $z \in V_k^*$, the triplet formed by $\operatorname{def}(z)$ is not small.

$\mathcal{Z}_k$: For every edge $z'z \in E_k^*$ with length $\Delta_{z'z}^*$, $\left|\Delta_{z'z}^* - \Delta_{z'z}\right| < 2\Delta_{\min}$, i.e., $\left|\mathtt{T}[z].\mathrm{length} - \Delta_{z'z}\right| < 2\Delta_{\min}$ where $z' = \mathtt{T}[z].\mathrm{parent}$ and $\Delta_{z'z}^* = \mathtt{T}[z].\mathrm{length}$.

Condition $\mathcal{X}_k$ states that $\Psi_k^*$ correctly represents the evolutionary relationships between its leaves. Condition $\mathcal{Y}_k$ states that $\Psi_k^*$ is built without using small triplets. Condition $\mathcal{Z}_k$ pronounces that the edge lengths in $\Psi_k^*$ are estimated within a small error.

The ensuing series of results stated by Lemmas 5.11, 5.12, and 5.13 prove that $\mathcal{E}_{\mathrm{g}}$ and $\mathcal{E}_{\mathrm{c}}$ imply $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$ for all $k = 3, \ldots, n$. The lemmas make the following assumptions for some $k < n$.

- The $(k-3)$-th iteration of the repeat loop at Line B5 has been completed.
- The tree $\Psi_k^*$ has been constructed, and the conditions $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$ hold.

- The BASIC-HGT algorithm is currently in the $(k-2)$-th iteration of the repeat.

**Lemma 5.11.** *Assume that the* HGT-SPLIT-EDGE *procedure is called with the relevant triplet $\langle uvw, z'z \rangle$. If $\mathcal{E}_c$ holds and $uvw$ is not small, then the test of Line S8 fails if and only if the center of $uvw$ is different from $z$ and $z'$.*

PROOF. Let $o$ be the center of $uvw$. From Line S7,

$$d = \Big(\Delta^*_{xo} - \Delta_{xo}\Big) - \Big(\Delta^*_{xz} - \Delta_{xz}\Big) + \Big(\Delta_{xo} - \Delta_{xz}\Big). \qquad (*)$$

If $z = o$ then $z$ is an inner node of $\Psi^*_k$. Since $\mathcal{E}_c$ holds, and neither $uvw$ nor $\mathrm{def}(z)$ is small,

$$\Big|\Delta^*_{xo} - \Delta_{xo}\Big| < \frac{\Delta_{\min}}{2}, \qquad \Big|\Delta^*_{xz} - \Delta_{xz}\Big| < \frac{\Delta_{\min}}{2}. \qquad (**)$$

Since $z = o$, $\Delta_{xo} = \Delta_{xz}$ and by Equations $(*)$ and $(**)$, $|d| < \Delta_{\min}/2 + \Delta_{\min}/2 + 0 < \Delta_{\min}$ and thus the test of Line S8 passes. Using a similar reasoning, if $z' = o$, then $|d'| < \Delta_{\min}$ and the test passes.

If $o \neq z, z'$, then $|\Delta_{xo} - \Delta_{xz}| \geq -\ln(1 - S_1) \geq 2\Delta_{\min}$ since the center of $o$ and $z$ are both on the path between $u$ and $v$ in $\Psi$. If $z$ is a leaf in $\Psi^*_k$, then $\Delta^*_{xz} = \Delta_{xz} = 0$. By $\mathcal{E}_c$ and Equation $(*)$, $|d| > \frac{3}{2}\Delta_{\min}$. If $z$ is an inner node in $\Psi^*_k$, then by $\mathcal{Y}_k$, $\mathcal{E}_c$, and Equation $(*)$, $|d| > \Delta_{\min}$. In either case, $|d| > \Delta_{\min}$. By symmetry, $|d'| > \Delta_{\min}$ also. Hence the test of Line S8 fails. ■

**Lemma 5.12.** *In addition to the assumptions of Lemma 5.11, also assume that $o \neq z, z'$, i.e, the test of Line S8 has failed. The test of Line S11 then fails if and only if the center of $uvw$ is on the path between $z$ and $z'$ in $\Psi$.*

PROOF. First, let us assume that the center of $uvw$ is on the path between $z$ and $z'$ in‘$\Psi$. From Lines S7 and S9, by $\mathcal{X}_k$,

$$\begin{aligned}(d_+ - d'_+) - (\Delta_{zo} - \Delta_{z'o}) = &\pm \Big((\Delta^*_{xo} - \Delta_{xo}) - (\Delta^*_{xz} - \Delta_{xz})\Big) \\ &\pm \Big((\Delta^*_{x'o} - \Delta_{x'o}) - (\Delta^*_{x'z'} - \Delta_{x'z'})\Big),\end{aligned}$$

where $\Delta^*_{xo}$ and $\Delta^*_{x'o}$ are calculated in Line S3, and $\Delta^*_{xz}$ and $\Delta^*_{x'z'}$ are obtained in Lines S5 and S6, respectively. Thus regardless whether $z$ and $z'$ are leaves or inner nodes in $\Psi^*_k$, by $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{E}_c$,

$$\Big|(d_+ - d'_+) - (\Delta_{zo} - \Delta_{z'o})\Big| < 2\Delta_{\min}. \tag{$*$}$$

By Line S10 and Equation ($*$), using the notation $\Delta^*_{z'z} = \mathtt{T}[z]$.length,

$$\begin{aligned}\Delta^*_{zo} <& \frac{(d_+ - d'_+) - (\Delta_{zo} - \Delta_{z'o}) + (\Delta_{zo} - \Delta_{z'o}) + \Delta^*_{z'z}}{2}\\ <& \frac{2\Delta_{\min} + (\Delta_{zo} - \Delta_{z'o}) + \Delta^*_{z'z}}{2}\\ =& \frac{2(2\Delta_{\min} - \Delta_{z'o}) + (-2\Delta_{\min} + \Delta_{z'z}) + \Delta^*_{z'z}}{2}.\end{aligned}$$

Since $o \neq z'$ and thus $\Delta_{z'o} \geq 2\Delta_{\min}$, and $\Big|\Delta^*_{z'z} - \Delta_{z'z}\Big| > 2\Delta_{\min}$ by $\mathcal{Z}_k$, we then obtain $\Delta^*_{zo} < \Delta^*_{z'z}$. By symmetry, $\Delta^*_{z'o} < \Delta^*_{z'z}$ also holds. Thus, the test of Line S11 fails.

Now let us assume that the center of $uvw$ is not on the path between $z$ and $z'$ in $\Psi$. By similar arguments as before, if $\Delta_{zo} > \Delta_{z'z}$ (respectively $\Delta_{z'o} > \Delta_{z'z}$), then $\Delta^*_{zo} > \Delta^*_{z'z}$ (respectively $\Delta^*_{z'o} > \Delta^*_{z'z}$). Thus, the test of Line S11 passes. ■

**Lemma 5.13.** *Assume that $z'z$ is an edge in $\Psi^*_k$ and there exists a node strictly between $z$ and $z'$ in $\Psi$; i.e., $z'z$ is not an edge in $\Psi$. There is subsequently a large triplet $u''v''w''$ with center $z''$ such that $w'' \notin \Psi^*_k$, $u'' \in \mathrm{def}(z)$, $v'' \in \mathrm{def}(z')$, and $z''$ is strictly between $z$ and $z'$. In other words, $\langle u''v''w'', z'z\rangle$ is a strongly relevant pair with the center of $u''v''w''$ falling between $z$ and $z'$.*

Proof. See at the end of the chapter. ■

### 5.2.5 Statistical efficiency of the Basic-HGT algorithm

**Lemma 5.14.** *The events $\mathcal{E}_g$ and $\mathcal{E}_c$ imply that $\mathcal{X}_k$, $\mathcal{Y}_k$ and $\mathcal{Z}_k$ hold for all $k = 3, \ldots, n$.*

Proof. The proof is by induction in $k$.

*Base case*: $k = 3$. By Lemma 5.7 and the greedy selection of Line B1, Line B2 constructs $\Psi^*_3$ for which $\mathcal{X}_3$ holds trivially, and $\mathcal{Y}_3$ follows from $\mathcal{E}_g$. $\mathcal{Z}_3$ follows from $\mathcal{Y}_3$, $\mathcal{E}_c$, and the use of Equation (5.1) at Line J4 of the HGT-Init procedure.

*Induction hypothesis:* $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$ hold for some $k < n$.

*Induction step.* The induction step is concerned with the $(k-2)$-th iteration of the repeat at Line B5. Right before this iteration, by the induction hypothesis, (since $k < n$,) some pair $\langle u''v''w'', z'z \rangle$ satisfies Lemma 5.13. Therefore, during this iteration, by $\mathcal{E}_{\mathrm{c}}$ and Lemmas 5.11, 5.12, and 5.4, $\mathcal{R}$ at Line B6 has a splitting pair in $\Psi^*_k$ that contains a triplet $uvw$ with $\hat{S}_{uvw} > \hat{S}_{u''v''w''}$. Furthermore, Line B7 finds such a pair. By $\mathcal{E}_{\mathrm{g}}$, $uvw$ is not small. Lines B9–B12 create $\Psi^*_3$ using this triplet. Thus $\mathcal{Y}_{k+1}$ follows from $\mathcal{Y}_k$. By Lemmas 5.11 and 5.12, $\mathcal{X}_{k+1}$ follows from $\mathcal{X}_k$. $\mathcal{Z}_{k+1}$ follows from $\mathcal{Z}_k$ since the triplets involved at Line S10 are not small. ∎

**Theorem 5.15.** *For every $0 < \delta < 1$ there exists a sample length*

$$\ell = O\left(\frac{\log\frac{1}{\delta} + \log n}{\vartheta^2 S_1^2 S_0^{4\varrho_{\mathrm{in}}+8}}\right) \tag{5.15}$$

*such that with probability at least $(1-\delta)$, the* Basic-HGT *algorithm outputs $\Psi^*$ represented by the tree structure* T *satisfying both statements below.*

*(i) The algorithm successfully recovers the topology, i.e.,* $\Psi^* \underset{L}{\sim} \Psi$.

*(ii) The edge lengths are recovered within $2\Delta_{\min}$ error, i.e., for all edges $z'z$ in $\Psi^*$, $\left|\Delta_{z'z} - \Delta^*_{z'z}\right| < 2\Delta_{\min}$, where $z' = \mathtt{T}[z].\mathrm{parent}$ and $\Delta^*_{z'z} = \mathtt{T}[z].\mathrm{length}$.*

Proof. By Equation (5.14a), $\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{g}}\right\} < \delta/2$ if

$$\ell \geq \ell_{\mathrm{g}} = 420\frac{3\ln n + \ln\frac{a}{3\delta}}{bS_{\mathrm{lg}}^2}.$$

Similarly, from Equation (5.14b), $\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{c}}\right\} < \delta/2$ if

$$\ell \geq \ell_{\mathrm{c}} = 162\frac{3\ln n + \ln\frac{7a}{\delta}}{b\vartheta^2 S_1^2 S_{\mathrm{lg}}^2}.$$

We choose $\ell = \left\lceil \max\{\ell_{\mathrm{g}}, \ell_{\mathrm{c}}\}\right\rceil$. Consequently, $\mathbb{P}\left\{\mathcal{E}_{\mathrm{g}} \text{ and } \mathcal{E}_{\mathrm{c}}\right\} \geq 1-\delta$. By Lemma 5.14, with probability at least $(1-\delta)$, the Basic-HGT algorithm outputs $\Psi_n^*$, for which $\mathcal{X}_n$ and $\mathcal{Z}_n$ hold, corresponding to the two statements of the theorem. ■

**Algorithm** Fast-HGT

**Input:** estimated tree metric $\hat{\boldsymbol{\Delta}}$, minimum distance parameter $\Delta_{\min}$.

**Output:** tree structure $\mathtt{T}$.

F1 Select and arbitrary leaf $u_0$ and find $u_0v_0w_0$ with the maximum $\hat{S}_{u_0v_0w_0}$.

F2 HGT-Init($\mathtt{T}, \hat{\boldsymbol{\Delta}}, u_0v_0w_0$).

F3 Set $o \leftarrow n+1$. $\quad \triangleright$ *(index for first inner node)*

F4 Set $\mathcal{R} \leftarrow \emptyset$.
**for** each $z'z \in \{u_0o, ov_0, ow_0\}$ **do** Update-Cand($\mathtt{T}, \hat{\boldsymbol{\Delta}}, \mathcal{R}, \Delta_{\min}, z$).

F5 **repeat**

F6     **if** $\mathcal{R} = \emptyset$ **then fail** and **stop**.

F7     Find $\langle uvw, z'z \rangle \in \mathcal{R}$ with maximum $\hat{S}_{uvw}$.

F8     Set $o \leftarrow o+1$. $\quad \triangleright$ *(index for new inner node)*

F9     Add-On-Edge($\mathtt{T}, z, w, o$).

F10     Set $(\Delta^*_{zo}, \Delta^*_{z'o}) \leftarrow$ HGT-Edge-Length($\mathtt{T}, \hat{\boldsymbol{\Delta}}, z, uvw$).

F11     Set-Length($\mathtt{T}, z, w, o, \Delta^*_{zo}, \Delta^*_{z'o}, \mathrm{TC}(\hat{\boldsymbol{\Delta}}, w, uv)$).

F12     Set $\mathrm{def}(o) \leftarrow \{u, v, w\}$.

F13     Remove from $\mathcal{R}$ all pairs containing $z'z$.

F14     **for** each $x'x \in \{z'o, oz, ow\}$ **do** Update-Cand($\mathtt{T}, \hat{\boldsymbol{\Delta}}, \mathcal{R}, \Delta_{\min}, x$).

F15         $\triangleright$ *(This line does no operation in contrast to B15.)*

F16 **until** all leaves are added.

F17 Output $\mathtt{T}$.

Figure 5.14: *The* Fast-HGT *algorithm. Line F12 calls the* Set-Deftrip *procedure after determining whether* $\overset{o}{\underset{u}{\uparrow}}$ *or* $\overset{o}{\underset{v}{\uparrow}}$.

**Procedure** UPDATE-CAND

**Input:** tree structure $\mathtt{T}$, estimated tree metric $\hat{\mathbf{\Delta}}$, candidate pair vector $\mathcal{R}$, minimum distance parameter $\Delta_{\min}$, non-root node $z$.

**Output:** none (updates $\mathcal{R}$)

U1 Set $z' \leftarrow \mathcal{T}[z]$.parent.

U2 **for** each strongly relevant pair $\langle uvw, z'z \rangle$ **do**

U3     **if** HGT-SPLIT-EDGE($\mathtt{T}, \hat{\mathbf{\Delta}}, \Delta_{\min}, z, uvw$)

U4     **then** set $\langle u'v'w, x'x \rangle \leftarrow \mathcal{R}[w]$.

U5         **if** $\mathcal{R}[w] = \mathtt{null}$ or $\hat{S}_{uvw} > \hat{S}_{u'v'w}$

U6         **then** set $\mathcal{R}[w] \leftarrow \langle uvw, z'z \rangle$.

FIGURE 5.15: *The* UPDATE-CAND *procedure. The candidate set* $\mathcal{R}$ *is implemented in the* FAST-HGT *algorithm as a vector* $\mathcal{R}[1 \dots n]$ *with one entry for each leaf not added. Line U2 performs the tests described in Figure 5.9.*

## 5.3 The Fast-HGT algorithm

This section presents the FAST-HGT algorithm and its subroutine UPDATE-CAND in Figures 5.14 and 5.15, respectively. The algorithm is parallel to the BASIC-HGT algorithm. Every line in Figure 5.14 performs the same function as the line with the same numbering in Figure 5.13. Also, for each leaf $w$, $\mathcal{R}[w]$ plays the same role as the column of $\mathcal{R}$ indexed by $w$ in the proof of Theorem 5.5.

The analysis of the FAST-HGT algorithm is also parallel to that of the BASIC-HGT algorithm. Hence, we adapt for the FAST-HGT algorithm in a straightforward manner the definitions of $\Psi^*$, $\Psi^*_k$, and Conditions $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$. A *strongly* splitting pair is a splitting pair whose triplet is strongly relevant. The UPDATE-CAND procedure selects the strongly splitting pairs for a given edge. In the FAST-HGT algorithm, each $\mathcal{R}[w]$ is either a single strongly splitting pair or null. In the former case, the pair is essentially equal to, but subtly different from the pair whose triplet has the largest estimated closeness in the column of $\mathcal{R}$ indexed by $w$ in the proof of Theorem 5.5.

To save running time and work space, the FAST-HGT algorithm differs from the BASIC-HGT algorithm in the following four key aspects:

1. At line F1, the triplet $u_0v_0w_0$ is selected for a fixed arbitrary $u_0$. This reduces the number of triplets considered at line F2 from $O(n^3)$ to $O(n^2)$. This improvement is supported by the fact that each leaf in $\Psi$ is contained in a large triplet.

2. At lines F4 and F14, $\mathcal{R}$ keeps only strongly splitting pairs. This decreases the number of triplets considered at lines B4 and B14 for each involved edge from $O(n^2)$ to $O(n)$. This modification is feasible since by Lemma 5.13, $\Psi$ can be recovered using only strongly relevant triplets.

3. At line F15, $\mathcal{R}$ includes no new strongly splitting pairs for the edges $x'x$ that already exist in $\Psi^*$ before $w$ is inserted. This entirely avoids the $O(n^2)$ triplets considered at line B15. This change is possible because the insertion of $w$ results in no new strongly relevant triplets for any $x'x$ at all.

4. At line F7, the best pair is chosen among $n$ pairs in contrast to $n^2$ at line B7. This new greedy policy is similar to choosing the best among the best pairs of individual columns of $\mathcal{R}$ in the proof of Theorem 5.5.

We now proceed to analyze the FAST-HGT algorithm.

**Lemma 5.16.** *Lemma 5.4(i) holds for the* FAST-HGT *algorithm, i.e., for every edge* $z'z \in \Psi_k^*$, $\operatorname{def}(z') \cap \operatorname{def}(z) \neq \emptyset$.

PROOF. The proof is parallel to that of Lemma 5.4(i) and follows from the fact that a strongly relevant triplet is also relevant. ■

**Theorem 5.17.** *The* FAST-HGT *algorithm runs in* $O(n^2)$ *time using* $O(n)$ *work space.*

PROOF. It is straightforward to show that the time and space complexities are as claimed based on the above four key differences between the FAST-HGT algorithm and the BASIC-HGT algorithm. Note that here $\mathcal{R}$ stores one pair for each leaf using $O(1)$ space instead of $O(n)$ space as in the BASIC-HGT algorithm. ■

**Lemma 5.18.**

*(i) Lemmas 5.11 and 5.12 hold for the* FAST-HGT *algorithm.*

*(ii) Lemma 5.13 holds for the* FAST-HGT *algorithm.*

PROOF. The proofs of Lemmas 5.11, 5.12, and 5.13 refer to lines in HGT-SPLIT-EDGE but to none in the BASIC-HGT algorithm. Thus they hold for the FAST-HGT algorithm based on the fact that a strongly relevant triplet is also relevant. ■

We need the following version of Lemma 5.4(ii) for the FAST-HGT algorithm. For $k = 3, \ldots, n-1$ and each leaf $w$, let $\mathcal{R}_k[w]$ be the version of $\mathcal{R}[w]$ at the start of the $(k-2)$-th iteration of the repeat at Line F5.

**Lemma 5.19.** *Assume that for a given $k \leq n-1$, $\mathcal{E}_{\mathrm{g}}$, $\mathcal{E}_{\mathrm{c}}$, $\mathcal{X}_{k'}$, $\mathcal{Y}_{k'}$, $\mathcal{Z}_{k'}$ hold for all $k' \leq k$.*

*(i) If $\mathcal{R}_k[w]$ is not null, then it is a strongly splitting pair for some edge in $\Psi^*_k$.*

*(ii) If an edge $z'z$ and a triplet $u''v''w''$ with $w'' \notin \Psi^*_k$ satisfy Lemma 5.18(ii), then $\mathcal{R}_k[w'']$ is a strongly splitting pair $\langle uvw'', z'z\rangle$ with $\hat{S}_{uvw''} > \hat{S}_{u''v''w''}$.*

PROOF. The two statements are proved as follows.

Statement 1. This statement follows directly from the initialization of $\mathcal{R}$ at Line F4, the deletions from $\mathcal{R}$ at Line F13, and the insertions into $\mathcal{R}$ at Lines F4 and F14.

