Université catholique de Louvain

Ecole polytechnique de Louvain

Models and Methods for Computational Phylogenetics under Minimum Evolution

Martin Frohn

Center for Operations Research and Econometrics (C.O.R.E.)

Louvain Institute of Data Analysis and Modeling in economics and statistics

Promoteur

Prof. Daniele Catanzaro

Année académique 2020-2021

Thèse présentée en vue de l'obtention
du titre de Docteur en Sciences de l'Ingénieur et technologie.

Abstract

In this thesis we study a specific combinatorial optimization problem, called the *Balanced Minimum Evolution Problem* (BMEP), which is an $\mathcal{APX}$-hard network design problem that consists of finding a minimum length unrooted binary tree (called a *phylogeny*) having as a leaf-set a given set of molecular sequences. The optimal solution to the BMEP (i.e., the optimal phylogeny) encodes the hierarchical evolutionary relationships of the input sequences. This information is crucial for a multitude of research fields, ranging from systematics to medical research, passing through drug discovery, epidemiology, ecology, biodiversity assessment and population dynamics. We provide new results on the combinatorial and computational aspects of the BMEP, present an algorithm that performs better than previously known exact solution algorithms for the BMEP and discuss the implications of these insights for other phylogenetic estimation problems based on minimum evolution.

Acknowledgements

I acknowledge financial support from the FRS-FNRS via the grant "Crédit de Recherche" ref. S/25-MCF/OL J.0026.17, the Université catholique de Louvain, Belgium via the "Fonds Spéciaux de Recherche" (FSR) 2017–2021, and the Fondation Louvain, Belgium via the grant "COALESCENS" belonging to the program "Le numérique au service de l'humain". Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

Thanks to Prof. Daniele Catanzaro for being a brilliant mentor, for his trust in me and his confidence whenever I was too doubtful about current developments. Without meeting him I would not have started a PhD project. He has an unique ability to inspire other people's own interests. His passion for research is very visceral, making long discussions and reviews of my work always fruitful for mind and heart. His teachings allowed me to learn and understand a lot about accurate and precise communication.

Thanks to Prof. Raffaele Pesenti for his insights on how to conduct research with an independent mindset, the broad knowledge he shared on different subjects and his willingness to continuously give useful hints to overcome obstacles arising during the development of this thesis over the years. I am grateful to have met him as my teacher.

Thanks to Prof. Maurice Queyranne for his suggestions, brainstorming sessions and his kind and forthcoming support during my stay at his chair.

Louvain-la-Neuve, October 11th 2021

Preface

This work consists of contributions in computational phylogenetics. Specifically:

An information theory perspective on phylogenetics: The thesis provides a reinterpretation of the Balanced Minimum Evolution Problem (BMEP) as a cross-entropy minimization problem. Moreover, it investigates the information theory aspects of phylogenetic estimation models under minimum evolution and connects the notions of statistical consistency and numerical stability for these models.

Computational aspects of the BMEP: The thesis shows that the $\mathcal{NP}$-hardness of the BMEP persists if we consider the assignment subproblem nested inside the BMEP and the limitations of this result for the development of approximation schemes.

Applications: The thesis derives a new family of dual bounds for the BMEP and provides a polynomial-time algorithm to calculate them. Furthermore, it proposes a massively parallel Branch-&-Bound algorithm for the BMEP and shows that this algorithm performs better than any previously known method for the problem.

In the first chapter we introduce the reader to the field of computational phylogenetics and give more details on fundamental structures in the subsequent chapter. Hence, in the second chapter, we introduce the reader to the Balanced Minimum Evolution Problem and provide a survey to give a historical perspective for the emergence of this problem 21 years ago. In particular, we introduce the concepts and problems in phylogenetics that lead to the development of the BMEP as well as details on the most important achievements for the BMEP that are of interest for the operations research community and include some challenges that still remain to be addressed.

The third chapter describes a new view of the problems and considerations made that lead to the development of the BMEP from an information theory perspective. Specifically, we study the connections between the combinatorial aspects of the BMEP and its interpretation as a cross-entropy minimization problem. This result adds to the insights previously

reserved only to connections between the BMEP and the Traveling Salesman Problem and the Huffman Coding Problem, respectively. Moreover, it allows us to develop an algorithm to calculate new lower bounds on the optimal solution to the BMEP in polynomial time. In addition, we evaluate the quality of these bounds in relation to the input data. Furthermore, the information theory interpretation of the BMEP opens up the discussion whether other phylogenetic estimation problems could be studied in the same framework. We provide evidence that this question is well founded and discuss the implications.