Statement 2. The proof is by induction on $k$.

*Base case*: $k = 3$. By $\mathcal{E}_{\mathrm{g}}$, $\mathcal{E}_{\mathrm{c}}$, $\mathcal{X}_3$, $\mathcal{Y}_3$, $\mathcal{Z}_3$, and Lemma 5.18(i), $u''v''w''$ is a strongly splitting pair for some edge $z'z \in \Psi^*_3$. Then, by the maximization in UPDATE-CAND at Line F4, $\mathcal{R}[w'']$ is a strongly splitting pair $\langle uvw'', x'x\rangle$ with $\hat{S}_{uvw''} > \hat{S}_{u''v''w''}$. By $\mathcal{E}_{\mathrm{g}}$, $uvw''$ is not small. Thus, by Lemma 5.18(i), $x'x = z'z$.

*Induction hypothesis:* Statement 2 holds for some $k < n-1$.

*Induction step.* We consider how $\mathcal{R}_{k+1}$ is obtained from $\mathcal{R}_k$ during the $(k-2)$-th iteration of the repeat at Line F5. There are two cases.

*Case 1*: $z'z$ also exists in $\Psi^*_k$. By $\mathcal{X}_k$, $z'z$ and $u''v''w''$ also satisfy Lemma 5.18(i) for $\Psi^*_k$. By the induction hypothesis, $\mathcal{R}_k[w'']$ is a strongly splitting pair for $z'z$ in $\Psi^*_k$ that contains a triplet $uvw''$ with $\hat{S}_{uvw''} > \hat{S}_{u''v''w''}$.

Then $\mathcal{R}_k[w'']$ is not reset to null at Line F13. Thus, it can be changed only through replacement at line F14 by a strongly splitting pair for some edge $x'x$ in $\Psi^*_{k+1}$ that contains a triplet $u'v'w''$. By $\mathcal{E}_{\mathrm{g}}$, $u'v'w''$ is not small. Thus, by $\mathcal{E}_{\mathrm{c}}$, $\mathcal{X}_{k+1}$, $\mathcal{Y}_{k+1}$, $\mathcal{Z}_{k+1}$, and Lemma 5.18(i), $x'x$ equals $z'z$.

*Case 2*: $z'z \in \Psi^*_k$. This case is similar to the base case but uses the maximization in UPDATE-CAND at line F14. ■

**Lemma 5.20.** *Lemma 5.14 holds for the* FAST-HGT *algorithm.*

PROOF. The proof is parallel to that of Lemma 5.14 with Lemma 5.19 replacing Lemma 5.4(ii). ■

**Theorem 5.21.** *For every* $0 < \delta < 1$ *there exists a sample length*

$$\ell = O\left(\frac{\log\frac{1}{\delta} + \log n}{\vartheta^2 S_1^2 S_0^{4\varrho_{\mathrm{in}}+8}}\right) \tag{5.16}$$

*such that with probability at least* $(1-\delta)$, *the* FAST-HGT *algorithm outputs* $\Psi^*$ *represented as the tree structure* $\mathtt{T}$ *satisfying both statements below.*

*(i) The algorithm successfully recovers the topology, i.e.,* $\Psi^* \underset{L}{\sim} \Psi$.

*(ii) The edge lengths are recovered within* $2\Delta_{\min}$ *error, i.e., for all edges* $z'z$ *in* $\Psi^*$, $\left|\Delta_{z'z} - \Delta^*_{z'z}\right| < 2\Delta_{\min}$, *where* $z' = \mathtt{T}[z].\mathrm{parent}$ *and* $\Delta^*_{z'z} = \mathtt{T}[z].\mathrm{length}$.

PROOF. The proof is parallel to that of Theorem 5.15. ■

## 5.4 A closer look at the minimum distance parameter

Both BASIC-HGT and FAST-HGT use a minimum distance parameter $\Delta_{\min}$ to recognize separate and identical triplet centers. Theorems 5.15 and 5.21 state that the algorithms are statistically efficient when $\Delta_{\min}$ is less than half the minimum edge length, i.e., when $\Delta_{\min} \leq \frac{-\ln(1-S_1)}{2}$. In what follows we show that the algorithms are statistically efficient for any $0 < \Delta_{\min} < -\ln(1-S_1)$. Furthermore, we describe how a deterministic setting of the minimum distance parameter can preserve statistical efficiency.

In order to prove that the range of the minimum distance parameter for which BASIC-HGT and FAST-HGT are statistically efficient can be extended, we employ the following technique. Let $0 < \Delta_{\text{input}} < -\ln(1-S_1)$ be arbitrary and define

$$\Delta_{\min} = \min\Big\{\Delta_{\text{input}}, (-\ln(1-S_1)) - \Delta_{\text{input}}\Big\}. \tag{5.17}$$

Consequently, $\Delta_{\min} \leq \frac{-\ln(1-S_1)}{2}$, and $2\Delta_{\min} \leq \Delta_{\text{input}} + \Delta_{\min} \leq -\ln(1-S_1)$. We claim that using $\Delta_{\text{input}}$ as the input parameter to BASIC-HGT or BASIC-HGT, the analysis of statistical efficiency remains valid with $\Delta_{\min}$. More precisely, after substituting $\Delta_{\text{input}}$ for $\Delta_{\min}$ in the procedures and algorithms of §5.2 and §5.3, the lemmas and theorems remain true without any textual change, using $\Delta_{\min}$ defined by Equation (5.17). Indeed, Lemma 5.13 does not depend on $\Delta_{\min}$, Lemmas 5.8, 5.9, and 5.12 hold for any $\Delta_{\min} > 0$, and Lemma 5.11 holds with minimal changes as shown here.

PROOF OF LEMMA 5.11. Let $o$ be the center of $uvw$. From Line S7,

$$d = \Big(\Delta^*_{xo} - \Delta_{xo}\Big) - \Big(\Delta^*_{xz} - \Delta_{xz}\Big) + \Big(\Delta_{xo} - \Delta_{xz}\Big). \tag{$*$}$$

If $z = o$ then $z$ is an inner node of $\Psi^*_k$. Since $\mathcal{E}_{\text{c}}$ holds, and neither $uvw$ nor $\text{def}(z)$ is small,

$$\Big|\Delta^*_{xo} - \Delta_{xo}\Big| < \frac{\Delta_{\min}}{2}, \qquad \Big|\Delta^*_{xz} - \Delta_{xz}\Big| < \frac{\Delta_{\min}}{2}. \tag{$**$}$$

Since $z = o$, $\Delta_{xo} = \Delta_{xz}$ and by Equations $(*)$ and $(**)$, $|d| < \Delta_{\min}/2 + \Delta_{\min}/2 + 0 < \Delta_{\min} \leq \Delta_{\text{input}}$ and thus the test of Line S8 passes. Using a similar reasoning, if $z' = o$, then $|d'| < \Delta_{\min}$ and the test passes.

If $o \neq z, z'$, then $|\Delta_{xo} - \Delta_{xz}| \geq -\ln(1-S_1) \geq \Delta_{\min} + \Delta_{\text{input}}$ since the center of $o$ and $z$ are both on the path between $u$ and $v$ in $\Psi$. If $z$ is a leaf in $\Psi^*_k$, then $\Delta^*_{xz} = \Delta_{xz} = 0$. By $\mathcal{E}_{\text{c}}$ and Equation $(*)$, $|d| > \Delta_{\text{input}} + \frac{\Delta_{\min}}{2}$. If $z$ is an inner node in $\Psi^*_k$, then by $\mathcal{Y}_k$, $\mathcal{E}_{\text{c}}$, and Equation $(*)$, $|d| > \Delta_{\text{input}}$. In either case, $|d| > \Delta_{\text{input}}$. By symmetry, $|d'| > \Delta_{\text{input}}$ also. Hence the test of Line S8 fails. ■

Consequently, the input parameter $\Delta_{\text{input}}$ to BASIC-HGT and FAST-HGT may take any positive value as long as it is less than the minimum edge length in the tree. A larger value decreases the error in recognizing identical triplet centers and increases the error of not recognizing separate triplet centers. A smaller value has the opposite effect. As a general observation, rejecting triplets with separate centers is a less severe mistake than adding triplets with identical centers, and thus larger values are usually more preferable. In particular, our sample length bounds are the same for $\Delta_{\text{input}} = \frac{-\ln(1-S_1)}{2} - c$ and $\Delta_{\text{input}} = \frac{-\ln(1-S_1)}{2} + c$ with arbitrary $c$, but the latter choice is more desirable. We return to this question in §5.6.2.

A theoretically interesting solution to setting $\Delta_{\text{input}}$ without the knowledge of $S_1$ is to set it as a function of the sample length $\ell$. For instance, let $\alpha > 0$ be an arbitrary value, and let $\Delta_{\text{input}} = \ell^{-1/(2+2\alpha)}$. If $\Delta_{\text{input}} < -\ln(1-S_1)$, then by Equations (5.15) and (5.16) there exists

$$\ell_\alpha = O\left(\left(\frac{\log(1/\delta) + \log n}{S_0^{4\varrho_{\text{in}}+8}}\right)^{1+\frac{1}{\alpha}}\right)$$

such that the Basic-HGT and Fast-HGT algorithms successfully recover the topology. Hence if

$$\ell \geq \max\left\{\ell_\alpha, S_1^{-2(1+\alpha)}\right\},$$

then the algorithms are successful. A choice of $\alpha = 1$ keeps the statistical efficiency, and so do other more interesting choices such as $\alpha = \ln n$ or $\alpha = \ln\ln n$.

## 5.5 Harmonic Greedy Triplets and the Four-Point Condition

New possibilities of employing the Harmonic Greedy Triplets principle arise if we replace the HGT-Split-Edge procedure in Basic-HGT or Fast-HGT with another reasonable way of deciding whether a triplet defines a new center on an edge. Here we describe the use of the relaxed four-point condition for that purpose. The relaxed four-point condition is applied to a relevant pair $\langle uvw, z'z\rangle$ in the following manner, as illustrated in Figure 5.16. Let $z$ be an internal node in $\Psi^*$, let $\operatorname{def}(z) = \{x, x', x''\}$, and assume that $z'$ lies on the path between $z$ and $x$ in $\Psi^*$ without loss of generality. We test

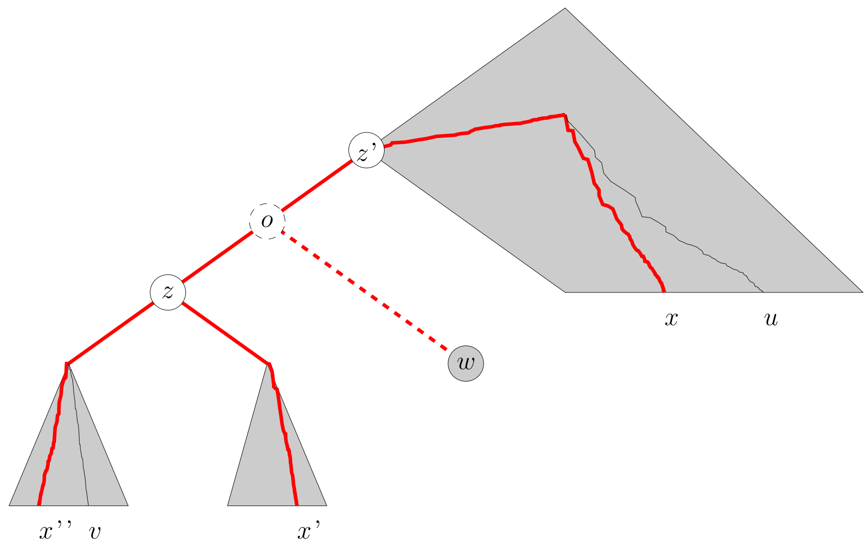



FIGURE 5.16: *Using the four-point condition to evaluate relevant pairs. The relaxed four-point condition is checked for the quartet topology $xw|x'x''$ to determine whether the center of the triplet $uvw$ may fall onto the edge $z'z$.*

whether the relaxed four-point condition holds for $xw|x'x''$, which signifies that for the center $o$ of $uvw$, either $o$ lies on the path between $z$ and $z'$, or $z'$ lies on the path between $z$ and $o$. Recall from Equation (4.11), that the relaxed four-point condition holds for $xw|x'x''$ if

$$\hat{\mathbf{\Delta}}[x,w] + \hat{\mathbf{\Delta}}[x',x''] < \min\Big\{\hat{\mathbf{\Delta}}[x,x'] + \hat{\mathbf{\Delta}}[w,x''], \hat{\mathbf{\Delta}}[x,x''] + \hat{\mathbf{\Delta}}[w,x']\Big\}. \quad (5.18)$$

The relaxed four-point condition is checked similarly for $z'$ if it is an internal node. If $z$ (or $z'$) is a leaf, the condition for $z$ (respectively, for $z'$) is not tested. The FPC-SPLIT-EDGE procedure detailed in Figure 5.17 implements this method.

Both the BASIC-HGT and FAST-HGT algorithms can be used with the FPC-SPLIT-EDGE procedure replacing HGT-SPLIT-EDGE. We discuss in

**Procedure** FPC-SPLIT-EDGE

**Input:** tree structure $\mathtt{T}$, estimated tree metric $\hat{\boldsymbol{\Delta}}$, non-root node $z$, triplet $uvw$.

**Output:** `true` or `false`
▷ *(whether the center of uvw falls onto the edge $z'z$)*

P1 Set $z' \leftarrow \mathtt{T}[z]$.parent.

P2 **if** $z$ is not a leaf
**then** ▷ *(cf. Figure 5.16)*

P3 set $x \leftarrow \mathtt{T}[z]$.def.up, $x' \leftarrow \mathtt{T}[z]$.def.left, $x'' \leftarrow \mathtt{T}[z]$.def.right.

P4 **if** the relaxed four-point condition does not hold for $xw|x'x''$
**then return** `false`.

P5 **if** $z'$ is not the root
**then**

P6 **if** $z$ is the left child

P7 **then** set $x \leftarrow \mathtt{T}[z']$.def.left, $x' \leftarrow \mathtt{T}[z']$.def.right;
**else** set $x \leftarrow \mathtt{T}[z']$.def.right, $x' \leftarrow \mathtt{T}[z']$.def.left.

P8 Set $x'' \leftarrow \mathtt{T}[z']$.def.up.

P9 **if** the relaxed four-point condition does not hold for $xw|x'x''$
**then return** `false`.

P10 **return** `true`.

FIGURE 5.17: *The* FPC-SPLIT-EDGE *procedure uses the relaxed four-point condition of Equation* (5.18) *in Lines P4 and P9 to evaluate relevant pairs.*

detail how to use FPC-SPLIT-EDGE with the FAST-HGT algorithm only. The resulting algorithm called HGT-FP is described in Figure 5.18 along with its FPC-UPDATE-CAND subroutine in Figure 5.19.

**Theorem 5.22.** *The* HGT-FP *algorithm runs in $O(n^2)$ time using $O(n)$ work space.*

PROOF. The proof is analogous to that of Theorem 5.17, using the fact that FPC-SPLIT-EDGE runs in $O(1)$ time and uses $O(1)$ work space. ■

**Algorithm** HGT-FP
*(Harmonic Greedy Triplets with Four Point Condition)*

**Input:** estimated tree metric $\hat{\mathbf{\Delta}}$.

**Output:** tree structure $\mathtt{T}$.

Q1 Select and arbitrary leaf $u_0$ and find $u_0v_0w_0$ with the maximum $\hat{S}_{u_0v_0w_0}$.

Q2 HGT-Init($\mathtt{T}, \hat{\mathbf{\Delta}}, u_0v_0w_0$).

Q3 Set $o \leftarrow n + 1$. ▷ *(index for first inner node)*

Q4 Set $\mathcal{R} \leftarrow \emptyset$.
**for** each $z'z \in \{u_0o, ov_0, ow_0\}$ **do** FPC-Update-Cand($\mathtt{T}, \hat{\mathbf{\Delta}}, \mathcal{R}, z$).

Q5 **repeat**

Q6 **if** $\mathcal{R} = \emptyset$ **then fail** and **stop**.

Q7 Find $\langle uvw, z'z \rangle \in \mathcal{R}$ with maximum $\hat{S}_{uvw}$.

Q8 Set $o \leftarrow o + 1$. ▷ *(index for new inner node)*

Q9 Add-On-Edge($\mathtt{T}, z, w, o$).

Q10 Set $(\Delta^*_{zo}, \Delta^*_{z'o}) \leftarrow$ HGT-Edge-Length($\mathtt{T}, \hat{\mathbf{\Delta}}, z, uvw$).

Q11 Set-Length($\mathtt{T}, z, w, o, \Delta^*_{zo}, \Delta^*_{z'o}$, TC($\hat{\mathbf{\Delta}}, w, uv$)).

Q12 Set $\mathrm{def}(o) \leftarrow \{u, v, w\}$.

Q13 Remove from $\mathcal{R}$ all pairs containing $z'z$.

Q14 **for** each $x'x \in \{z'o, oz, ow\}$ **do** FPC-Update-Cand($\mathtt{T}, \hat{\mathbf{\Delta}}, \mathcal{R}, x$).

Q15 **until** all leaves are added.

Q16 Output $\mathtt{T}$.

Figure 5.18: *The* HGT-FP *algorithm. Line Q12 calls the* Set-Deftrip *procedure after determining whether* $\overset{o}{\underset{u}{\uparrow}}$ *or* $\overset{o}{\underset{v}{\uparrow}}$.

The statistical analysis of the HGT-FP algorithm is parallel to that of the Fast-HGT algorithm. We claim that the success of the HGT-FP algorithm is implied by the events $\mathcal{E}_\mathrm{g}$ and $\mathcal{E}_\mathrm{c}$ used in the analysis of the Basic-HGT and Fast-HGT algorithms, and the specific event $\mathcal{E}_\mathrm{q}$, described by Definition 5.3.

**Procedure** FPC-UPDATE-CAND

**Input:** tree structure $\mathtt{T}$, estimated tree metric $\hat{\boldsymbol{\Delta}}$, candidate pair vector $\mathcal{R}$, non-root node $z$.

**Output:** none (updates $\mathcal{R}$)

C1 Set $z' \leftarrow \mathtt{T}[z]$.parent.

C2 **for** each strongly relevant pair $\langle uvw, z'z \rangle$ **do**

C3     **if** FPC-SPLIT-EDGE($\mathtt{T}, \hat{\boldsymbol{\Delta}}, z, uvw$)

C4     **then** set $\langle u'v'w, x'x \rangle \leftarrow \mathcal{R}[w]$.

C5         **if** $\mathcal{R}[w] = \mathtt{null}$ or $\hat{S}_{uvw} > \hat{S}_{u'v'w}$

C6         **then** set $\mathcal{R}[w] \leftarrow \langle uvw, z'z \rangle$.

FIGURE 5.19: *The* FPC-UPDATE-CAND *procedure. Line C2 performs the tests described in Figure 5.9.*

**Definition 5.3.** *Let $\mathcal{E}_{\mathrm{q}}$ denote the event that for every leaf pair $u, v$, if*

$$\mathbf{\Delta}[u,v] \le 2\Big(-\ln\frac{S_{\mathrm{sm}}}{3}\Big),$$

*then*

$$\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| < \frac{-\ln(1-S_1)}{4}.$$

**Lemma 5.23.** *The probability of the complementary event to $\mathcal{E}_{\mathrm{q}}$ is bounded from above as*

$$\mathbb{P}\Big\{\bar{\mathcal{E}}_{\mathrm{q}}\Big\} \le 2a\binom{n}{2}\exp\Big(-\frac{b}{5184}\ell S_{\mathrm{lg}}^4 S_1^2\Big). \tag{5.19}$$

Proof. Let $u, v$ be two leaves with $\mathbf{\Delta}[u,v] \le 2\Big(-\ln(S_{\mathrm{sm}}/3)\Big)$. Trivially,

$$\begin{aligned}
&2\mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| \ge \frac{-\ln(1-S_1)}{4}\Big\}\\
&= \mathbb{P}\Big\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \ge \frac{-\ln(1-S_1)}{4}\Big\} + \mathbb{P}\Big\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \le -\frac{-\ln(1-S_1)}{4}\Big\}.
\end{aligned}$$

Since $\hat{\mathbf{\Delta}}$ is an $(a,b)$-regular estimator of $\mathbf{\Delta}$, by Equation (5.5),

$$\begin{aligned}
&2\mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| \ge \frac{-\ln(1-S_1)}{4}\Big\}\\
&\le a\exp\Big(-b\ell S_{uv}^2\Big(1-(1-S_1)^{1/4}\Big)^2\Big) + a\exp\Big(-b\ell S_{uv}^2\Big((1-S_1)^{-1/4}-1\Big)^2\Big).
\end{aligned}$$

By Taylor's expansion,

$$\Big(1-(1-S_1)^{1/4}\Big)^2 \ge \frac{1}{16}S_1^2; \qquad \Big((1-S_1)^{-1/4}-1\Big)^2 \ge \frac{1}{16}S_1^2.$$

Thus

$$\mathbb{P}\Big\{\Big|\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v]\Big| \ge \frac{-\ln(1-S_1)}{4}\Big\} \le 2a\exp\Big(-\frac{b}{16}\ell S_{uv}^2 S_1^2\Big).$$

The lemma follows from the facts that there are $\binom{n}{2}$ leaf pairs, and that for every leaf pair involved in $\mathcal{E}_{\mathrm{q}}$, $S_{uv} \ge S_{\mathrm{lg}}^2/18$. ■

The analysis leading to bounding the sample size is analogous to the one for FAST-HGT. The events $\mathcal{E}_{\mathrm{g}}$, $\mathcal{E}_{\mathrm{c}}$, and $\mathcal{E}_{\mathrm{q}}$ imply the invariants $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$ for $k = 3, \ldots, n$. We use the event $\mathcal{E}_{\mathrm{c}}$ only to show that the edge lengths are estimated correctly, i.e., that $\mathcal{Z}_k$ holds, by setting $\Delta_{\min}$ arbitrarily for the argument's sake. Note that the edge length estimation can be omitted from the algorithm entirely. Arguments in the analysis relying on correct decisions made by HGT-SPLIT-EDGE as stated by Lemmas 5.11, 5.12, and 5.18(i) need to refer to the following lemma.

**Lemma 5.24.** *Assume that the* FPC-SPLIT-EDGE *procedure is called with the relevant pair* $\langle uvw, z'z\rangle$. *If* $\mathcal{X}_k$, $\mathcal{Y}_k$, *and* $\mathcal{E}_{\mathrm{q}}$ *hold, and* $uvw$ *is not small, then the procedure returns* `true` *if and only if the center of* $uvw$ *is on the path between* $z$ *and* $z'$ *in* $\Psi$.

PROOF. By $\mathcal{Y}_k$, $\operatorname{def}(z)$ and $\operatorname{def}(z')$ are not small. Since $uvw$ is not small either, the distance between $w$ and an arbitrary member of $\operatorname{def}(z)$ or $\operatorname{def}(z')$, as well as between members $\operatorname{def}(z)$ or $\operatorname{def}(z')$ is bounded from above by $2\Big(-\ln(S_{\mathrm{sm}}/3)\Big)$. Since $\mathcal{E}_{\mathrm{q}}$ holds, the tests at Lines P4 and P9 correctly establish the quartet topologies using the relaxed four-point condition. ■

**Theorem 5.25.** *For all* $0 < \delta < 1$ *and* $0 < \vartheta < 1/2$, *there exists a sample length*

$$\ell = O\left(\frac{\log\frac{1}{\delta} + \log n}{\vartheta^2 S_1^2 S_0^{8\varrho_{\mathrm{in}}+16}}\right) \tag{5.20}$$

*such that with probability at least* $(1-\delta)$, *the* HGT-FP *algorithm outputs* $\Psi^*$ *represented by the tree structure* `T` *satisfying both ensuing statements.*

*(i) The algorithm successfully recovers the topology, i.e.,* $\Psi^* \underset{L}{\sim} \Psi$.

*(ii) The edge lengths are recovered within* $2\vartheta\Big(-\ln(1-S_1)\Big)$ *error, i.e., for all edges* $z'z$ *in* $\Psi^*$, $\Big|\Delta_{z'z} - \Delta^*_{z'z}\Big| < 2\vartheta\Big(-\ln(1-S_1)\Big)$, *where* $z' =$ `T`$[z]$.parent *and* $\Delta^*_{z'z} =$ `T`$[z]$.length.

PROOF. By Equation (5.14a), $\mathbb{P}\Big\{\bar{\mathcal{E}}_{\mathrm{g}}\Big\} < \delta/3$ if

$$\ell \geq \ell'_{\mathrm{g}} = 420\frac{3\ln n + \ln\frac{a}{2\delta}}{bS_{\mathrm{lg}}^2}.$$

Similarly, from Equation (5.14b), $\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{c}}\right\} < \delta/3$ if

$$\ell \geq \ell'_{\mathrm{c}} = 162\frac{3\ln n + \ln\frac{21a}{2\delta}}{b\vartheta^2 S_1^2 S_{\mathrm{lg}}^2}.$$

By Equation (5.19), $\mathbb{P}\left\{\bar{\mathcal{E}}_{\mathrm{q}}\right\} < \delta/3$ if

$$\ell \geq \ell_{\mathrm{q}} = 5184\frac{2\ln n + \ln\frac{a}{3\delta}}{b\vartheta^2 S_1^2 S_{\mathrm{lg}}^4}.$$

Set $\ell = \left\lceil \max\{\ell'_{\mathrm{g}}, \ell'_{\mathrm{c}}, \ell_{\mathrm{q}}\}\right\rceil$. Subsequently, $\mathbb{P}\left\{\mathcal{E}_{\mathrm{g}} \text{ and } \mathcal{E}_{\mathrm{c}} \text{ and } \mathcal{E}_{\mathrm{q}}\right\} \geq 1-\delta$. The proof that $\mathcal{X}_k$, $\mathcal{Y}_k$, and $\mathcal{Z}_k$ are implied by $\mathcal{E}_{\mathrm{g}}$, $\mathcal{E}_{\mathrm{c}}$, and $\mathcal{E}_{\mathrm{q}}$ is analogous to the one for Fast-HGT, and uses Lemma 5.24 instead of Lemmas 5.11, 5.12, and 5.18(i). ■

# 5.6 Experimental results

## 5.6.1 Robinson-Foulds distance

Simulated experiments are often used to assess the statistical efficiency of evolutionary tree reconstruction algorithms (Hillis *et al.* 1994; Hillis 1995). Simulation consists of generating sample sequences with the distribution defined by an evolutionary tree $\mathcal{P}$. The output topology $\Psi^*$ of the algorithm is compared to the topology $\Psi$ of $\mathcal{P}$ using distance measures between unrooted binary trees (Day 1983b). We use the Robinson-Foulds distance (Robinson and Foulds 1981) for this purpose, defined as follows. Let $\mathcal{T}$ be an unrooted binary tree with leaf set $L$. A *split* generated by an edge $e$ is the unordered pair $(L_1, L_2)$, where $L_1$ and $L_2$ are the leaf sets of the two subtrees obtained by removing $e$ from $\mathcal{T}$. The *split set* Splits($\mathcal{T}$) is the set of all splits generated by edges of $\mathcal{T}$. Let $\mathcal{T}_1$, $\mathcal{T}_2$ be two unrooted trees with the same leaf set $L$ and let $n = |L|$. The *normalized Robinson-Foulds distance* between $\mathcal{T}_1$ and $\mathcal{T}_2$ is defined as

$$\mathsf{RF\%}(\mathcal{T}_1, \mathcal{T}_2) = \frac{\left|\mathrm{Splits}(\mathcal{T}_1)\right| + \left|\mathrm{Splits}(\mathcal{T}_2)\right| - 2\left|\mathrm{Splits}(\mathcal{T}_1) \cap \mathrm{Splits}(\mathcal{T}_2)\right|}{2(n-3)},$$

which is always between 0 and 100%. We say that a topology reconstruction algorithm has $\delta$ *Robinson-Foulds error* on a given sample sequence generated by a phylogeny with topology $\Psi$, if for the algorithm's output $\Psi^*$, $\mathsf{RF\%}(\Psi^*, \Psi) = \delta$. The algorithm is successful, i.e., $\Psi^* \underset{L}{\sim} \Psi$, if and only if it has zero Robinson-Foulds error.

### 5.6.2 Using the minimum evolution heuristic

In order to estimate the topology accurately, the Fast-HGT algorithm uses the minimum distance parameter $\Delta_{\min}$ to recognize triplets with the same center. Taking the arguments of §5.4 into account, Theorem 5.21 shows that if $\Delta_{\min}$ is smaller than the minimum edge length in the tree, then Fast-HGT is statistically efficient. In order to optimize the performance of the Fast-HGT algorithm on actual data, we conducted a series of simulated experiments on a 135-leaf tree in the Jukes-Cantor model of evolution. The tree (see Figure 5.33 for the topology) is based on a phylogeny derived from mitochondrial DNA sequences in the course of debating the African origin of humans (Maddison *et al.* 1992). We scaled the edge lengths linearly from the originally calculated number of character changes per edge, so that all edge lengths fall into the interval $[0.125, 1.0]$. This same scaled tree was also used by Huson *et al.* (1999) in similar experiments. Figure 5.20 shows the results of three experiments with sample length $\ell = 1500$. The optimal choice of $\Delta_{\min}$ is slightly larger than half the minimum edge length in the tree in all three cases.

The graphs in Figure 5.20 suggest that $\Delta_{\min}$ should be set to around 2/3 of the shortest edge length to achieve the minimum $\mathsf{RF\%}$ error. In practice, however, the shortest edge length is unknown. When a lower bound estimation is available, one can use that to set $\Delta_{\min}$. Here we report that the minimum evolution heuristic can be used to set $\Delta_{\min}$.

While building the output tree $\Psi^*$, the Fast-HGT algorithm keeps track of edge length estimates $\Delta^*_{z'z} = \mathtt{T}[z].\mathrm{length}$. By keeping track of the tree length $\mathrm{ME}(\Psi^*, \Delta^*)$ together with the Robinson-Foulds error, we found a strong correlation between them. Figure 5.21 shows the Robinson-Foulds error as a function of $\mathrm{ME}(\Psi^*, \Delta^*)$, with error values at the same choice of $\Delta_{\min}$ clustered close to each other. Figure 5.22 plots the trajectories described by the curve $\langle \mathrm{ME}(\Psi^*, \Delta^*), \mathsf{RF\%}(\Psi^*, \Psi)\rangle$ with $\Delta_{\min}$ as a parameter in ten different experiments. Not only do $\mathrm{ME}(\Psi^*, \Delta^*)$ and $\mathsf{RF\%}(\Psi^*, \Psi)$ both verge

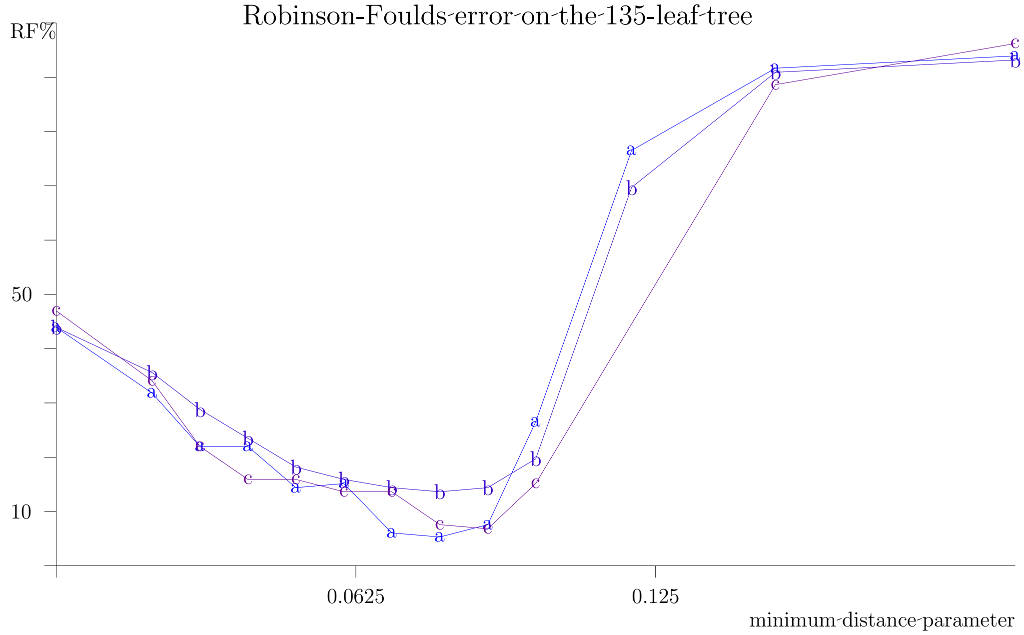



FIGURE 5.20: *Error of the* FAST-HGT *algorithm as a function of the minimum distance parameter* $\Delta_{\min}$ *in three simulated experiments (a,b,c) with sample length* $\ell = 1500$. *The normalized Robinson-Foulds distance (as percentage) between the output and the real topology is shown on the ordinate. The shortest edges in the tree have length* 0.125. *The smallest* RF% *error is achieved at setting* $\Delta_{\min}$ *slightly larger than half the shortest edge length.*

on being unimodal functions of $\Delta_{\min}$ but their minima are also very close to each other. Based on these observations, we designed an iterative procedure that uses FAST-HGT as a subroutine. The FAST-HGT algorithm is run in each iteration with a different $\Delta_{\min}$ value. The tree length of the output $\Psi^*$ is calculated, which we aim to minimize using Golden Section search (Press *et al.* 1992) which quickly finds the minimum of a unimodal function. Since the number of $\Delta_{\min}$ values giving different results is determined by the granularity of $\hat{\Delta}$, only $O(\log \ell)$ iterations are needed and they all use the same input matrix. The resulting HGT-ME algorithm is described in Figure 5.23.

**Theorem 5.26.** *The running time of the* HGT-ME *algorithm building a tree with $n$ leaves is $O(n^2 \log \ell)$. The algorithm uses $O(n)$ work space.*

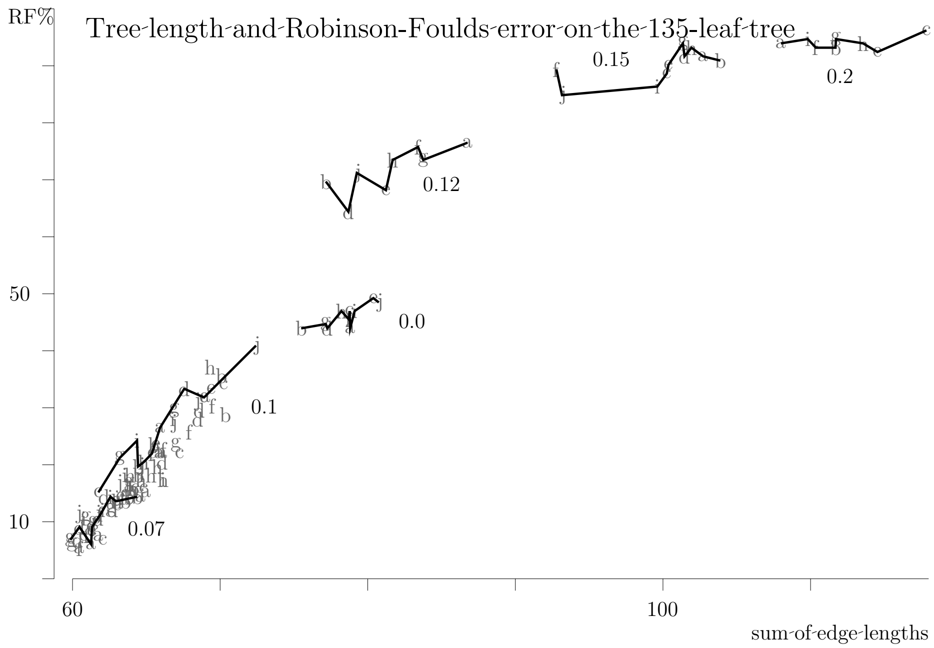



FIGURE 5.21: *Error of the* FAST-HGT *algorithm as the function of the output tree length in ten simulated experiments (a,...,j) with different choices of* $\Delta_{\min}$*, at sample length* $\ell = 1500$*. The normalized Robinson-Foulds distance between the output and the real topology is shown on the ordinate. The* RF% *error values at the same choice of* $\Delta_{\min}$ *in different experiments are clustered closely together. Moreover, the* RF% *error value is an "almost monotone" function of the output tree length. Figure 5.22 shows the shape of the trajectories traversed.*

PROOF. By setting the minimum bracketing value $\epsilon = \ell^{-1}/2$, FAST-HGT is executed at most $\lfloor \log_2 \ell \rfloor$ times in Step M1, and at most $2 + \lceil \log_\beta(2\ell) \rceil$ times in Step M2 with $\beta = (1 + \sqrt{5})/2$. By Theorem 5.17, the running time of FAST-HGT is $O(n^2)$, hence the running time of HGT-ME is as stated by the theorem. FAST-HGT needs $O(n)$ space, and HGT-ME stores at most four topologies at a time to carry out the minimization. ■

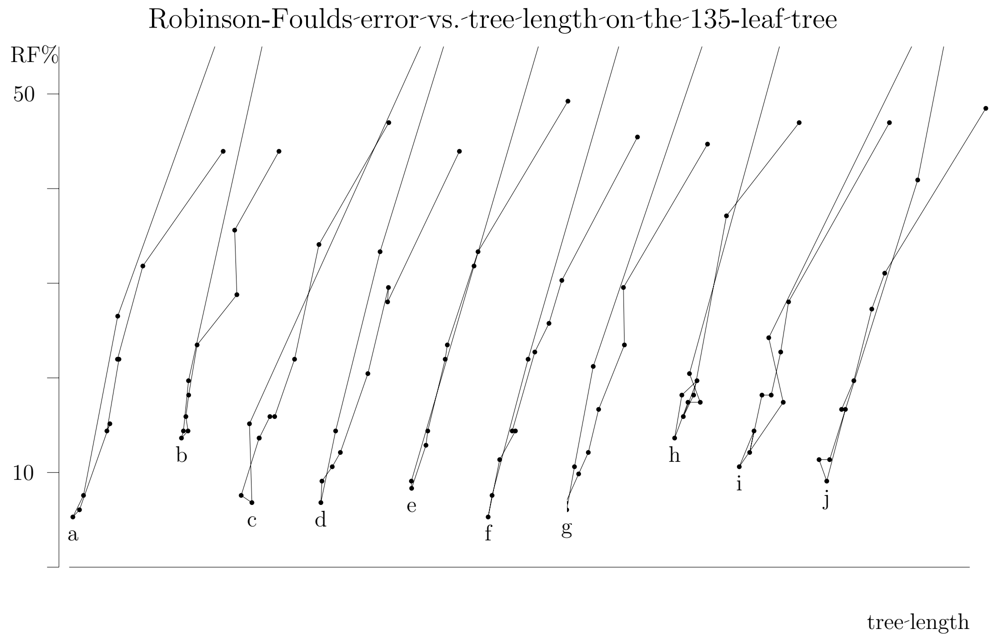



FIGURE 5.22: *This figure is the result of ten simulated experiments on the 135-leaf tree with topology* $\Psi$*, for sample length* $\ell = 1500$*. In each experiment, both the length of the output tree* $\Psi^*$ *and the* RF% *error were calculated at different choices of* $\Delta_{\min}$*. The shape of trajectories* $\langle \mathrm{ME}(\Psi^*, \Delta^*), \mathsf{RF\%}(\Psi^*, \Psi)\rangle$ *are depicted, showing that* ME *and* RF% *are approximately unimodal functions of* $\Delta_{\min}$ *taking their minima close to each other.*

### 5.6.3 Computational efficiency in experiments

Computational aspects of tree reconstruction methods, such as time and space requirements, become accentuated when the number $n$ of taxa is in the order of hundreds or thousands. Exhaustive search among all possible topologies is not feasible since the number of different topologies is super-exponential in $n$ by Equation (4.1). In fact, even computationally efficient algorithms, which run in polynomial time, may be too slow if the order of the polynomial in $n$ is more than three. In order to illustrate this point, we measured the execution time of a number of algorithms on a desktop com-

**Algorithm** HGT-ME
*(Harmonic Greedy Triplets with Minimum Evolution Heuristic)*

**Input:** estimated tree metric $\hat{\mathbf{\Delta}}$, and a small positive bracketing value $\epsilon$, such as $\epsilon = \ell^{-1}/2$.

**Output:** Hypothetical topology $\Psi^*$ with edge length estimates $\Delta^*$.

M1 Find $a = 2^k\epsilon$ with the smallest $k = 1, 2, \ldots, \lfloor -\log_2 \epsilon \rfloor$ such that Fast-HGT builds a full tree with $\Delta_{\min} = 2^k\epsilon$.