Subsequently, we consider the ordering of feasible solutions in cross-entropy problems. Specifically, in the fourth chapter we show that restricting the BMEP to a fixed phylogeny constitutes a nested cross-entropy subproblem of the BMEP and a special case of the General Assignment Problem as well as preserve the $\mathcal{APX}$-hardness of the BMEP. Moreover, we present arguments in support of the claim that differences in complexity between the BMEP and its restriction can neither be exploited computationally nor are apparent.

Finally, in the fifth chapter we present a massively parallel exact solution algorithm for the BMEP and show empirically that it succeeds in solving more and larger instances than previously known methods by utilizing a series of theoretical insights discussed in this thesis. We also show which limitations any exact solution algorithm for the BMEP faces in terms of numerical stability and statistical consistency and provide evidence that our algorithm takes a significant step in closing the gap towards the best possible exact solution algorithm.

Some chapters of this thesis are partially based on articles that have been published or prepared for publication in a number of international peer-reviewed journals. Specifically:

Chapter 2 is contained in:

Daniele Catanzaro, Martin Frohn, Olivier Gascuel, Raffaele Pesenti A tutorial on the balanced minimum evolution problem. Invited Review, European Journal of Operational Research, 2021.

Chapter 3 is contained in:

Daniele Catanzaro, Martin Frohn, Raffaele Pesenti An information theory perspective on the balanced minimum evolution problem. Operations Research Letters, 2020.

Chapter 4 is contained in:

Martin Frohn On the approximability of the fixed-tree balanced minimum evolution problem. Optimization Letters, 2021.

Chapter 5 is contained in:

Daniele Catanzaro, Martin Frohn, Raffaele Pesenti A massively parallel exact solution algorithm for the balanced minimum evolution problem. In preparation.

This thesis is intended for readers who are familiar with some fundamental concepts from phylogenetics and discrete optimization.

Contents

1

Introduction

Phylogenetics is the discipline concerned with the study of hierarchical evolutionary relationships among organisms through the reconstruction of phylogenetic trees from molecular data, e.g., DNA or RNA nucleotide sequences, or from morphological data, e.g., specific plants or animal bones. Figure 1.1 shows a phylogenetic tree that describes the evolutionary relationships between different rabbit breeds by introducing hypothetical common ancestors. For example, the Chinchilla rabbit and the Japanese hare are at an estimated evolutionary distance of 267 and 131, respectively, from their latest common ancestor.

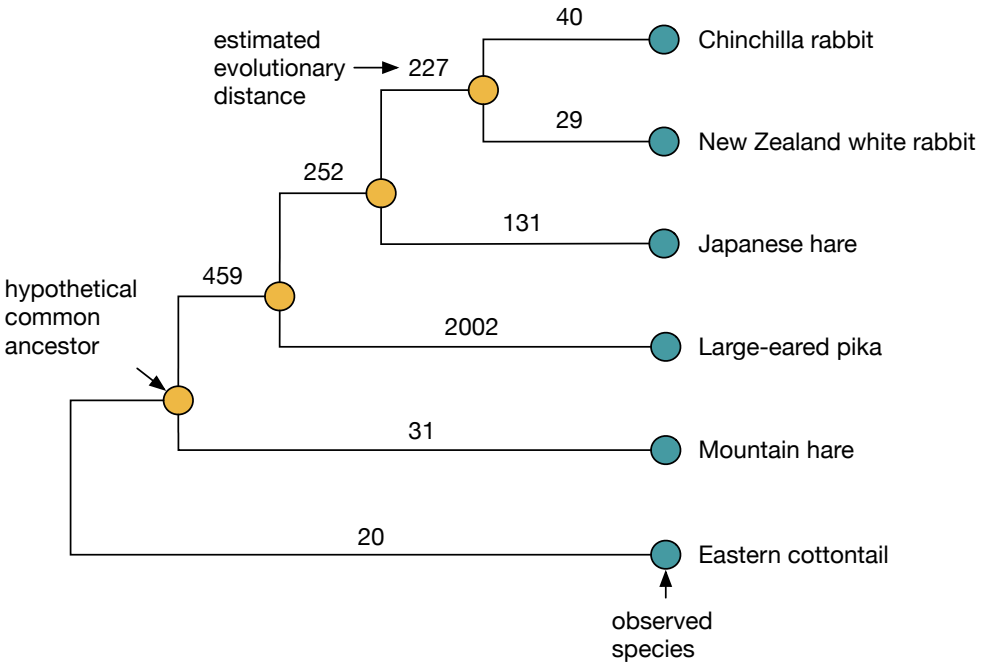


Fig. 1.1. An example of a phylogenetic tree for different species encoded by RNA nucleotide sequences from the GenBank/ NCBI database. The observed species, hypothetical common ancestors and estimated evolutionary distances are represented by external vertices, internal vertices and edge weights of the tree, respectively.

By providing core insights on the evolutionary dynamics of many fine-scale molecular data, phylogenetic trees prove of considerable assistance in a multitude of research fields and applications ranging from systematics to medical research, passing through drug discovery, epidemiology, ecology, biodiversity assessment and population dynamics [Pachter and Sturmfels, 2007, Lemmon and Lemmon, 2013, McGuire et al., 2014, Duellman et al., 2016, Valiente-Banuet and Verdú, 2013, Jäger, 2018, Kadam et al., 2016]. For example, phylogenetic trees have been used to predict evolution of human influenza A [Bush et al., 1999, Perovic, 2013]; to understand the relationships between the virulence and evolution of HIV [Ross and Rodrigo, 2002, Ou et al., 1992, Castro-Nallar et al., 2012, Poon et al., 2015]; to identify emerging viruses such as SARS [Marra et al., 2003, Amiroch et al., 2017, Ge et al., 2013]; to recreate and investigate ancestral proteins [Chang and Donoghue, 2000, Washburne et al., 2018]; to design neuropeptides causing smooth muscle contraction [Bader et al., 2001]; to relate geographic patterns to ecological and macro-evolutionary processes [Leibold et al., 2010, Kress et al., 2010, Harvey et al., 1996, McCormack et al., 2013]; or to uncover similarities in evolution of a number of human languages [Jäger, 2018, Bowern and Atkinson, 2012]. Phylogenetic trees have also been used to study the evolutionary processes underlying the genetic factors involved in common human diseases [Pennington et al., 2006, Sridhar et al., 2007, 2008, Misra et al., 2011, Catanzaro et al., 2009, 2013b] as well as those at the core of the progression of carcinomas over time [Riester et al., 2010b, Subramanian et al., 2013, Catanzaro et al., 2016, Chowdhury et al., 2013, Riester et al., 2010a, Schwartz, 2019, Myers et al., 2019, Lähnemann et al., 2020]. In particular, in the cancer context, phylogenetic trees allowed the remarkable classification of tumor cells of given pathologies in subfamilies characterized by specific evolutionary traits [Schwartz, 2019]. A shared hope is that this classification could enable a better understanding of cellular atypia over time and, in the long

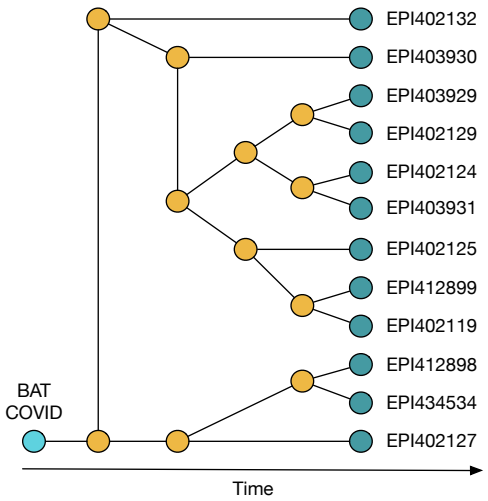


Fig. 1.2. An example of the phylogenetic tree of SARS-CoV-2 Chinese genomes sampled from affected patients during December 2019. Data, tree, and taxa labels refer to GISAID nomenclature (`https://www.gisaid.org/epiflu-applications/next-hcov-19-app/`). The BAT COVID label refers to the bat (R. affinis) coronavirus isolate BatCoVRaTG13 from Yunnan Province, used as an outgroup in [Forster et al., 2020].