M2 Minimize $\mathrm{ME}(\Psi^*, \Delta^*)$ on the interval $\Delta_{\min} \in [0, a]$ with Golden Section Search using $\epsilon$ as the minimum bracketing size, and output $\Psi^*$ built by Fast-HGT at that value.

Figure 5.23: *The* HGT-ME *algorithm.*

puter [1]. We measured the execution time of the following algorithms: FastDNAML (Olsen *et al.* 1994), implemented by its authors; Heuristic parsimony implemented in Phylip (Felsenstein 1993) as DNAPARS; Neighbor-Joining (Studier and Keppler 1988), implemented in qclust (Brzustowski 1998); Unweighted Neighbor-Joining (Gascuel 1997b) (UNJ), implemented in T-REX (Makarenkov and Casgrain 1999); ADDTREE (Sattath and Tversky 1977), implemented in T-REX; Fitch-Margoliash method (Fitch and Margoliash 1967), implemented in Phylip as FITCH; BioNJ (Gascuel 1997a), implemented by its author; Weighbor (Bruno *et al.* 2000), implemented by its authors; UPGMA (Sokal and Michener 1957; Sokal and Sneath 1963), implemented in qclust; and HGT-FP, our implementation.

Figure 5.24 shows the processor times of these programs as a function of $n$. The disadvantage of maximum likelihood is immediately clear from the graphs. The optimization of the likelihood function is extremely time-consuming, even with heuristic approaches such as FastDNAML. As a consequence, maximum likelihood cannot be used for the recovery of trees with more than about forty terminal taxa.

[1] We used a PC with a Pentium III 550 MHz CPU and 128 MB memory, running Windows NT 4.0, compiling executing, and timing the programs with `cygwin` v1.1 utilities. We compiled the source codes using `gcc` or `g++`, with maximum optimization (`-O9`) enabled. We edited some of the I/O functions in the sources to conform to the Phylip formats.

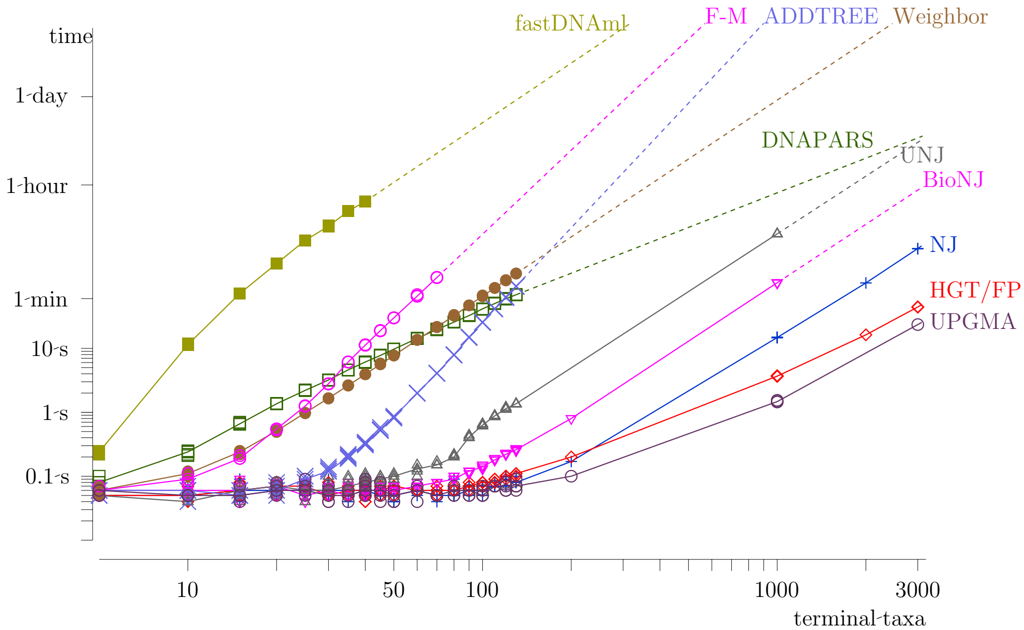



FIGURE 5.24: *Running time of various reconstruction methods as a function of the number of terminal taxa. For each size $n$, the timing was repeated 5 times; the graphs go through the median values. The trees had random topologies, drawn by the Yule-Harding distribution. Each edge length was set to* 0.1*, and random DNA sequences of length 2000 were generated using the Jukes-Cantor model. The dashed lines denote extrapolated values. NJ: Neighbor-Joining, F-M: Fitch-Margoliash.*

As for maximum parsimony, the optimization is also extremely time-consuming, but recent advances in heuristic parsimony optimization make it a viable method for trees with up to at least a few hundred taxa. However, the running time bounds of heuristic optimization are not clear, and we found that for large mutation probabilities permitting many equally parsimonious trees, DNAPARS is very slow.

Distance-based methods have been favored for their computational speed. The classic Fitch-Margoliash (Fitch and Margoliash 1967) algorithm runs in $O(n^4)$ time for trees with $n$ leaves. The ADDTREE (Sattath and Tversky

1977) algorithm also runs in $O(n^4)$ time. Neighbor-Joining (Saitou and Nei 1987), and a number of related methods, such as UNJ (Gascuel 1997b), BioNJ (Gascuel 1997a), and Weighbor (Bruno *et al.* 2000), run in $O(n^3)$ time. To our knowledge, the only topology reconstruction algorithm with $O(n^2)$ running time is UPGMA, which is a hierarchical clustering method rarely used nowadays for phylogeny reconstruction. UPGMA is statistically inconsistent and performs very poorly in simulated experiments.

Our graphs in Figure 5.24 illustrate that while polynomial running time is a necessary requirement for large-scale phylogeny reconstruction, it is not sufficient in itself, as $O(n^4)$ algorithms become too slow when $n$ is in the order of thousands. Even Weighbor with $O(n^3)$ running time is not fast enough due to expensive computations hidden by the asymptotic notation. In contrast, UPGMA and HGT-FP run in $O(n^2)$ time, delivering their output in less than two minutes, even for a tree with 3000 terminal taxa, thus posing virtually no computing constraint on the tree reconstruction.

### 5.6.4 Statistical efficiency in experiments

In addition to the 135-leaf tree of §5.6.2, we used three model trees in the simulation studies. The 500-leaf tree in Figure 5.34 has the topology of a seed plant phylogeny based on rbcL gene sequences from the Green Plant Phylogeny Project (Brown 1999) by Chase *et al.* (1993). The 1895-leaf in Figure 5.35 tree is derived from the evolutionary tree of Eukaryotes based on 12S sequences in the Ribosomal Database Project (Maidak *et al.* 2000). We removed a subtree containing distantly related taxa from the original tree of 2055 leaves. The 3135-leaf tree in Figure 5.36 is based on the subtree of Proteobacteria within the phylogeny of Prokaryotes in the Ribosomal Database Project. We scaled the edge lengths of the original trees using a linear transformation to evaluate the performance of various reconstruction methods. Specifically, for each simulation experiment we chose a maximum and minimum edge length $d_{\max}$ and $d_{\min}$, and calculated the values $c_0, 0 < c_1$ such that after replacing each edge length $d$ in the original tree by $c_0 d + c_1$, the edge lengths in the resulting tree range from $d_{\min}$ to $d_{\max}$. We used the Jukes-Cantor model for the DNA alphabet in all experiments.

In a set of experiments, we compared the accuracy of topology recovery algorithms as a function of sample length. The performance of HGT-ME and HGT-FP on the 135-leaf tree is compared to that of Neighbor-Joining in Figure 5.25. HGT-ME and HGT-FP perform slightly better starting at

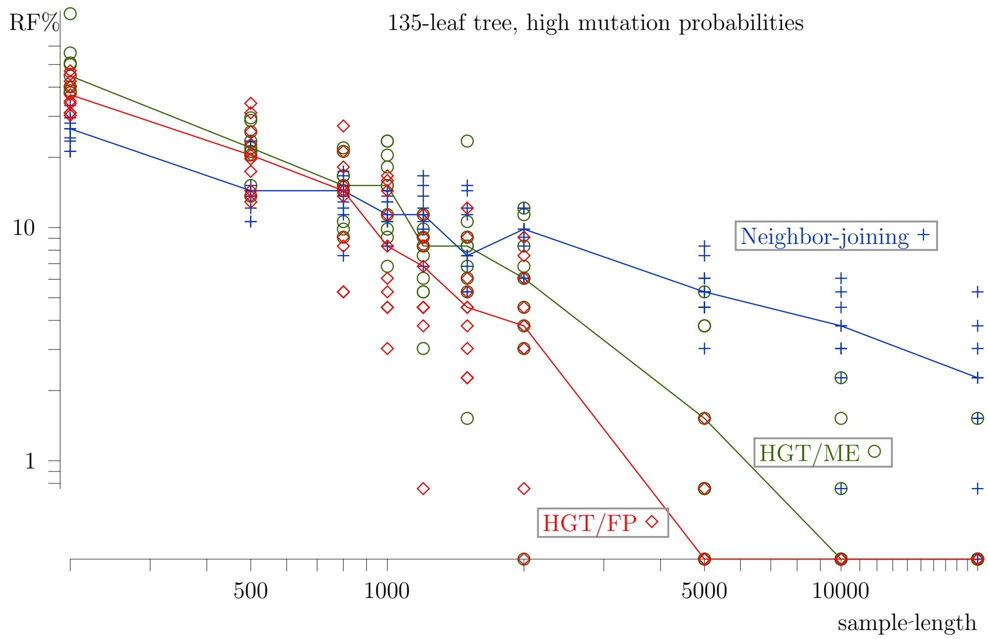



FIGURE 5.25: *Experimental results on the 135-leaf tree with edge mutation probabilities between 0.09 and 0.47. The plot shows the Robinson-Foulds error of the algorithms observed on ten separate samples for each length. The graphs go through the median values.*

sequence lengths of 1500 and converge faster to recover the topology than Neighbor-Joining.

As Figure 5.26 shows, our algorithms outperform Neighbor-Joining on the 500-leaf tree from around sample length $\ell = 1200$, and miss only 3% of the edges at $\ell = 2000$. It is worth pointing out that deriving the original tree took several months in computer time employing parsimony methods, while HGT-FP, HGT-ME, and Neighbor-Joining produce their output in a few seconds on a desktop computer.

Figure 5.27 shows the experimental results on the 1895-leaf tree. We scaled the edge lengths so that they fell into the interval $[0.1, 1.0]$. HGT-FP and HGT-ME converge quickly so that they miss only one edge on the majority of 2000 length samples, and steadily recover the topology from samples of length 5000. Neighbor-Joining's performance improves only threefold between sample lengths of 200 and 10000, and still misplaces 120–150

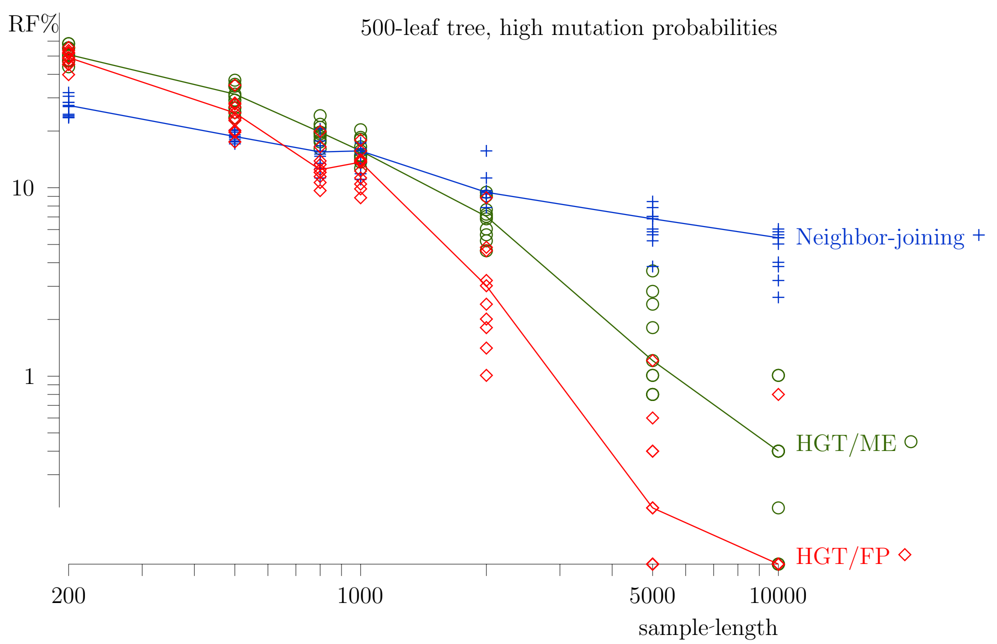



FIGURE 5.26: *Simulation results on the 500-leaf tree with edge mutation probabilities ranging between 0.07 and 0.47.*

edges at 5000 length sequences. We also conducted experiments on a differently scaled version of the 1895-leaf tree, in which the edge lengths were linearly mapped onto the $[0.01, 1.0]$ interval. Most of the edges in this tree are short, with around 60% of them having the shortest edge length corresponding to 0.007 mutation probability. The results of the experiments are shown in Figure 5.27. In accordance with previous findings in simulations with different smaller trees (e.g., Saitou and Imanishi (1989)), Neighbor-Joining performs well, achieving high success rates at relatively short sample sequences. HGT-FP and HGT-ME are more sensitive to short edge lengths due to the greedy selection of triplets lying at their core. At large sample sizes, however, they do converge more quickly than Neighbor-Joining on the highly divergent tree and misplace very few edges from $\ell = 5000$ on. The following simple example may shed some light on the philosophical differences underlying our algorithm and Neighbor-Joining. Let $\eta_{11}, \eta_{12}, \ldots, \eta_{1\ell}$ be a series of independent identically distributed random variables with unknown mean $m$ and variance $\sigma_1^2$. Let $\eta_{21}, \ldots, \eta_{2\ell}$ be another series of inde-

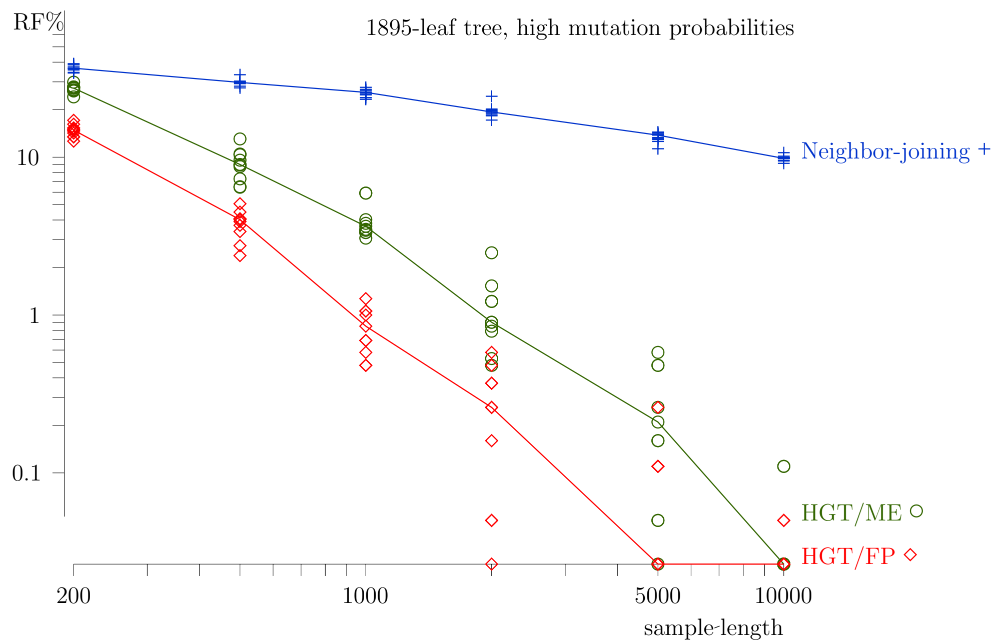



FIGURE 5.27: *Experimental results on the 1895-leaf tree with edge mutation probabilities ranging between 0.07 and 0.47.*

pendent identically distributed random variables with the same mean $m$ but larger variance $\sigma_2^2 > \sigma_1^2$. When $\sigma_1$ is close to $\sigma_2$, then $m$ is best estimated by $A = \ell^{-1} \sum_{i=1}^{\ell} (\eta_{1i} + \eta_{2i})/2$. However, if $\sigma_2 \gg \sigma_1$, then $B = \ell^{-1} \sum_{i=1}^{\ell} \eta_{1i}$ is a better estimator. The variance of $A$ equals $(\sigma_1^2 + \sigma_2^2)/(4\ell)$, while the variance of $B$ equals $\sigma_1^2/\ell$. Hence if $\sigma_2 > \sigma_1\sqrt{3}$, then $B$ has smaller variance than $A$. Neighbor-Joining, similarly to the estimator $A$, averages many estimated distances. When the edge mutation probabilities are small, the distances are small and do not differ by much, so the average provides more accurate information about the topology than the one obtained from a greedy approach. On the other hand, the error committed while calculating the average is governed by the error in the estimation of the largest distance in the expression, which may be significant when the mutation probabilities are large. As a result, the statistical performance of Neighbor-Joining is less stable, and a greedy algorithm may provide better efficiency in the case of large mutation probabilities.

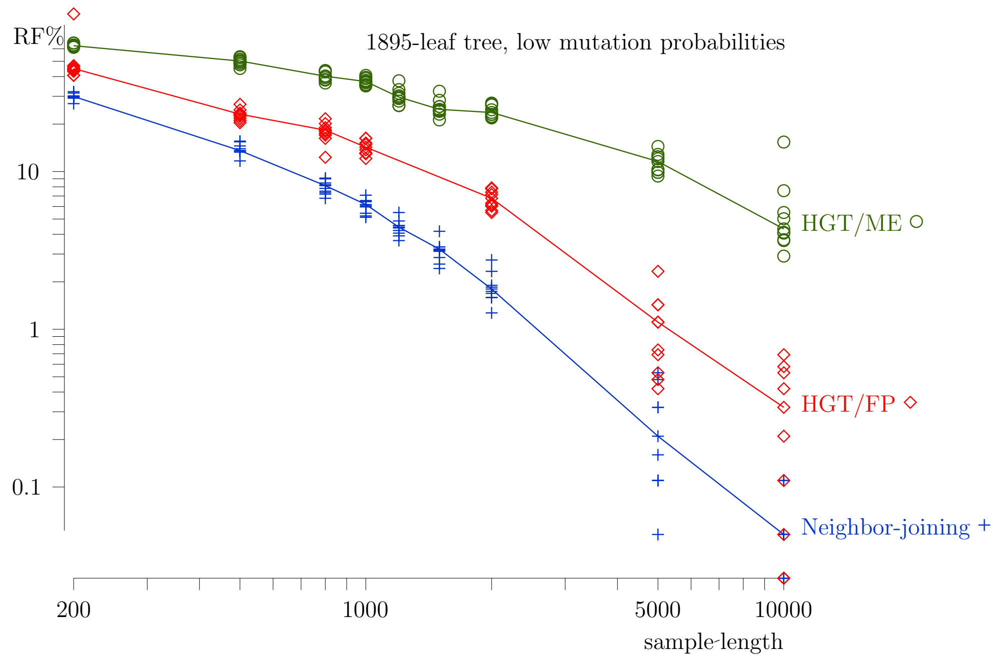



Figure 5.28: *Experimental results on the 1895-leaf tree with edge mutation probabilities ranging between 0.007 and 0.47.*

In order to further explore the effect that the range of mutation probabilities has on the recovery accuracy, we conducted another set of experiments for fixed sample lengths and variable scalings. In each experiment we scaled the edge lengths linearly to fall into the range $[d_{\min}, d_{\max}]$ where $d_{\min}$ is a function of $d_{\max}$. Specifically, we carried out two groups of experiments for every tree, in one set we selected $d_{\min} = d_{\max}/10$, and in the other set we chose $d_{\min} = d_{\max}/100$. The figures refer to the former as high mutation probabilities, and to the latter as low mutation probabilities.

Figure 5.29 compares the results of the experiments on the 500-leaf tree for UNJ, BioNJ, Neighbor-Joining, HGT-FP, and parsimony. Other algorithms considered in Figure 5.24 take hours if not days to build one tree on this size. In accordance with previous findings (Hillis 1996), parsimony performs very well, although when the maximum edge lengths are close to 1, its running time is increased to several hours. This phenomenon can be attributed to the fact that the sample sequences differ by much, and thus the optimization of the parsimony function becomes very difficult. It is worth pointing out that this tree, as well as some other trees in simulation studies (Hillis 1996; Rice and Warnow 1997) where parsimony performed well, were built using heuristic parsimony methods, so the simulation may have a certain bias in favor of parsimony. For both high and low mutation probabilities, BioNJ performs slightly better than Neighbor-Joining, and UNJ simply fails if the mutation probabilities are not small. HGT-FP performs better than Neighbor-Joining and BioNJ in the case of high mutation probabilities, while the Neighbor-Joining methods are better for low mutation probabilities, even though they do not recover the tree completely.

Figure 5.30 shows the experimental results on the 1895-leaf tree, for Neighbor-Joining, HGT-FP, and in case of high mutation probabilities, BioNJ. We omitted UNJ from the experiments here because it performs much worse than either Neighbor-Joining or BioNJ. In its defense we must mention that it was developed for input matrices in which the estimation error is uniformly distributed, which is not the true for distances computed from sequence data. We abandoned tracking the performance of BioNJ as its behavior is barely distinct from that of Neighbor-Joining. Parsimony is very slow for this tree size, taking hours or more to build a single tree. Again, HGT-FP outperforms Neighbor-Joining for high mutation probabilities, recovering the tree reliably for a large part of the region, whereas low mutation probabilities seem favorable for Neighbor-Joining. However, low mutation probabilities make the recovery more difficult and Neighbor-Joining needs

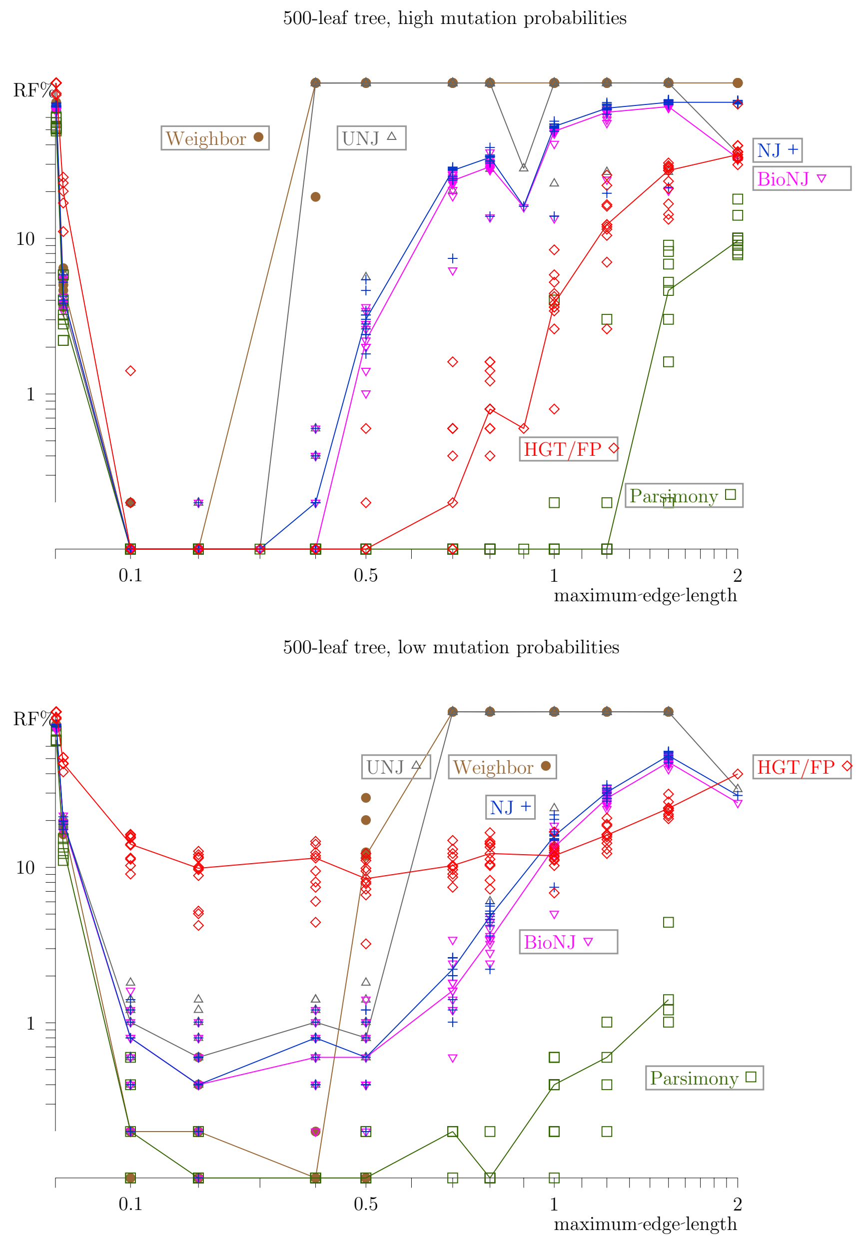
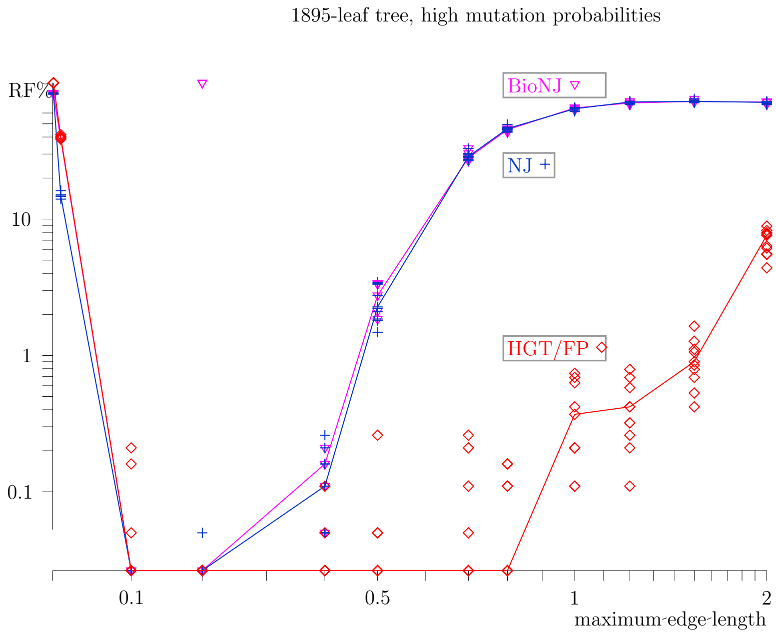



FIGURE 5.29: *Simulations on the 500-leaf tree. The plot shows the Robinson-Foulds error of the algorithms observed in ten separate 2000 bp long samples at each scaling value. The graphs go through the median values.*

longer sequences for accurate reconstruction.

Finally, Figure 5.31 shows our simulation results on the 3135-leaf tree for HGT-FP and Neighbor-Joining. The HGT-FP algorithm recovers the tree with large success rates for high mutation probabilities and performs better than Neighbor-Joining at part of the low mutation probability region. It is worth noticing that HGT-FP actually performs better on larger trees, while Neighbor-Joining does not take advantage of the richer information on the tree structure.



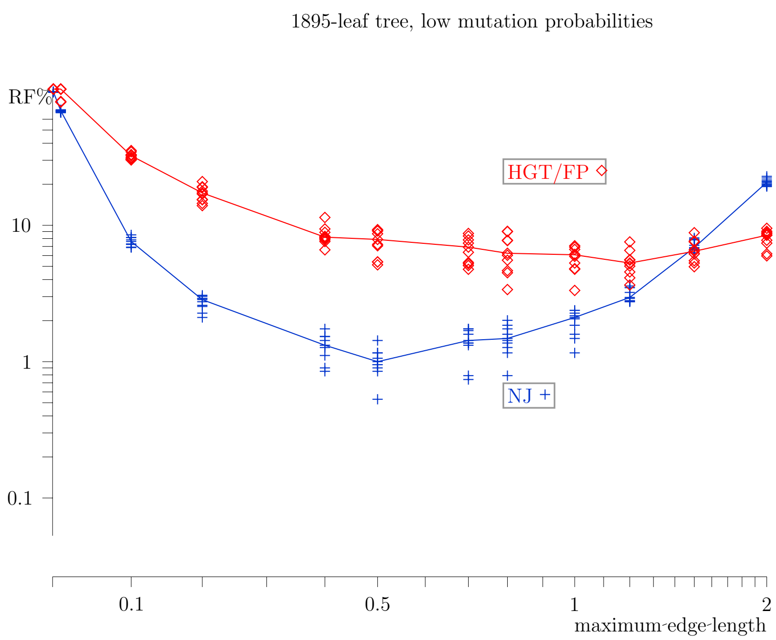
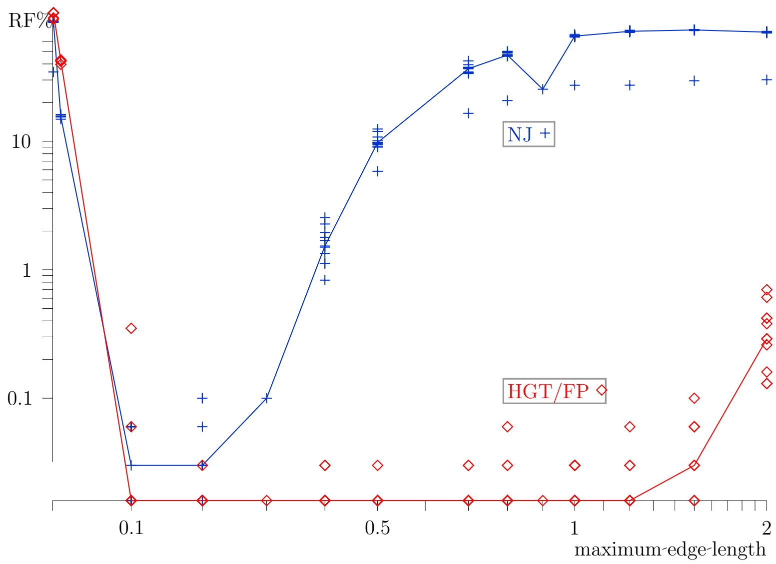




FIGURE 5.30: *Simulations on the 1895-leaf tree.*

3135-leaf tree, high mutation probabilities




3135-leaf tree, low mutation probabilities

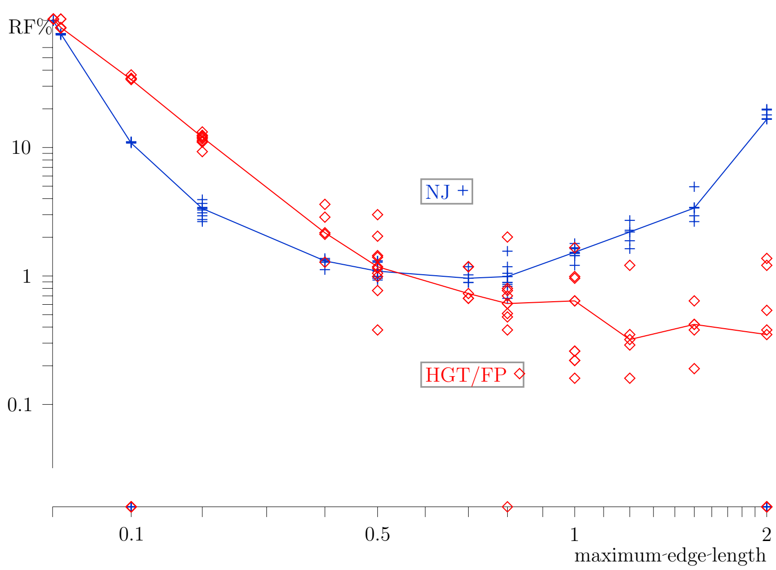



FIGURE 5.31: *Simulations on the 3135-leaf tree.*

## 5.A Proof of Lemma 5.2

Let

$$h_{uv} = \frac{\hat{S}_{uv}}{S_{uv}}, \qquad h_{uw} = \frac{\hat{S}_{uw}}{S_{uw}}, \qquad h_{vw} = \frac{\hat{S}_{vw}}{S_{vw}}.$$

Then

$$\mathbb{P}\left\{ \mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw) \geq \frac{-\ln(1-\epsilon)}{2} \right\} = \mathbb{P}\left\{ \frac{h_{uv}h_{uw}}{h_{vw}} \leq 1-\epsilon \right\}.$$

By conditioning on the events

$$\left\{ h_{uw} \leq 1-r \right\} \qquad \text{and} \qquad \left\{ h_{vw} \geq 1+s \right\}$$

for some $r, s > 0$,

$$\begin{aligned}
\mathbb{P}&\left\{ \frac{h_{uv}h_{uw}}{h_{vw}} \leq 1-\epsilon \right\} \\
&= \mathbb{P}\left\{ \frac{h_{uv}h_{uw}}{h_{vw}} \leq 1-\epsilon \;\middle|\; h_{uw} \leq 1-r \right\} \mathbb{P}\left\{ h_{uw} \leq 1-r \right\} \\
&\quad + \mathbb{P}\left\{ \frac{h_{uv}h_{uw}}{h_{vw}} \leq 1-\epsilon \;\middle|\; h_{uw} > 1-r \right\} \mathbb{P}\left\{ h_{uw} > 1-r \right\} \\
&\leq \mathbb{P}\left\{ h_{uw} \leq 1-r \right\} + \mathbb{P}\left\{ \frac{h_{uv}}{h_{vw}} \leq \frac{1-\epsilon}{1-r} \right\} \\
&= \mathbb{P}\left\{ h_{uw} \leq 1-r \right\} \\
&\quad + \mathbb{P}\left\{ \frac{h_{uv}}{h_{vw}} \leq \frac{1-\epsilon}{1-r} \;\middle|\; h_{vw} \geq 1+s \right\} \mathbb{P}\left\{ h_{vw} \geq 1+s \right\} \\
&\quad + \mathbb{P}\left\{ \frac{h_{uv}}{h_{vw}} \leq \frac{1-\epsilon}{1-r} \;\middle|\; h_{vw} < 1+s \right\} \mathbb{P}\left\{ h_{vw} < 1+s \right\} \\
&\leq \mathbb{P}\left\{ h_{uw} \leq 1-r \right\} + \mathbb{P}\left\{ h_{vw} \geq 1+s \right\} + \mathbb{P}\left\{ h_{uv} \leq (1-\epsilon)\frac{1+s}{1-r} \right\}.
\end{aligned}$$

Setting $\frac{1-r}{1+s} > 1-\epsilon$, by Equation 5.5,

$$\begin{aligned}\mathbb{P}\left\{\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw) \geq \frac{-\ln(1-\epsilon)}{2}\right\}& \\ \leq & a \exp\Big(-b\ell S_{uw}^2 r^2\Big) \qquad (*) \\ &+ a \exp\Big(-b\ell S_{vw}^2 s^2\Big) \\ &+ a \exp\left(-b\ell S_{uv}\left(1-(1-\epsilon)\frac{1+s}{1-r}\right)^2\right).\end{aligned}$$

Equating these exponential terms yields a second-order system of equations for $r$ and $s$. The solution for $r$ is

$$r = \frac{t - \sqrt{t^2 - y}}{2S_{uw}S_{vw}}$$

where

$$\begin{aligned} t &= S_{uv}S_{vw} + S_{uw}S_{vw} + (1-\epsilon)S_{uv}S_{uw}; \\ y &= 4S_{uv}S_{vw}^2S_{uw}\epsilon. \end{aligned}$$

By Taylor's expansion,

$$\left(t - \sqrt{t^2 - y}\right)^2 > \frac{y^2}{4t^2}.$$

Thus,

$$r^2 > \frac{\epsilon^2}{\left(\frac{1}{S_{uw}} + \frac{1-\epsilon}{S_{vw}} + \frac{1}{S_{uv}}\right)^2 S_{uw}^2} > \frac{\epsilon^2 S_{uvw}^2}{9S_{uw}^2}. \qquad (**)$$

Consequently, by Equations $(*)$ and $(**)$,

$$\begin{aligned}\mathbb{P}\left\{\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw) \geq \frac{-\ln(1-\epsilon)}{2}\right\}& \\ \leq 3a \exp\Big(-b\ell S_{uw}^2 r^2\Big) < 3a \exp\left(-\frac{b}{9}\ell S_{uvw}^2 \epsilon^2\right),&\end{aligned}$$

proving Equation (5.7).

## 5.B Proof of Lemma 5.8

We use the following basic inequalities.

$$\min\left\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\right\} \leq \frac{\hat{S}_{uvw}}{S_{uvw}} \leq \max\left\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\right\}; \tag{$*$}$$

$$\frac{S_{uvw}}{3} \leq \min\left\{S_{uv}, S_{uw}, S_{vw}\right\}. \tag{$**$}$$

We first prove Equation (5.11a). Pick $\lambda \geq 1$ such that $S_{uvw} = \lambda S_{\mathrm{lg}}$. Without loss of generality, we may suppose

$$\min\left\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\right\} = \frac{\hat{S}_{uv}}{S_{uv}}.$$

Then by Equations ($*$), ($**$), and (5.5a),

$$\begin{aligned}
\mathbb{P}\left\{\hat{S}_{uvw} \leq S_{\mathrm{md}}\right\} = \mathbb{P}&\left\{\frac{\hat{S}_{uvw}}{S_{uvw}} \leq \frac{S_{\mathrm{md}}}{\lambda S_{\mathrm{lg}}}\right\} \\
&\leq \mathbb{P}\left\{\frac{\hat{S}_{uv}}{S_{uv}} \leq \frac{S_{\mathrm{md}}}{\lambda S_{\mathrm{lg}}}\right\} \leq a\exp\left(-b\ell S_{uv}^2\left(1 - \frac{S_{\mathrm{md}}}{\lambda S_{\mathrm{lg}}}\right)^2\right) \\
&\leq a\exp\left(-\frac{b\left(1 - \frac{S_{\mathrm{md}}}{S_{\mathrm{lg}}}\right)^2}{9}\ell S_{\mathrm{lg}}^2\right).
\end{aligned}$$

By the choice of $S_{\mathrm{md}}$,

$$\frac{\left(1 - \frac{S_{\mathrm{md}}}{S_{\mathrm{lg}}}\right)^2}{9} = \frac{(\sqrt{2}-1)^2}{72},$$

and thus Equation (5.11a) holds.

We next prove Equation (5.11b) similarly. Pick $\lambda \leq 1$ such that $S_{uvw} =$

$\lambda S_{\mathrm{sm}}$. Without loss of generality, we may assume

$$\max\left\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\right\} = \frac{\hat{S}_{uv}}{S_{uv}}.$$

Then by Equations (∗), (∗∗), and (5.5b),

$$\begin{aligned}\mathbb{P}\left\{\hat{S}_{uvw} \geq S_{\mathrm{md}}\right\} = \mathbb{P}&\left\{\frac{\hat{S}_{uvw}}{S_{uvw}} \geq \frac{S_{\mathrm{md}}\sqrt{2}}{\lambda S_{\mathrm{lg}}}\right\} \\ &\leq \mathbb{P}\left\{\frac{\hat{S}_{uv}}{S_{uv}} \leq \frac{S_{\mathrm{md}}\sqrt{2}}{\lambda S_{\mathrm{lg}}}\right\} \leq a\exp\left(-b\ell S_{uv}^2\left(\frac{S_{\mathrm{md}}\sqrt{2}}{\lambda S_{\mathrm{lg}}}\right)^2\right) \\ &\leq a\exp\left(-\frac{b\left(\frac{S_{\mathrm{md}}\sqrt{2}}{S_{\mathrm{lg}}}\right)^2}{9}\ell S_{\mathrm{lg}}^2\right).\end{aligned}$$

By the choice of $S_{\mathrm{md}}$,

$$\frac{\left(\frac{S_{\mathrm{md}}\sqrt{2}}{S_{\mathrm{lg}}}\right)^2}{9} = \frac{(\sqrt{2}-1)^2}{72},$$

and thus Equation (5.11b) holds.