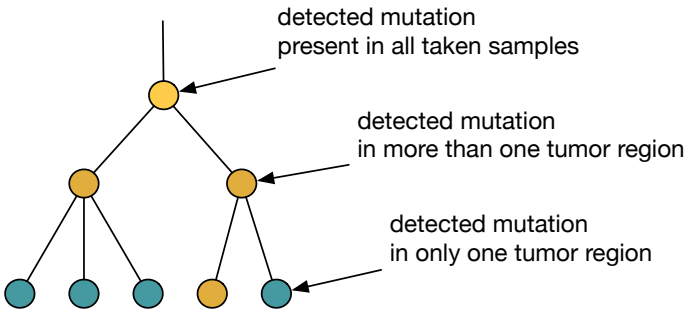


Fig. 1.3. An example of part of the topology of a phylogenetic tree for nucleotide variants arising in the sample of a lung cancer patient (see patient CRUK0062 in Abbosh et al. [2017] for more details).

run, suggest new therapeutic targets [Beerenwinkel et al., 2015, Schwartz, 2019, Myers et al., 2019, Lähnemann et al., 2020].

More recently, phylogenetic trees are proving of fundamental support to study evolution of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) as well as to track its spreading during the COVID19 pandemic (see, e.g., `https://nextstrain.org/ncov/global` and `https://www.gisaid.org/epiflu-applications/phylodynamics/`). By describing core mutations of the virus over time and through the continents, the phylogenetic tree of SARS-CoV-2 (see, e.g., Figure 1.2) shows how distinctive migratory histories and founder events may modulate the clinical presentation in affected patients [Lai et al., 2020, Zhou et al., 2020, Mavian et al., 2020]. This key information, in turn, aids the search for potential vaccines.

We can observe, apart from the measure being used to estimate the evolutionary distance between adjacent vertices in a phylogenetic tree (see Figure 1.1), that the shape, i.e., *topology*, of the tree and its corresponding biological meaning may depend on the specific application or use. In our example on rabbit breeds we notice that all observed species are known a priori and that external and internal vertices of the tree have distinct interpretations. This does not have to be the case when constructing a phylogenetic tree. For example, in the context of tumor evolution (see Figure 1.3), the topology of a phylogenetic tree can be represented as an arborescence of point mutations where the set of external vertices is unknown when clinical samples are taken. In systematics (central for this thesis), instead, the topology of a phylogenetic tree is usually encoded as an unrooted binary tree in which the external vertices (or *leaves*) represent the observed molecular sequences (e.g., sequences over a fixed alphabet); *internal vertices* represent speciation events that occurred throughout evolution of these sequences; *edges* represent estimated evolutionary relationships; and *edge weights* represent a measure of the similarity between pairs of observed molecular sequences [Catanzaro, 2009].

There exists a wide varity of methods to compute the edge weights and topology of a phylogenetic tree that fit the given data well. The development of such estimation methods reaches back to Darwin [1859]. The author studied the evidence that evolutionary relationships between species can be attributed to hierarchical branching patterns of evolution, e.g.,

the notion of natural selection: "[The] preservation of favourable variations and the rejection of injurious variations, I call Natural Selection." Following these early studies, for a long time, phylogenetic trees for a group of species were computed by relying on morphological features and the experience of the naturalist studying the group of species. The first quantitative approaches started to emerge with the work of Sokal and Michener [1957] which allowed for the development of increasingly automized procedures to determine the edge weights and topology of phylogenetic trees. Furthermore, many different methods have been developed in the past 60 years to achieve this task, some of which will be detailed in this thesis. At their core, estimation methods for phylogenetic trees aim to calculate the *true* phylogenetic tree, i.e., a tree one constructs if the information inferred by the molecular or morphological data would be complete. Hence, these estimation methods can be seen as optimization algorithms looking for the true tree of the observed data while differing in their criteria used to measure the differences between tree and data. However, finding the true phylogenetic tree for a general set of data remains a big computational challenge to this day.

In the next chapter we will go into more detail on the characteristics of phylogenetic trees, commonly called *phylogenies*, and some important criteria used to estimate them from molecular data.