## 5.C Proof of Lemma 5.9

Since Lemma 5.2 can help establish only one half of the desired inequality, we split the probability on the left-hand side of Equation (5.13). Define

$$\begin{aligned}\hat{\Delta}_{uo} &= \mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) &\qquad \Delta_{uo} &= \mathrm{TC}(\mathbf{\Delta}, u, vw) \\ \hat{\Delta}_{vo} &= \mathrm{TC}(\hat{\mathbf{\Delta}}, v, uw) &\qquad \Delta_{vo} &= \mathrm{TC}(\mathbf{\Delta}, v, uw)\end{aligned}$$

Then

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\mathrm{TC}(\hat{\mathbf{\Delta}}, u, vw) - \mathrm{TC}(\mathbf{\Delta}, u, vw)\right| \geq \frac{\Delta_{\min}}{2}\right\} = \mathbb{P}\left\{\left|\hat{\Delta}_{uo} - \Delta_{uo}\right| \geq \frac{\Delta_{\min}}{2}\right\} \\
=&\mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{2}\right\} + \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \leq -\frac{\Delta_{\min}}{2}\right\} \\
\leq\ & \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{6}\right\} \\
&+ \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \leq -\frac{\Delta_{\min}}{2} \,\middle|\, \hat{\Delta}_{vo} - \Delta_{vo} \geq \frac{\Delta_{\min}}{6}\right\}\mathbb{P}\left\{\hat{\Delta}_{vo} - \Delta_{vo} \geq \frac{\Delta_{\min}}{6}\right\} \\
&+ \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \leq -\frac{\Delta_{\min}}{2} \,\middle|\, \hat{\Delta}_{vo} - \Delta_{vo} < \frac{\Delta_{\min}}{6}\right\}\mathbb{P}\left\{\hat{\Delta}_{vo} - \Delta_{vo} < \frac{\Delta_{\min}}{6}\right\} \\
\leq\ & \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{6}\right\} \\
&+ \mathbb{P}\left\{\hat{\Delta}_{vo} - \Delta_{vo} \geq \frac{\Delta_{\min}}{6}\right\} \\
&+ \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \leq -\frac{\Delta_{\min}}{2} \,\middle|\, \hat{\Delta}_{vo} - \Delta_{vo} < \frac{\Delta_{\min}}{6}\right\} \\
&.
\end{aligned}
$$

Then, since

$$
\hat{\mathbf{\Delta}}[u, v] - \mathbf{\Delta}[u, v] = \left(\hat{\Delta}_{uo} - \Delta_{uo}\right) - \left(\hat{\Delta}_{vo} - \Delta_{vo}\right),
$$

we have

$$
\begin{aligned}
&\mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \leq -\frac{\Delta_{\min}}{2} \,\middle|\, \hat{\Delta}_{vo} - \Delta_{vo} < \frac{\Delta_{\min}}{6}\right\} \\
&\qquad\qquad\qquad\qquad \leq \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u, v] - \mathbf{\Delta}[u, v] \leq -\frac{\Delta_{\min}}{3}\right\}.
\end{aligned}
$$

Consequently,

$$\begin{aligned}\mathbb{P}\left\{\left|\hat{\Delta}_{uo} - \Delta_{uo}\right| \geq \frac{\Delta_{\min}}{2}\right\} \leq\ & \mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{6}\right\} \\ & + \mathbb{P}\left\{\hat{\Delta}_{vo} - \Delta_{vo} \geq \frac{\Delta_{\min}}{6}\right\} \\ & + \mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \leq -\frac{\Delta_{\min}}{3}\right\}.\end{aligned} \tag{5.21}$$

By Equation (5.7),

$$\mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3a\exp\left(-b\frac{\vartheta^2}{9}\ell S_{uvw}^2\left(1 - e^{-\frac{\Delta_{\min}}{3}}\right)^2\right).$$

By Taylor's expansion,

$$\left(1 - e^{-\frac{\Delta_{\min}}{3}}\right)^2 \geq \left(1 - (1 - S_1)^{\frac{\vartheta}{3}}\right)^2 > \frac{\vartheta^2}{9}S_1^2,$$

and thus

$$\mathbb{P}\left\{\hat{\Delta}_{uo} - \Delta_{uo} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3a\exp\left(-b\frac{\vartheta^2}{81}\ell S_{uvw}^2 S_1^2\right). \tag{5.22}$$

By symmetry,

$$\mathbb{P}\left\{\hat{\Delta}_{vo} - \Delta_{vo} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3a\exp\left(-b\frac{\vartheta^2}{81}\ell S_{uvw}^2 S_1^2\right). \tag{5.23}$$

From Equation (5.5b),

$$\mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v] - \mathbf{\Delta}[u,v] \leq -\frac{\Delta_{\min}}{3}\right\} \leq \exp\left(-b\ell S_{uv}\left(e^{-\frac{\Delta_{\min}}{3}} - 1\right)^2\right).$$

Since

$$\frac{S_{uvw}}{3} \leq \min\left\{S_{uv}, S_{uw}, S_{vw}\right\},$$

and $S_{uv} > S_{\mathrm{sm}}$,

$$S_{uv} > \frac{S_{\mathrm{lg}}}{3\sqrt{2}}.$$

By Taylor's expansion,

$$\left(e^{\frac{\Delta_{\min}}{3}}-1\right)^2 \geq \left((1-S_1)^{-\frac{\vartheta}{3}}-1\right)^2 > \frac{\vartheta^2}{9}S_1^2.$$

Therefore,

$$\mathbb{P}\left\{\hat{\mathbf{\Delta}}[u,v]-\mathbf{\Delta}[u,v] \leq -\frac{\Delta_{\min}}{3}\right\} \leq \exp\left(-b\frac{\vartheta^2}{162}\ell S_{\lg}^2 S_1^2\right). \qquad (***)$$

Putting Equations (5.21), $(*)$, $(**)$, and $(***)$ together,

$$\mathbb{P}\left\{\left|\hat{\Delta}_{uo}-\Delta_{uo}\right| \geq \frac{\Delta_{\min}}{2}\right\} \leq 7a\exp\left(-b\frac{\vartheta^2}{162}\ell S_{\lg}^2 S_1^2\right),$$

which is tantamount to Equation (5.13).

## 5.D Proof of Lemma 5.13

By Lemma 5.7, for every node $z''$ strictly between $z$ and $z'$, there exists a leaf $w'' \notin \Psi_k^*$ with $S_{w''z''} \geq S_0^{\varrho_{\text{in}}+1}$. To choose $z''$, there are two cases: (1) both $z$ and $z'$ are inner nodes, and (2) $z$ or $z'$ is a leaf.

Case 1. By Lemma 5.4, let $\text{def}(z) = \{u, v, w\}$ and $\text{def}(z') = \{u, v', w'\}$. By $\mathcal{Y}_k$, neither $uvw$ nor $uv'w'$ is small. To fix the notation for $\text{def}(z)$ and $\text{def}(z')$ with respect to their topological layout, we assume without loss of generality that Figure 5.32 or equivalently the following statements hold.

- In $\Psi_k^*$ and thus in $\Psi$, $z'$ is on the paths between $z$ and $v'$, between $z$ and $w'$, and between $z$ and $v$, respectively.
- Similarly, $z$ is on the paths between $z'$ and $w$ and between $z'$ and $u$.
- $\Delta_{z'v'} \leq \Delta_{z'w'}$.

Both $uv'w''$ and $wv'w''$ define $z''$ and the target triplet is one of these two for some suitable $z''$. To choose $z''$, we further divide Case 1 into three subcases.

*Case* 1a: $S_{uz'} < S_{v'z'}S_0$ and $S_{v'z} < S_{uzS_0}$. The target triplet is $uv'w''$. Since $S_{uv'} \leq \sqrt{S_{uv'}}$, by Lemma 5.6 let $z''$ be a node on the path between $u$ and $v'$ in $\Psi$ with $\sqrt{S_{uv'}S_0} \leq S_{uz''} \leq \sqrt{S_{uv'}S_0^{-1}}$ and thus $\sqrt{S_{uv'}S_0} \leq S_{v'z''} \leq$

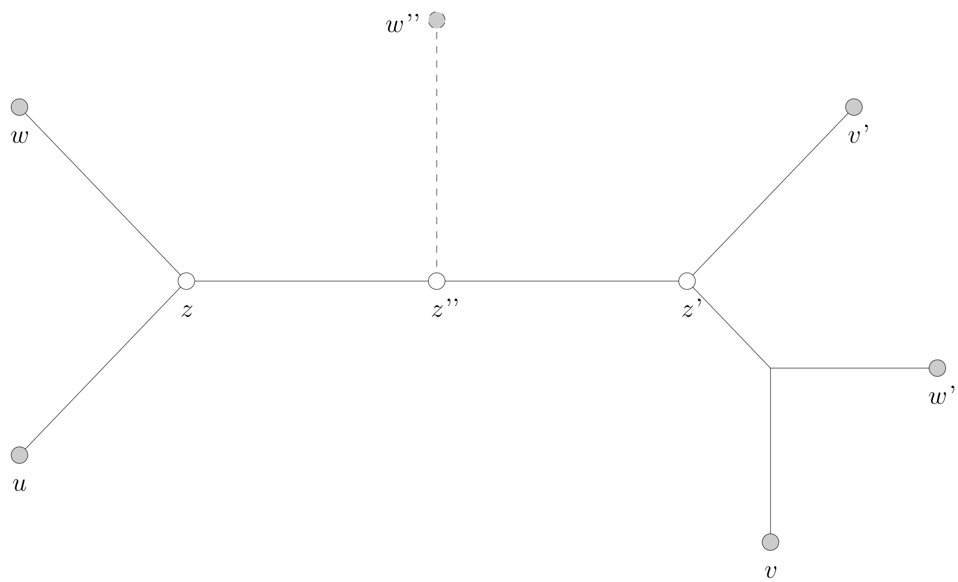



FIGURE 5.32: *This subgraph of $\Psi$ fixes some notation used in the proof of Lemma 5.13. The location of $v$ relative to $v'$ and $w'$ is not essential, for instance, $v$ can be even the same as $v'$. In $\Psi_k^*$,* $\mathrm{def}(z) = \{u, v, w\}$ *and* $\mathrm{def}(z') = \{u, v', w'\}$*. Neither $uvw$ nor $uv'w'$ is small, and $\Delta_{z'v'} \leq \Delta_{z'w'}$. We aim to prove that there is a leaf $w'' \notin \Psi_k^*$ such that $uv'w''$ or $wv'w''$ is large and defines a node $z''$ strictly between $z$ and $z'$.*

$\sqrt{S_{uv'} S_0^{-1}}$. By the condition of Case 1a, $z''$ is strictly between $z$ and $z'$ in $\Psi$. Also, by Lemma 5.3, $S_{uv'} \geq \frac{2}{3} S_{uv'w'}$. Thus, since $uv'w'$ is not small,

$$\begin{aligned} S_{uv'w''} &= \frac{3}{\frac{1}{S_{uz''}S_{z''w''}} + \frac{1}{S_{v'z''}S_{z''w''}} + \frac{1}{S_{uv'}}} \\ &\geq \frac{1}{\sqrt{\frac{2}{3}} S_{uv'w'}^{-1/2} {S_0}^{-\varrho_{\mathrm{in}}-3/2} + \frac{1}{2} S_{uv'w'}^{-1}} \\ &> S_{\mathrm{lg}}. \end{aligned} \tag{5.24}$$

So $uv'w''$ is as desired for Case 1a.

*Case* 1b: $S_{uz'} \geq S_{v'z'} S_0$. The target triplet is $uv'w''$. Let $z''$ be the first node after $z'$ on the path from $z'$ toward $z$ in $\Psi$. Then, $S_{v'z''} \geq S_{v'z'}S_0$. By Lemma 5.3, $S^2_{v'z''} \geq S_{uv'w'}S_0^2/3$. Next, since $S_{uv'} \geq S_{uw'}$ and $S_{z'v'} \geq S_{z'w'}$,

$$\begin{aligned} S_{uv'w'} &\leq \frac{3}{2S^{-1}_{uv'} + S^{-1}_{v'z'}S^{-1}_{z'w'}} \\ &\leq \frac{3}{2S^{-1}_{uz'}S^{-1}_{v'z'} + S^{-2}_{v'z'}} \\ &\leq \frac{3S^2_{uz'}}{2S_0 + S_0^2}. \end{aligned}$$

Thus, $S^2_{uz''} > S^2_{uz'} > S_{uv'w'}S_0^2$. Since $S_{uv'} \geq \frac{2}{3}S_{uv'w'}$ and $uv'w'$ is not small,

$$\begin{aligned} S_{uv'w''} &= \frac{3}{\frac{1}{S_{uz''}S_{z''w''}} + \frac{1}{S_{v'z''}S_{z''w''}} + \frac{1}{S_{uv'}}} \\ &> \frac{1}{\left(\frac{1+\sqrt{3}}{3}\right)S^{-1/2}_{uv'w'}S_0^{-\varrho_{\rm in}-2} + \frac{1}{2}S^{-1}_{uv'w'}} \\ &> S_{\rm lg}. \end{aligned} \tag{5.25}$$

So $uv'w''$ is as desired for Case 1b.

*Case* 1c: $S_{v'z} \geq S_{uz}S_0$. If $S_{wz} > S_{uz}$, the target triplet is $wv'w''$; otherwise, it is $uv'w''$. The two cases are symmetric, and we assume $S_{uz} \geq S_{wz}$. Let $z''$ be the first node after $z$ on the path from $z$ toward $z'$ in $\Psi$. Then, $S_{uz''} \geq S_{uz}S_0$. By Lemma 5.3, $S^2_{uz''} \geq S^2_{uz}S_0^2 \geq S_{uvw}S_0^2/3$. Since $S_{uv'} \geq S_{uw'}$ and $S_{v'w'} > 0$,

$$S_{uv'w'} < \frac{3}{2S^{-1}_{uv'}} \leq \frac{3}{2S^{-1}_{v'z}S^{-1}_{uz}} \leq \frac{3S^2_{v'z}}{2S_0}.$$

Hence $S^2_{v'z''} > S^2_{v'z} > 2S_{uv'w'}S_0/3$. Then, since neither $uv'w'$ nor $uvw$ is small and $S_{uv'} \geq \frac{2}{3}S_{uv'w'}$,

$$\begin{aligned} S_{uv'w''} &= \frac{3}{\frac{1}{S_{uz''}S_{z''w''}} + \frac{1}{S_{v'z''}S_{z''w''}} + \frac{1}{S_{uv'}}} \\ &> \frac{1}{\frac{1}{\sqrt{3}}S^{-1/2}_{uvw}S_0^{-\varrho_{\rm in}-2} + \frac{1}{\sqrt{6}}S^{-1/2}_{uv'w'}S_0^{-\varrho_{\rm in}-3/2} + \frac{1}{2}S^{-1}_{uv'w'}} \\ &> S_{\rm lg}. \end{aligned} \tag{5.26}$$

So $uv'w''$ is as desired for Case 1c with $S_{uz} \geq S_{wz}$.

*Case* 2. By symmetry, assume that $z' = u$ is a leaf in $\Psi_k^*$. Then, since $k \geq 3$, $z$ is an inner node in $\Psi_k^*$. Let $\mathrm{def}(z) = \{u, v, w\}$. By symmetry, further assume $S_{vz} \geq S_{wz}$. There are two subcases. If $S_{uz} < S_{vz}S_0$, then the proof is similar to that of Case 1a and the desired $z''$ is in the middle of the path between $u$ and $v$ in $\Psi$. Otherwise, the proof is similar that of Case 1b and $z''$ is the first node after $z$ on the path from $z$ toward $u$ in $\Psi$. In both cases, the desired triplet is $uvw''$.

## 5.E Trees used in the experiments

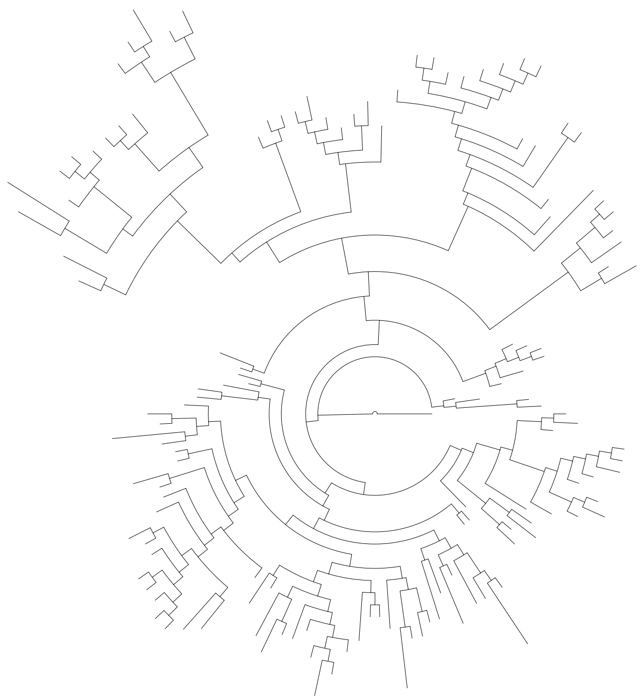


FIGURE 5.33: *The 135-leaf tree of Maddison* et al. *(1992) — a phylogeny of human populations*

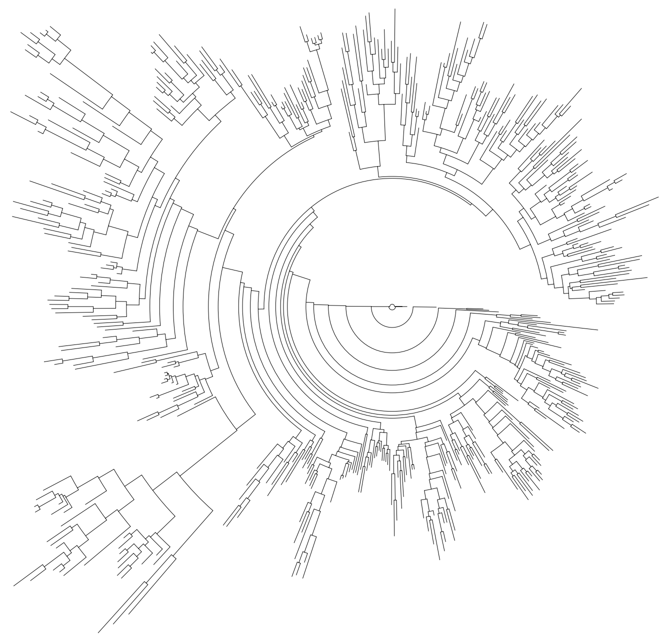


FIGURE 5.34: *The 500-leaf tree of Chase* et al. *(1993) — a phylogeny of green plants*

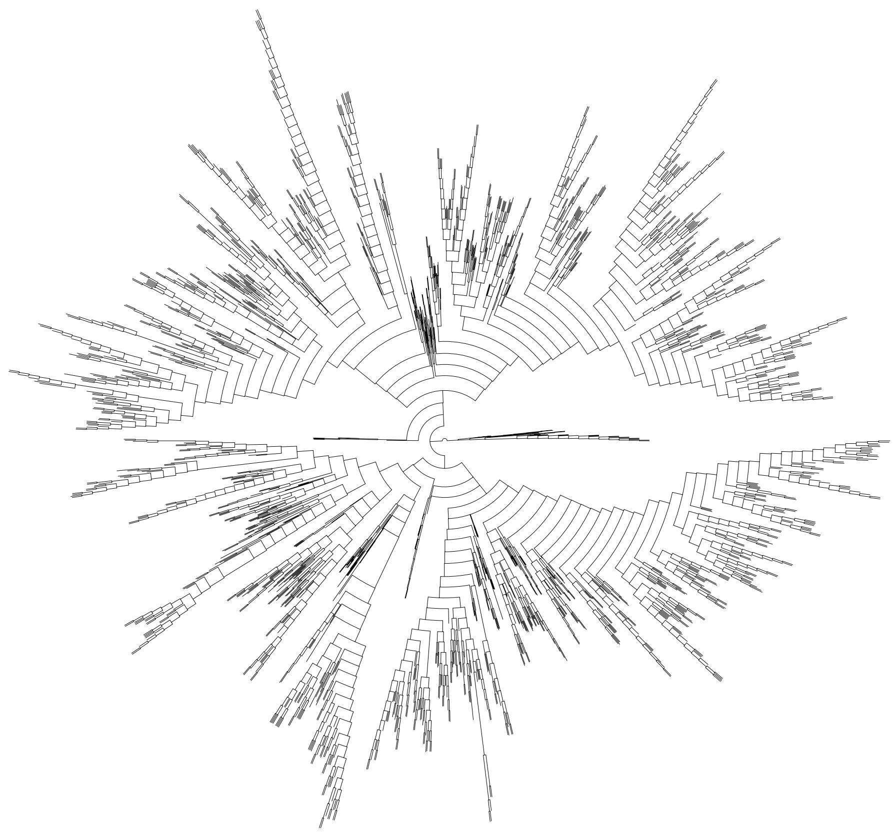


FIGURE 5.35: *The 1895-leaf tree based on the phylogeny of Eukaryotes (Maidak* et al. *2000)*

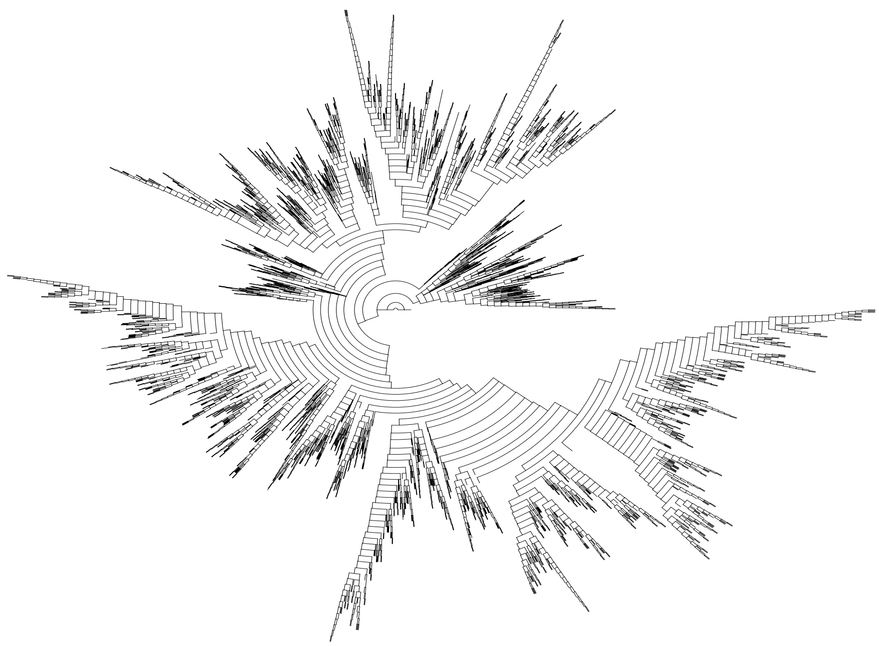


FIGURE 5.36: *The 3135-leaf tree based on the phylogeny of Proteobacteria (Maidak* et al. *2000)*

# Chapter 6

# Summary

Our comparative analysis of evolutionary tree reconstruction concludes by recapitulating what factors have made our results possible. We began our work by explaining the pertinence of using evolutionary trees in conjunction with biomolecular sequences such as genes and proteins. The first cursory overview of our context set the framework for viewing the construction of evolutionary trees from homologous sequences as a probabilistic learning problem. In particular, we stated the focal problem of our study as that of learning evolutionary tree topologies from the sample sequences they generate.

The viability of the Markov model for sequence evolution was then explored. Two features of evolution made our recourse to Markov models particularly appropriate: first, that evolution is "memoryless" — inheritance depends solely on the parents and not on the entire history of ancestors —, and secondly, that mutations along different evolutionary branches occur independently from each other at the molecular level. Considering how random taxon sequences form a Markov chain along any path in the tree, we offered an axiomatic definition of phylogeny, noting importantly that the distribution is fully determined by the root sequence distribution and the sequence transition probabilities. A particularly relevant implication of this definition is that the evolutionary tree topology is a function of the joint taxon sequence distribution, making topology recovery from sample sequences at least hypothetically possible. We then set out to investigate the i. i. d. Markov model in which the taxon sequence distribution is a product distribution of identical and independently distributed taxon labels corresponding to sequence characters. An important feature of the i. i. d. Markov model is that the number of parameters defining the phylogeny is finite, and sample sequences of increasing length convey an increasing amount of information about them. We presented a number of subclasses of the i. i. d. Markov model as natural extensions of commonly used nucleotide substitution rate models, accompanied by novel results on the closedness of corresponding transition matrix classes. We specifically discussed the Jukes-Cantor model, Kimura's two and three parameter models, the Hasegawa-Kishino-Yano model, and the Gojobori-Ishii-Nei model.

We further delved into the problem of topology recovery by discussing the nature of evolutionary distances and similarities, where we defined distance as the logarithm of similarity. By treating distances as functionals of distributions over sequence pairs, we defined several axiomatic properties that evolutionary distances possess, such as additivity along paths and symmetry. We presented the Jukes-Cantor distance, Kimura's three parameter distance, and the paralinear distance, and proved that they exhibit the properties of evolutionary distances. Further, we stated our novel result concerning the uniqueness of evolutionary distances, specifically, that evolutionary distance functions differ by only a constant factor in time-reversible mutation models with constant substitution rates. The additive property of evolutionary distances was particularly important for our purposes, since it implied that topologies could be recovered from distances between sample sequences. This recognition led us to scrutinize methods for estimating evolutionary distances from finite sample sequences. We derived novel upper bounds on the probabilities of large deviations in the cases of Jukes-Cantor distance, Kimura's three parameter distance, and paralinear distance. In each case we showed that the tail probabilities decrease exponentially with the sequence length and the square of the similarities between the sequences involved.

We examined existing algorithmic approaches to evolutionary tree topology reconstruction. We defined computational efficiency as polynomial running time in tree size, and statistical efficiency as successful topology recovery from polynomially long sequences. Both efficiency requirements are essential for recovery of large trees with hundreds or thousands of nodes. We offered a comprehensive overview of maximum-likelihood, character-based, and distance-based algorithms. We noted that exact optimization algorithms that select their output by minimizing a penalty function inevitably address to NP-hard problems, and are not computationally efficient. This difficulty is encountered with maximum-likelihood and character-based methods, as well as with numerical taxonomy- and minimum evolution-related distance-based algorithms. We pointed out the lack of statistical efficiency in the case of character-based methods due to their statistical inconsistency. We described existing theoretical guarantees for successful topology recovery, specifically, the three-point and four-point conditions. We presented the LogDet metric, which is not an evolutionary distance according to our definition, but satisfies the four-point condition, and can thus serve as a basis for topology recovery with a distance-based algorithm. We analyzed the convergence speed of the LogDet metric estimated from sample sequences. In particular, we derived upper bounds on tail probabilities of the estimation error in a similar form to our upper bounds for empirical evolutionary distances. We used our error bounds to extend existing results on the sample length requirements for distance-based algorithms to recover the topology. We remarked that the sample length bounds for popular distance-base algorithms, including Neighbor-Joining, are generally exponential in tree size.

Given the computational and statistical inadequacies of most existing algorithms, we designed a family of novel distance-based methods satisfying the criteria for statistical and computational efficiency. Our algorithms build evolutionary trees by using triplets of leaves. The algorithms are based on the "Harmonic Greedy Triplets" principle, which originates from our result that in the case of the studied evolutionary distances and the LogDet metric, the error committed in estimating the triplet centers depends on the harmonic average of pairwise similarities between the triplet members. We presented the Basic-HGT and Fast-HGT algorithms, where the former runs in cubic time, and the latter in quadratic time in the number of tree nodes. The algorithms use an input parameter that determines the shortest distance between tree nodes. We presented another quadratic-time algorithm, called HGT-FP, which uses the four-point condition, and eliminates the need for the minimum distance input parameter. We proved that all three algorithms are statistically efficient, and the sample length bounds for the first two match the best asymptotic bounds of other statistically efficient algorithms. In fact, our algorithms are the only known topology recovery algorithms that are provably statistically efficient and run in cubic or quadratic time. Based on simulation experiments, we offered a heuristic way of setting the minimum distance parameter of Fast-HGT by employing the minimum evolution principle. The resulting algorithm, called HGT-ME, runs in $O(n^2 \log \ell)$ time for a tree with $n$ leaves and sample sequences of length $\ell$. We compared the computational and statistical efficiency of our algorithms to the efficiency of many existing methods in simulated experiments. Our goal in the experiments was to evaluate the algorithms' appropriateness for large-scale phylogeny reconstruction. Running time measurements of existing implementations showed the superior speed of HGT-FP requiring a few seconds to reconstruct trees with thousands of leaves on a desktop computer, in contrast to several minutes, hours, or even days in the case of other algorithms. We compared the success of topology recovery between the HGT-FP, HGT-ME, heuristic parsimony, and several distance-based algorithms, such as Neighbor-Joining. In the experiments we simulated sequence evolution in the Jukes-Cantor model along biologically motivated trees with 135, 500, 1895, and 3135 leaves, with varying mutation probabilities and sample sequence lengths. In summary, we found that heuristic parsimony performs very well, but its slow speed hinders its use with large trees; that Neighbor-Joining is viable in the case of small mutation probabilities but still fails to recover about 1% of the edges from realistic sample lengths, and performs poorly when mutation probabilities are large; and that HGT-FP achieves high success rates when mutation probabilities are large, and fails to recover 5–7 times as many edges as Neighbor-Joining when mutation probabilities are small. Moreover, HGT-FP tends to achieve even higher success rates as the tree size increases. The theoretical results and the simulation experiments show that large-scale phylogeny recovery is feasible with distance-based methods in Markov models of evolution, and our HGT-FP algorithm is particularly useful where other distance-based methods fail. The success of our algorithms is attributable to the greedy selection lying at their core. They do not aim to optimize any explicit penalty function but strive to recover the topology as correctly as possible. As a result, they avoid theoretical and experimental weaknesses of optimization methods.

The path we followed in our dissertation led from molecular sequences and mathematical sequence evolution models, to the design of algorithms with superior efficiency within these models. We hope that in the future we will be able to close this conceptual circle, and that the algorithms will prove useful for molecular evolutionary studies based on biomolecular sequences.

# Notations and abbreviations

## Notations

$\mathbb{I}\{\cdot\}$ indicator variable

$\mathcal{G}$ graph (page 7)

$\mathcal{T}$ tree (page 8)

$V$ set of all vertices or taxa

$L$ subset of all vertices, usually the set of leaves

$E$ set of all edges

$u, v, w, z$ vertices or taxa

$\simeq$ graph isomorphism (page 10)

$\underset{L}{\simeq}$ graph isomorphism with equality on node set $L$ (page 10)

$\mathcal{A}, \mathcal{A}^+$ alphabet, set of positive length sequences

$m = |\mathcal{A}|$ alphabet size

$s, t$ sequences

$\mathcal{S}$ language: possible values of taxon sequences

$X^{(u)}$ random taxon sequence associated with node $u$ ( page 10)

$X_i^{(u)}$ $i$-th character of the random taxon sequence

$\mathcal{P}$ evolutionary tree (page 10)

$\Psi(\mathcal{P})$ topology — graph obtained from $\mathcal{P}$ by removing the direction of the edges (page 11)

$\mathcal{C}$ hypothesis class of evolutionary trees (page 13)

$\mathbf{M}_e$ edge mutation matrix (page 20)

$\mathbf{M}_{uv}^{(k)}$ mutation matrix for $k$-th sequence position (page 21)

$\mathbf{M}_{uv}$ mutation matrix in i. i. d. Markov model (page 25)

$\mathcal{M}$ class of mutation matrices

$\xi^{(u)}$ random taxon label taking values in $\mathcal{A}$ (page 25)

$\boldsymbol{\pi}^{(0)}$ root symbol distribution in i. i. d. Markov model (page 22)

$\boldsymbol{\pi}^{(u)} = \langle \pi_1^{(u)}, \ldots, \pi_m^{(u)} \rangle$ taxon label distribution (page 25)

$\mathbf{Q}$ constant substitution rate matrix (page 26)

$\tau$ evolutionary time

$\ell$ sequence length in i. i. d. Markov model

$S$ similarity (page 47)

$D$ distance (page 47)

$\mathbf{J}_{uv}$ matrix of joint distribution for nodes $u, v$ (page 59)

$\hat{S}$ empirical similarity

$\hat{D}$ empirical distance

$\hat{\mathbf{M}}$ empirical mutation matrix (page 77)

$\varrho_{\mathrm{in}}, \varrho_{\mathrm{out}}$ inner and outer radius of a tree (see page 115)

$\boldsymbol{\Delta}$ distance matrix, tree metric (page 95)

$\hat{\boldsymbol{\Delta}}$ estimated distance matrix (page 95)

${}^{u}\searrow_{v}$ node $v$ is in the right subtree of node $u$ (page 126)

${}_{v}\swarrow^{u}$ node $v$ is in the left subtree of node $u$ (page 126)

$\overset{v}{\underset{u}{\uparrow}}$ node $v$ is neither in the left not in the right subtree of node $u$ (page 126)

RF% Robinson-Foulds error (page 164)

# Abbreviations

i. i. d. independent identically distributed (page 22)

PAM point accepted mutation (page 26)

JC Jukes-Cantor model

K2P Kimura's two parameter model

K3P Kimura's three parameter model

TK Takahata-Kimura model

HKY Hasegawa-Kishino-Yano model

GIN Gojobori-Ishii-Nei model

UNJ Unweighted Neighbor Joining algorithm

# Bibliography


Agarwala, R., V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup (1999). On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing 28*, 1073–1085. Preliminary version at SODA '96.

Agarwala, R. and D. Fernández-Baca (1994). A polynomial time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM Journal on Computing 23*, 1216–1224. Also available as DIMACS TR93-03.

Albert, V. A., B. D. Mishler, and M. W. Chase (1992). Character-site weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. In P. S. Soltis, D. E. Soltis, and J. J. Doyle (Eds.), *Molecular Systematics of Plants*, Chapter 16, pp. 369–403. New York: Chapman and Hall.

Alon, N. and J. H. Spencer (1992). *The Probabilistic Method*. New York: John Wiley & Sons.

Ambainis, A., R. Desper, M. Farach, and S. Kannan (1997). Nearly tight bounds on the learnability of evolution. In *38th Annual Symposium on Foundations of Computer Science*, pp. 524–533. IEEE.

Atteson, K. (1997). The performance of neighbor-joining algorithms of phylogeny reconstruction. See Jiang and Lee (1997), pp. 101–110.

Avise, J. C. and K. Wollenberg (1997). Phylogenetics and the origin of species. *Proceedings of the National Academy of Sciences of the USA 94*, 7748–7755.

Ayala, F. J. (1997). Vagaries of the molecular clock. *Proceedings of the National Academy of Sciences of the USA 94*, 7776–7783.

Bandelt, H.-J. (1990). Recognition of tree metrics. *SIAM Journal on Discrete Mathematics 3*, 3–6.

Bandelt, H.-J. and A. Dress (1986). Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics 7*, 309–343.

Barry, D. and J. A. Hartigan (1987). Asynchronous distance between homologous DNA sequences. *Biometrics 43*, 261–276.

Barthélemy, J.-P. and A. Guénoche (1991). *Trees and Proximity Representations*. New York: John Wiley & Sons.

Berry, V. and D. Bryant (1999). Faster reliable phylogenetic analysis. See Istrail, Pevzner, and Waterman (1999), pp. 59–68.

Berry, V. and O. Gascuel (1997). Inferring evolutionary trees with strong combinatorial evidence. See Jiang and Lee (1997), pp. 111–123.

Bodlaender, H. L., M. R. Fellows, and T. Warnow (1992). Two strikes against perfect phylogeny. In W. Knieh (Ed.), *Automata, Languages and Programming, 19th International Colloquium*, Volume 623 of *Lecture Notes in Computer Science*, Berlin, pp. 273–283. Springer-Verlag.

Bollobás, B. (1979). *Graph Theory: an Introductory Course*. New York: Springer-Verlag.

Bollyky, P. L. and E. C. Holmes (1999). Reconstructing the complex evolutionary history of the hepatitis B virus. *Journal of Molecular Evolution 49*, 130–141.

Bondy, J. A. and U. S. R. Murty (1976). *Graph Theory with Applications*. New York: Elsevier Science.

Bonfield, J. K. and R. Staden (1995). The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Research 23*, 1406–1410.

Brown, J. K. M. (1994). Probabilities of evolutionary trees. *Systematic Biology 43*, 78–91.

Brown, K. S. (1999). Deep Green rewrites evolutionary history of plants. *Science 285*, 990.

Brown, W. M. (1985). The mitochondrial genome of animals. In R. J. MacIntyre (Ed.), *Molecular Evolutionary Genetics*, pp. 95–124. New York: Plenum Press.

Bruno, W. J., N. D. Socci, and A. L. Halpern (2000). Weighted Neighbor-Joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution 17*, 189–197.

Bryant, D. and P. Waddell (2000). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution 15*, 1346–1359.

Brzustowski, J. (1998). *qclust V0.2*. ⟨http://www.biology.ualberta.ca/jbrzusto/⟩.

Bulmer, M. (1991). Use of the method of generalized least-squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution 8*, 868–883.

Buneman, P. (1971). The recovery of trees from dissimilarity matrices. In F. R. Hodson, D. G. Kendall, and P. Tautu (Eds.), *Mathematics in the Archaelogical and Historical Sciences: Proceedings of the Anglo-Romanian Conference, Mamaia, 1970*, Edinburgh, pp. 387–395. Edinburgh University Press.

Camin, J. H. and R. R. Sokal (1965). A method for deducing branching sequences in phylogeny. *Evolution 19*, 311–326.

Cavalli-Sforza, L. L. and A. W. H. Edwards (1967a). Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics 19*, 233–267. Reprinted as (Cavalli-Sforza and Edwards 1967b).

Cavalli-Sforza, L. L. and A. W. H. Edwards (1967b). Phylogenetic analysis models and estimation procedures. *Evolution 32*, 550–570.

Cavender, J. (1978). Taxonomy with confidence. *Mathematical Biosciences 40*, 271–280.

Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences 137*, 51–73.

Chang, J. T. and J. A. Hartigan (1991). Reconstruction of evolutionary trees from pairwise distributions on current species. In E. M. Keramidas (Ed.), *Computing Science and Statistics: Proceedings of the Second Symposium on the Interface*, pp. 254–257.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron,

J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrn, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. M. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert (1993). Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden 80*, 528–580.

Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics 23*, 493–507.

Cohen, J. and M. Farach (1997). Numerical taxonomy on data: experimental results. *Journal of Computational Biology 4*, 547–558.

Colonius, H. and H.-H. Schulze (1981). Tree structure for proximity data. *British Journal of Mathematical & Statistical Psychology 34*, 167–180.

Cormen, T. H., C. E. Leiserson, and R. D. Rivest (1990). *Introduction to Algorithms*. Cambridge, Mass: MIT Press.

Cryan, M., L. A. Goldberg, and P. W. Goldberg (1998). Evolutionary trees can be learned in polynomial time in the two-state general Markov-model. Technical Report RR347, Department of Computer Science, University of Warwick, UK. Preliminary version at FOCS '98.

Csűrös, M. and M.-Y. Kao (1999). Recovering evolutionary trees through Harmonic Greedy Triplets. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Baltimore, MD, pp. 261–270.

Culberson, J. and P. Rudnicki (1989). A fast algorithm for constructing trees from distance matrices. *Information Processing Letters 30*, 215–220.

Cunningham, J. P. (1978). Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology 17*, 165–188.

Day, W. H. E. (1983a). Computationally difficult parsimony problems in phylogenetic systematics. *Journal of Theoretical Biology 103*, 429–438.

Day, W. H. E. (1983b). Distributions of distances between pairs of classifications. In J. Felsenstein (Ed.), *Numerical Taxonomy*, New York, pp. 127–131. NATO ASI: Springer.

Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology 49*, 461–467.

Day, W. H. E., D. S. Johnson, and D. Sankoff (1986). The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences 81*, 33–42.

Day, W. H. E. and D. Sankoff (1986). Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology 35*, 224–229.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure*, Volume 5, pp. 345–352.

de Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika 48*, 621–626.

de Soete, G., J. D. Carroll, and W. S. DeSarbo (1987). Least squares algorithms for constructing ultrametric and additive tree representations of symmetric proximity data. *Journal of Classification 4*, 155–173.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Dress, A. and M. A. Steel (1993). Convex tree realizations of partitions. *Advances in Applied Mathematics 5*, 3–6.

Edwards, A. W. F. and L. L. Cavalli-Sforza (1963). Reconstructing evolutionary trees. In V. H. Heywood and J. McNeill (Eds.), *Phenetic and phylogenetic classification*, Volume Publication No. 6, pp. 67–76. London: Systematics Association.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research 8*, 163–167.

Erdős, P. L., K. Rice, M. A. Steel, L. A. Székely, and T. Warnow (1998). The Short Quartet Method. *Mathematical Modeling and Scientific Computing*. to appear.