2

A Review of the Balanced Minimum Evolution Problem

Consider a set $\Gamma = \{1, 2, \ldots, n\}$ of $n \geq 3$ distinct aligned molecular sequences over an alphabet Σ (such as DNA, RNA, codon sequences or whole genomes) hereafter referred to as *taxa* (see Figure 2.3 for an example of molecular sequences over the alphabet $\Sigma = \{A, T, C, G\}$). A *phylogeny* T of Γ is an ordered triplet $(U, \phi, \mathbf{w})$ such that U is an *Unrooted Binary Tree* (UBT) (also known as *cubic tree*) having n leaves, ϕ is a bijection between the leaves of U and the taxa in Γ, and $\mathbf{w}$ is a vector of weights associated to the edges of U. For example, Figure 2.1 shows the phylogeny of a set of seven biosafety level 2, 3 and 4 pathogens (taxa), including the whole genomes of the Crimean-Congo Hemorrhagic Fever (CCHF) orthonairovirus, Ebolavirus, the Lassa mammarenavirus, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the Human Immunodeficiency Viruses (HIVs) 1 and 2, and the Nipah virus. The complete genomes for these taxa are available at GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) via the reference numbers GCF_000854165.1, NC_002549.1, GCF_000851705.1, NC_004777.1 NC_045512.2, NC_001802.1, NC_001722.1, and NC_002728.1, respectively. The genome of Lassa virus has been obtained by concatenating the segment S (NCBI Reference Sequence NC_004296.1) with the segment L (NCBI Reference Sequence NC_004297.1). The genome of the Crimean-Congo orthonairovirus has been obtained by concatenating, in this order, the segment S (NCBI Reference Sequence NC_005302.1) with the segments M (NCBI Reference Sequence NC_005300.2) and L (NCBI Reference Sequence NC_005301.3).

Let $\mathbf{D} = \mathbf{D}(\Gamma)$ denote a given $n \times n$ symmetric distance matrix associated to Γ, whose generic entry d_{ij} – equal to zero on the main diagonal and strictly positive otherwise – en-

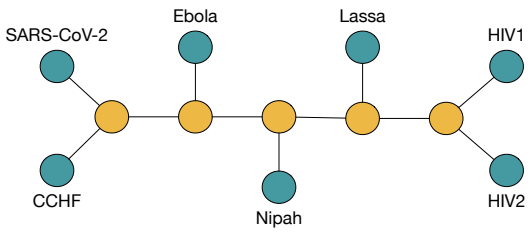


Fig. 2.1. An example of a phylogeny of a set of seven taxa (blue vertices). The internal vertices of the phylogeny are marked in yellow and edge weights have been removed for the sake of readability.

	CCHF	Ebola	Lassa	SARS-CoV-2	HIV1	HIV2	Nipah
CCHF	0.0000000	0.0048683	0.0094948	0.0139234	0.0109681	0.0114508	0.0081327
Ebola	0.0048683	0.0000000	0.0081206	0.0118547	0.0164171	0.0184611	0.0038627
Lassa	0.0094948	0.0081206	0.0000000	0.0154856	0.0326514	0.0306672	0.0148241
SARS-CoV-2	0.0139234	0.0118547	0.0154856	0.0000000	0.0338913	0.0406092	0.0143275
HIV1	0.0109681	0.0164171	0.0326514	0.0338913	0.0000000	0.0025788	0.0141298
HIV2	0.0114508	0.0184611	0.0306672	0.0406092	0.0025788	0.0000000	0.0203479
Nipah	0.0081327	0.0038627	0.0148241	0.0143275	0.0141298	0.0203479	0.0000000

Fig. 2.2. The distance matrix computed by Alfree (`http://150.254.123.165/alfree/`) when analyzing the taxa considered in Figure 2.1 by using d^{EVOL1} metrics (i.e., Stuart *et al.*' alignment-free *angle cosine metric*, see [Stuart et al., 2002a,b]).

codes a measure of the dissimilarity between the input taxa i and j in Γ, which we call the *evolutionary distance* between taxa i and j. For example, Figure 2.2 shows a distance matrix for the seven taxa in Figure 2.1. In the literature, multiple methods have been proposed to compute a distance matrix [Notredame, 2004, Ng et al., 2017, Felsenstein, 2004]. Figure 2.3 shows four examples of such methods, including the one cited in Figure 2.2. In general, given pairs of taxa, methods for the calculation of evolutionary distances can require an *alignment process* [Rosenberg, 2011] before being analyzed, i.e., a process that generates two

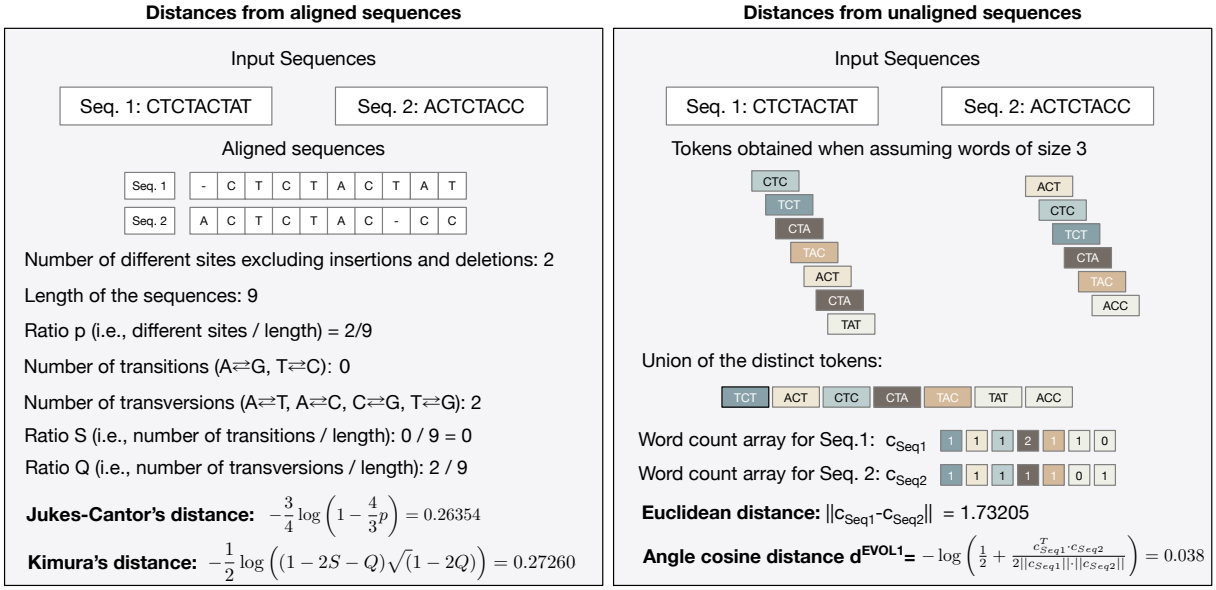


Fig. 2.3. Some examples of how to derive the evolutionary distance for a pair of dummy molecular sequences, namely "CTCTACTAT" and "ACTCTACC". In particular, the subfigure on the left shows two possible distances derived from the dissimilarity measures proposed by Jukes and Cantor [1969] and Kimura [1980], respectively. The subfigure on the right shows two possible distances derived from the alignment-free dissimilarity measures proposed by Blaisdell [1986] and Stuart et al. [2002a,b], respectively.

equal length molecular sequences for each pair of taxa by maximizing the region of similarity (i.e., the character matching of the sequences associated to the taxa). The methods shown in Figure 2.3 on the left assume that the taxa have undergone an alignment process and they are based on the time-reversible Markov model of DNA sequence evolution (see [Felsenstein, 2004, Tavare, 1987, Waddell and Steel, 1997, Lanave et al., 1984, Rodriguez et al., 1990, Hasegawa et al., 1981, Yang, 1994, Catanzaro et al., 2006b,a]). The Jukes and Cantor's distance can be considered as a sort of Hamming distance as it takes into account the proportion of variant sites between the two aligned sequences. Kimura's distance refines Jukes and Cantor's one, by weighting appropriately the occurrence of transitions and transversions (see [Felsenstein, 2004] for details). Because solving the multiple alignment problem is notoriously $\mathcal{NP}$-hard [Rosenberg, 2011], the family of methods shown on the right of Figure 2.3, in which the requirement of alignment is not present, may prove particularly useful to process very large molecular datasets. We refer the reader to [Rosenberg, 2011] and [Notredame, 2004, Ng et al., 2017, Lemoine et al., 2020, Simonsen et al., 2008] for an introduction and recent developments on the alignment problem as well as to [Felsenstein, 2004, Page and Holmes, 1998] for a review of the numerous alignment-based distances currently discussed in the literature. We refer instead the reader to [Zielezinski et al., 2017, Vinga and Almeida, 2003, Vinga, 2014] for an introduction and recent surveys on alignment-free distances.

Then, the *Balanced Minimum Evolution Problem* (BMEP) consists of finding a *phylogeny* T of Γ that minimizes the following *length function*

$$L(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{d_{ij}}{2^{\tau_{ij}}}, \tag{2.1}$$

where τ_{ij} represents the *path-length* between taxa i and j in T, i.e., the number of edges belonging to the (unique) path in T connecting taxon i to taxon j [Pardi, 2009, Catanzaro et al., 2012, Aringhieri et al., 2011, Catanzaro et al., 2020a,b]. Given the evolutionary distances d_{ij} shown in Figure 2.2, the phylogeny in Figure 2.1 is provably optimal for the *Balanced Minimum Evolution Problem* (BMEP). On this phylogeny, the path-length between Ebola and Lassa virus is four, whereas the path-length between SARS-CoV-2 and HIV-1 is six. More in general, the whole path-length matrix $\tau = \{\tau_{ij}\}$ reads as follows