Erdős, P. L., M. Steel, L. A. Székely, and T. J. Warnow (1997). Constructing big trees from short sequences. In P. Degano, R. Gorrieri, and A. Marchetti-Spaccamela (Eds.), *Automata, Languages and Programming, 24th International Colloquium*, Volume 1256 of *Lecture Notes in Computer Science*, pp. 827–837. Springer-Verlag. Also available as DIMACS TR97-17.

Erdős, P. L., M. A. Steel, L. A. Székely, and T. J. Warnow (1999a). A few logs suffice to build (almost) all trees (I). *Random Structures and Algorithms 14*, 153–184. Preliminary version as DIMACS TR97-71.

Erdős, P. L., M. A. Steel, L. A. Székely, and T. J. Warnow (1999b). A few logs suffice to build (almost) all trees (II). *Theoretical Computer Science 221*, 77–118. Preliminary version as DIMACS TR97-72.

Estabrook, G. (1972). Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history. *Annual Review of Ecology and Systematics 3*, 427–456.

Estabrook, G. F., C. S. Johnson, Jr., and F. R. McMorris (1975). An idealized concept of the true cladistic character. *Mathematical Biosciences 23*, 263–272.

Ewing, B., L. Hillier, M. C. Wendt, and P. Green (1998). Base-calling of automated sequencer traces using Phred. I. accuracy assessment. *Genome Research 8*, 175–185.

Farach, M. and S. Kannan (1999). Efficient algorithms for inverting evolution. *Journal of the ACM 46*, 437–449. Preliminary version at STOC '96.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Zoology 19*, 83–92.

Farris, J. S. (1972). Estimating phylogenetic trees from dissimilarity matrices. *The American Naturalist 106*, 645–668.

Farris, J. S. (1973). A probability model for inferring evolutionary trees. *Systematic Zoology 22*, 250–256.

Farris, J. S. (1977). Phylogenetic analysis under Dollo's law. *Systematic Zoology 26*, 77–88.

Farris, J. S. (1988). *Hennig86 version 1.5. Distributed by the author.* Port Jefferson, NY.

Felsenstein, J. (1973). Maximum likelihood and minimum-step method for estimating evolutionary trees from data on discrete characters. *Systematic Zoology 22*, 240–249.

Felsenstein, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology 22*, 240–249.

Felsenstein, J. (1978b). The number of evolutionary trees. *Systematic Zoology 27*, 27–33.

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

Felsenstein, J. (1983). Statistical inference of phylogenies. *Journal of the Royal Statistical Society Series A 146*, 246–272.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics 22*, 521–565.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author*. Seattle, Wash.: University of Washington Department of Genetics.

Felsenstein, J. (2000). Phylogeny programs. 17 January 2000. ⟨http://evolution.genetics.washington.edu/phylip/software.html⟩.

Fernández-Baca, D. and J. Lagergren (1998). On the approximability of the Steiner tree problem in phylogeny. *Discrete Applied Mathematics 88*, 129–145.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum changes for a specific tree topology. *Systematic Zoology 20*, 406–416.

Fitch, W. M. and E. Margoliash (1967). Construction of phylogenetic trees. *Science 155*, 279–284.

Foulds, L. R. and R. L. Graham (1982). The Steiner tree problem in phylogeny is NP-complete. *Advances in Applied Mathematics 3*, 43–49.

Galtier, N., N. Tourasse, and M. Gouy (1999). A nonhyperthermophylic common ancestor to extant life forms. *Science 283*, 220–221.

Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn (1999). Origin of HIV-1 in the chimpanzee *pan troglodytes troglodytes*. *Nature 397*, 436—441.

Gascuel, O. (1994). A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology and Evolution 11*, 961–963.

Gascuel, O. (1997a). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution 14*(7), 685–695.

Gascuel, O. (1997b). Concerning the NJ algorithm and its unweighted version, UNJ. In B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky (Eds.), *Mathematical Hierarchies and Biology*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 149–170. Providence, RI: American Mathematical Society.

Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular Biology and Evolution 17*, 401–405.

Gojobori, T., K. Ishii, and M. Nei (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Journal of Molecular Evolution 18*, 414–423.

Goloboff, P. A. (1996). Methods for faster parsimony analysis. *Cladistics 12*, 199–220.

Goodman, M. (1976). Protein sequences in phylogeny. In F. J. Ayala (Ed.), *Molecular Evolution*, pp. 141–159. Sunderland, Mass.: Sinauer Associates.

Graham, R. L. and L. R. Foulds (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Mathematical Biosciences 60*, 133–142.

Gu, X. and W.-H. Li (1996). Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. *Molecular Biology and Evolution 13*, 1375–1383.

Gusfield, D. (1984). The Steiner tree problem in phylogeny. Technical Report TR-334, Yale University Department of Computer Science.

Gusfield, D. (1991). Efficient algorithms for inferring evolutionary history. *Networks 21*, 19–28.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Hakimi, S. L. and S. S. Yau (1964). Distance matrix of a graph and its realizability. *Quarterly of Applied Mathematics 22*, 305–317.

Harding, E. F. (1971). The probabilities of rooted tree shapes generated by random bifurcation. *Advances in Applied Probability 3*, 44–77.

Hartigan, J. (1973). Minimum mutation fits to a given tree. *Biometrics 29*, 53–65.

Hartigan, J. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*, 160–174.

Hein, J. (1989). An optimal algorithm to reconstruct trees from additive distance data. *Bulletin of Mathematical Biology 51*, 597–603.

Hendy, M. D. and D. Penny (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences 59*, 277–290.

Hendy, M. D. and D. Penny (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology 38*, 297–309.

Henikoff, S. and J. G. Henikoff (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA 89*, 10915–10919.

Hennig, W. (1950). *Grundzüge einer Theorie der phylogenetischen Systematik*. Berlin: Deutscher Zentralverlag.

Hennig, W. (1966). *Phylogenetic Systemtics*. Chicago, Ill.: University of Illinois Press. English translation of (Hennig 1950).

Hillis, D. M. (1995). Approaches for assessing phylogenetic accuracy. *Systematic Biology 44*, 3–16.

Hillis, D. M. (1996). Inferring complex phylogenies. *Nature 383*, 130–131.

Hillis, D. M. (1997). Biology recapitulates phylogeny. *Science 276*, 218–219.

Hillis, D. M., J. P. Huelsenback, and C. W. Cunningham (1994). Application and accuracy of molecular phylogenies. *Science 264*, 671–677.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Hubert, L. and P. Arabie (1995). Iterative projection strategies for the least squares fitting of tree structures to proximity data. *Psychometrika 48*, 281–317.

Hunter, L. (1999). Molecular biology for computer scientists. In L. Hunter (Ed.), *Artificial Intelligence and Molecular Biology*, pp. 1–45. AAAI Press.

Huson, D., S. Nettles, K. Rice, T. Warnow, and S. Yooseph (1998). Hybrid tree reconstruction methods. In K. Mehlhorn (Ed.), *Proceedings of the Second Workshop on Algorithm Engineering*, Saarbrücken, Germany, pp. 1–15. Max-Planck-Institut für Informatik. ⟨http://www.mpi-sb.mpg.de/~wae98/⟩.

Huson, D. H., S. Nettles, and T. J. Warnow (1999). Obtaining highly accurate topology estimates of evolutionary trees from very short sequences. See Istrail, Pevzner, and Waterman (1999), pp. 198–309.

Istrail, S., P. Pevzner, and M. Waterman (Eds.) (1999). *Proceedings of the Third Annual International Conference on Computational Biology*, Lyon. ACM Press.

Jiang, T. and D. T. Lee (Eds.) (1997). *Computing and Combinatorics, Third Annual International Conference*, Volume 1276 of *Lecture Notes in Computer Science*, Berlin. Springer-Verlag.

Jin, L. and M. Nei (1991). Relative efficiencies of the maximum-parsimony and distance-based methods of phylogeny construction for restriction data. *Molecular Biology and Evolution 8*, 356–365.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism*, Volume III, Chapter 24, pp. 21–132. New York: Academic Press.

Kannan, S. and T. Warnow (1994). Inferring evolutionary history from DNA sequences. *SIAM Journal on Computing 23*, 713–737.

Kannan, S. and T. Warnow (1997). A fast algorithm for the computation and enumeration of perfect phylogenies. *SIAM Journal on Computing 26*, 1749–1763. Preliminary version at SODA '95.

Kearns, M. J. and U. V. Vazirani (1994). *An Introduction to Computational Learning Theory*. Cambridge, Mass.: MIT Press.

Kidd, K. K. and L. A. Sgaramella-Zonta (1972). Phylogenetic analysis: concepts and methods. *American Journal of Human Genetics 23*, 235–252.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16*, 116–120.

Kimura, M. (1981a). Estimation of evolutionary differences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the USA 78*, 454–458.

Kimura, M. (1981b). Was globin evolution very rapid in its early stages? A dubious case against the rate-constancy hypothesis. *Journal of Molecular Evolution 17*, 110–113.

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

Kluge, A. R. and J. S. Farris (1969). Quantitative phyletics and the evolution of anurans. *Systematic Zoology 18*, 1–32.

Křivánek, M. and J. Morávek (1986). NP-hard problems in hierarchical-tree clustering. *Acta Informatica 23*, 311–323.

Kruskal, Jr., J. B. (1956). On the shortest spanning subtree of a graph and the travelling saleman problem. *Proceedings of the American Mathematical Society 7*, 48–50.

Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. *Molecular Biology and Evolution 13*, 584–593.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences of the USA 91*, 1455–1459.

Lanave, C., G. Preparata, C. Saccone, and G. Serio (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution 20*, 86–93.

Le Quesne, W. J. (1969). A method of selection of characters in numerical taxonomy. *Systematic Zoology 18*, 201–205.

Le Quesne, W. J. (1972). Further studies based on the uniquely derived character concept. *Systematic Zoology 21*, 281–288.

Le Quesne, W. J. (1974). The uniquely evolved character concept and its cladistic application. *Systematic Zoology 23*, 513–517.

Leitner, T., D. Escanilla, C. Franzén, M. Uhlén, and J. Albert (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences of the USA 93*, 10864–10869.

Li, W.-H. and T. Gojobori (1983). Rapid evolution of goat and sheep globin genes following gene duplication. *Molecular Biology and Evolution 1*, 94–108.

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

Lodish, H., D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell (1995). *Molecular Cell Biology* (3rd ed.). New York: Scientific American Books, Inc.

Maddison, D. R., M. Ruovolo, and D. L. Swofford (1992). Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Systematic Biology 41*, 111–124.

Maidak, B. L., J. R. Cole, T. G. Lilburn, J. Charles T. Parker, P. R. Saxman, J. M. Stredwick, G. M. Garrity, B. Li, G. J. Olsen, S. Pramanik, T. M. Schmidt, and J. M. Tiedje (2000). The RDP (Ribosomal Database Project) continues. *Nucleic Acids Research 28*, 173–174.

Makarenkov, V. and P. Casgrain (1999). *Program T-REX (Tree and reticulogram Reconstruction).* ⟨http://www.fas.umontreal.ca/BIOL/Casgrain/en/labo/t-rex/index.html⟩.

Margoliash, E. (1963). Primary structure and evolution of cytochrome *c*. *Proceedings of the National Academy of Sciences of the USA 50*, 672–679.

Matisoo-Smith, E., R. M. Roberts, G. J. Irwin, J. S. Allen, D. Penny, and D. M. Lambert (1998). Patterns of prehistoric human mobility in

Polynesia indicated by mtDNA from the Pacific rat. *Proceedings of the National Academy of Sciences of the USA 95*, 15145–15150.

Matsuda, H., G. J. Olsen, R. Overbeek, and Y. Kaneda (1994). Fast phylogenetic analysis on a massively parallel machine. In *Proceedings of the Eighth International Conference on Supercomputing — ICS '94*, pp. 297–302.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pp. 148–188. Cambridge: Cambridge University Press.

Moilanen, A. (1999). Searching for the most parsimonious tree with simulated evolutionary optimization. *Cladistics 15*, 39–50.

Nei, M., S. Kumar, and K. Takahashi (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of the National Academy of Sciences of the USA 95*, 12390–12397.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In S. S. Gupta and J. Yackel (Eds.), *Statistical Decision Theory and Related Topics*, pp. 1–27. New York: Academic Press.

Noro, M., R. Masuda, I. A. Dubrovo, M. C. Yoshida, and M. Kato (1998). Molecular phylogenetic inference of the Woolly Mammoth *mammuthus primigenius*, based on complete sequences of mitochondrial cytochrome *b* and 12S ribosomal genes. *Journal of Molecular Evolution 46*, 314–326.

Olsen, G. J., J. H. Matsuda, R. Hagstrom, and R. Overbeek (1994). FastDNAml: a tool for construction of phylogenetic tees of DNA sequences using maximum likelihood. *CABIOS 10*, 41–48.

Ou, C.-Y., C. A. Cieselski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. MacInnes, J. W. Curran, and H. W. Jaffe (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science 256*, 1165–1171.

Patrinos, A. N. and S. L. Hakimi (1972). The distance matrix of a graph and its realization. *Quarterly of Applied Mathematics 30*, 255–269.

Penny, D. and M. Hasegawa (1997). Molecular systematics: the platypus put in its place. *Nature 387*, 549–550.

Penny, D., P. J. Lockhart, M. A. Steel, and M. D. Hendy (1994). The role of models in reconstructing evolutionary trees. In R. W. Scotland, D. J. Siebert, and D. M. Williams (Eds.), *Models in Phylogeny Reconstruction*, Volume 52 of *Systematic Association Series*, pp. 211–230. Oxford: Clarendon Press.

Press, W. H., S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed.). Cambridge, UK: Cambridge University Press.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Systems Technical Journal 36*, 1389–1401.

Purdom, Jr., P. W., P. G. Bradford, K. Tamura, and S. Kumar (2000). Single column discrepancy and dynamic max-mini optimizations for quickly finding the most parsimonious evolutionary trees. *Bioinformatics 16*, 140–151.

Rényi, A. (1970). *Foundations of Probability.* San Francisco: Holden-Day.

Rice, K. and T. Warnow (1997). Parsimony is hard to beat! See Jiang and Lee (1997), pp. 124–133.

Robinson, D. F. and L. R. Foulds (1981). Comparison of phylogenetic trees. *Mathematical Biosciences 53*, 131–147.

Rodríguez, F., J. L. Oliver, A. Marín, and J. R. Medina (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology 142*, 485–501.

Rohlf, F. J. (1983). Numbering binary trees with labeled terminal vertices. *Bulletin of Mathematical Biology 45*, 33–40.

Rzhetsky, A. and M. Nei (1992a). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution 9*, 945–967.

Rzhetsky, A. and M. Nei (1992b). Statistical properties of the ordinary least-squares, generalized least-squares and mminimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution 35*, 367–375.

Rzhetsky, A. and M. Nei (1993). Theoretical foundation of the minimum evolution method or phylogenetic inference. *Molecular Biology and Evolution 10*, 1073–1095.

Saitou, N. and T. Imanishi (1989). Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum evolution, and neighbor-joining methods of phylogenetic tree reconstruction in obtaining the correct tree. *Molecular Biology and Evolution 6*, 514–525.

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

Sankoff, D. D. (1975). Minimum mutation trees of subsequences. *SIAM Journal on Applied Mathematics 28*, 35–42.

Sankoff, D. D. and P. Rousseau (1975). Locating the vertices of a Steiner tree in arbitrary space. *Mathematical Programming 9*, 240–246.

Sattath, S. and A. Tversky (1977). Additive similarity trees. *Psychometrika 42*, 319–345.

Setubal, J. C. and J. Meidanis (1997). *Introduction to Computational Molecular Biology*. Boston, Mass.: PWS Publishing Company.

Siddall, M. E. (1998). Success of parsimony in the four-taxon case: long branch repulsion by likelihood in the Farris-zone. *Cladistics 14*, 209–220.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood (1986). Fluorescence detection in automated dna sequence analysis. *Nature 321*, 674–679.

Smith, T. J. (1998). A comparison of three additive tree algorithms that rely on a least-squares loss criterion. *British Journal of Mathematical & Statistical Psychology 51*, 269–288.

Smolenskiĭ, E. A. (1962). Об одном способе линейной записи графов (On a method of linear recording of graphs). Журнал вычислителной математики и математической физики *2*(№ 2), 371–372.

Sneath, P. H. A. and R. R. Sokal (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco, Cal.: W. H. Freeman.

Sokal, R. R. and C. D. Michener (1957). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin 38*, 1409–1438.

Sokal, R. R. and P. H. A. Sneath (1963). *Principles of Numerical Taxonomy*. San Francisco, Cal.: W. H. Freeman.

Sourdis, J. and M. Nei (1988). Relative efficiencies of the maximum parsimony and distance-based methods in obtaining the correct phylogenetic tree. *Molecular Biology and Evolution 5*, 298–311.

Steel, M., L. A. Székely, and P. L. Erdős (1996). The number of nucleotide sites needed to accurately reconstruct large evolutionary trees. Technical Report 96-19, DIMACS.

Steel, M. A. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification 9*, 91–116.

Steel, M. A. (1994a). The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology 43*, 560–564.

Steel, M. A. (1994b). Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters 7*, 19–24.

Strimmer, K. and A. von Haeseler (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution 13*, 964–969.

Studier, J. A. and K. J. Keppler (1988). A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology and Evolution 5*, 729–731.

Suzuki, T. and K. Imai (1998). Evolution of myoglobin. *Cellular and Molecular Life Sciences 54*, 979–1004.

Swofford, D. L. (1990). *PAUP: Phylogeny Analysis Using Parsimony, version 3*. Champaign, Ill.: Illinois Natural History Survey.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis (1996). Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.), *Molecular Systematics* (2nd ed.)., Chapter 11, pp. 407–514. Sunderland, Mass.: Sinauer Associates.

Takahata, N. and M. Kimura (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics 98*, 644–657.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on mathematics in the life sciences*, Volume 17, Providence, Rhode Island, pp. 57–86. American Mathematical Society.

Tavaré, S. (1995). Calibrating the clock: using stochastic processes to measure the rate of Evolution. In E. S. Lander and M. S. Waterman (Eds.), *Calculating the Secrets of Life*, pp. 114–152. National Academy Press.

Trelles, O., C. Ceron, H. C. Wang, J. Dopazo, and J. M. Carazo (1998). New phylogenetic venues opened by a novel implementation of the DNAml algorithm. *Bioinformatics 14*, 544–545.

Tuffley, C. and M. Steel (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology 59*(3), 581–607.

Vach, W. and P. O. Degens (1991). Least-squares approximation of additive trees to dissimilarities. *Computational Statistics Quarterly 3*, 203–218.

Wang, L. and T. Jiang (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology 1*, 337–348.

Wang, L., T. Jiang, and E. L. Lawler (1996). Approximation algorithms for tree alignment with a given phylogeny. *Algorithmica 16*, 302–315. Preliminary version at STOC '94.

Warnow, T. (1994). Tree compatibility and inferring evolutionary history. *Journal of Algorithms 16*, 388–407.

Waterman, M. S., T. F. Smith, M. Singh, and W. A. Beyer (1977). Additive evolutionary trees. *Journal of Theoretical Biology 64*, 199–213.

Wayne, R. K., J. A. Leonard, and A. Cooper (1999). Full of sound and fury: the recent history of ancient DNA. *Annual Review of Ecology and Systematics 30*, 457–477.

Wilson, A. C., S. S. Carlson, and T. J. White (1977). Biochemical evolution. *Annual Review of Biochemistry 46*, 573–639.

Wolf, M. J., S. Easteal, M. Kahn, B. D. McKay, and L. S. Jermiin (2000). TrExML: a maximum-likelihood approach for extensive tree-space exploration. *Bioinformatics 16*, 383–394.

Wu, C.-I. and W.-H. Li (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the USA 82*, 1741–1745.

Zaretskiĭ, K. A. (1965). Построение дерева по набору расстояний между висячими вершинами (Construction of a tree from the distances between its pending vertices). Успехи математических наук *20*(№ 6), 90–92.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution 39*, 315–329.

Zuckerkandl, E. and L. Pauling (1962). Molecular disease, evolution and genic heterogeneity. In M. Kasha and B. Pullman (Eds.), *Horizons in Biochemistry*, pp. 189–225. New York: Academic Press.

Zuckerkandl, E. and L. Pauling (1965). Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel (Eds.), *Evolving Genes and Proteins*, pp. 97–116. New York: Academic Press.

# Index