CCHF	Ebola	Lassa	SARS-CoV-2	HIV1	HIV2	Nipah	
0	3	5	2	6	6	4	CCHF
3	0	4	3	5	5	3	Ebola
5	4	0	5	3	3	3	Lassa
2	3	5	0	6	6	4	SARS-CoV-2.
6	5	3	6	0	2	4	HIV1
6	5	3	6	2	0	4	HIV2
4	3	3	4	4	4	0	Nipah

In this chapter we introduce the reader to the biological and statistical foundations of the BMEP, to its combinatorics and to its optimization aspects. Moreover, we review and

classify the recent advances described in the literature, by highlighting the theoretical and computational challenges that still resist to the attacks of the scientific community.

2.1 Basic properties of phylogenies

In the context of systematics, the internal vertices of a phylogeny of n taxa usually have degree three [Felsenstein, 2004, Semple and Steel, 2003, Page and Holmes, 1998, Gascuel, 2005, Catanzaro, 2011]. This constraint is biologically founded for all the evolutionary processes that are *bifurcating*, i.e., those processes for which the occurrence of mutations over time gives rise to two distinct taxa at each speciation event [Felsenstein, 2004, Page and Holmes, 1998, Volkenstein and Livshits, 1989]. It is worth noting, however, that this constraint can also model evolutionary processes that include *polytomies*, i.e., multi-furcations during evolution of taxa [Page and Holmes, 1998]. In fact, any m-ary tree can be transformed into an unrooted binary tree by just adding "dummy" edges and vertices as shown, e.g., in Figure 2.4. The degree constraint on the internal vertices of a phylogeny T implicitly fixes the overall number of its edges and internal vertices to $2n-3$ and $n-2$, respectively. Indeed, according to classical results from graph theory (see [Reinhard, 2005]) it holds that

$$|E_i(T)| + |E_e(T)| = |V_i(T)| + |V_e(T)| - 1, \tag{2.2}$$

where $E_e(T)$, $E_i(T)$, $V_e(T)$ and $V_i(T)$ denote the sets of external edges, internal edges, external vertices and internal vertices of T, respectively. Moreover, because internal vertices have degree three, it holds also that

$$2|E_i(T)| + 2|E_e(T)| = 3|V_i(T)| + |V_e(T)|. \tag{2.3}$$

By combining (2.2) and (2.3), it follows that $|V_i(T)| = (n-2)$ and $|E_i(T)| = (n-3)$. Hence, the topology of a phylogeny looks unrooted and binary (see, e.g., Figure 2.1). The degree constraint on the internal vertices of a phylogeny also implies that the overall number of possible phylogenies for a set of n taxa is equal to $(2n-5)!! = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-5)$ [Catanzaro, 2009, Felsenstein, 2004]. This exponential number of plausible alternatives entails the search for specific estimation criteria to select the proper phylogeny of a given set of taxa [Catanzaro, 2011, Catanzaro et al., 2013a, 2015].

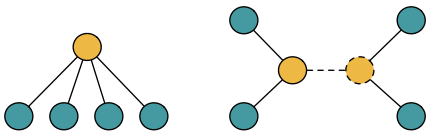


Fig. 2.4. The 4-ary tree (on the left) can be transformed into an unrooted binary tree by adding a dummy vertex and a dummy edge (dashed, on the right).

2.2 Estimating phylogenies from molecular data

The literature on phylogenetics proposes several estimation criteria for systematic purposes, differing from one another in respect of the nature and type of observed inherited traits considered, the evolutionary hypotheses and assumptions at their core, and the statistical properties of the phylogeny provided [Gascuel, 2005, Felsenstein, 2004, Page and Holmes, 1998]. Examples include the *parsimony criterion*, the criteria based on *distance methods*, the *likelihood criterion*, and the criteria based on *Bayesian inference* [Felsenstein, 2004, Gascuel, 2005, Page and Holmes, 1998, Huelsenbeck et al., 2001, 2002]. Discussing all of these estimation criteria goes beyond the scope of this introduction. The reader interested in such a topic may refer however both to comprehensive sources such as [Felsenstein, 2004, Page and Holmes, 1998, Gascuel, 2005, Yang, 2014, Nei, 2000, Semple and Steel, 2003, Albert, 2005] and surveys, such as those discussed in [Catanzaro, 2009, 2011]. Here, we just limit to mention that these criteria are designed so as to enable the estimation of a phylogeny $(U, \phi, \mathbf{w})$ for Γ based on the molecular sequences of the considered taxa, which incidentally often is the only evolutionary information available. The vast majority of these criteria can be stated in terms of $\mathcal{NP}$-hard network design problems defined over UBTs, whose general paradigm reads as follows [Catanzaro, 2011, Catanzaro et al., 2015, Gascuel, 2000]:

Problem 1 *(The phylogenetic estimation paradigm)*
Given a set Γ of n taxa, find a phylogeny T^* that solves the problem

$$\underset{T \in \mathcal{T}}{\text{optimize}} \quad \Lambda_\Gamma(T) \tag{2.4}$$

$$s.t. \quad \Psi_\Gamma(T) = 0 \tag{2.5}$$

$$\Omega_\Gamma(T) \geq 0 \tag{2.6}$$

where $\mathcal{T}$ is the set of the $(2n-5)!!$ phylogenies of Γ; $\Lambda_\Gamma : \mathcal{T} \to \mathbb{R}$ is a function that models a selected estimation criterion; $\Psi : \Gamma \times \mathcal{T} \to \mathbb{R}^{m_1}$, for some $m_1 \geq 1$, and $\Omega : \Gamma \times \mathcal{T} \to \mathbb{R}^{m_2}$, for some $m_2 \geq 1$, are auxiliary functions that impose additional constraints on the structure of feasible phylogenies on the basis of given evolutionary assumptions. As an example, in the case of the BMEP, the function Λ_Γ encodes the length function of a phylogeny; in the context of the likelihood criterion, instead, the function Λ_Γ encodes the likelihood score of a phylogeny [Felsenstein, 2004]. $\Psi_\Gamma(T)$ and Ω_Γ may encode e.g., constraints on the paths between pairs of distinct taxa and may be even not present as in the case of the BMEP.

A specific estimation criterion (or, equivalently, optimization problem) is completely characterized by defining the functions Λ_Γ, Ψ_Γ and Ω_Γ. A phylogeny $T^* \in \mathcal{T}$ that optimizes Λ_Γ and satisfies the additional constraints described by $\Psi_\Gamma(T)$ and Ω_Γ is referred to as *optimal*.

One of the most successful estimation criteria described in the literature on phylogenetics (in particular in that of distance methods) is *Balanced Minimum Evolution* (BME) [Felsenstein, 2004, Gascuel, 2005]. This criterion was introduced by Desper and Gascuel [2002] based

on Pauplin's *Direct Calculation Formulae of branch-lengths* (DCF) discussed in [Pauplin, 2000] (see also the next section). The corresponding optimization problem, i.e., the BMEP, can be seen as: (i) a restriction of the *Minimum Evolution Problem*, discussed in [Catanzaro et al., 2015]; (ii) a particular version of the *network design problem* [Johnson et al., 1978, Pop, 2012] defined over the class of the UBTs; and (iii) a particular version of the *Steiner tree problem* in which the length function depends upon the topology of the tree and the number of Steiner nodes is a priori known [Cheng and Du, 2001, Gusfield, 1984, Lu et al., 2003, Prömel and Steger, 2002, Du et al., 2000, Hwang et al., 1992, Cieslik, 1998, Li and Mao, 2016, Wu and Chao, 2004, Du and Hu, 2008]. BME is proven to be one of the most accurate estimation criteria in phylogenetics [Lefort et al., 2015]. It inherits from distance methods the benefit of being able to handle virtually any kind of data for which a dissimilarity measure is available [Schwartz, 2019]; moreover, because it works on evolutionary distances of taxa rather than on their molecular sequences it results to be computationally less demanding than other estimation criteria such as the parsimony criterion, the likelihood criterion, or Bayesian inference. These properties make BME (and the related BMEP) particularly suitable for general-purpose phylogenetic analyses.

By celebrating the first twenty years of the BMEP, in this chapter we want to achieve a twofold pedagogical and scientific goal: we wish to provide the necessary background to familiarize the reader with the problem as well as to review and classify the recent combinatorial and computational advances described in the literature. To this end, in Section 2.3 we review the theoretical foundations of the BMEP, by presenting a brief historical perspective on the mathematical development that led to Pauplin's DCF [Pauplin, 2000]. In Section 2.4 we review the main combinatorial and optimization aspects of the BMEP, by reporting on the recent advances on its polyhedral combinatorics [Eickmeyer et al., 2008, Haws et al., 2011, Catanzaro et al., 2012, Forcey et al., 2015, 2017, Catanzaro et al., 2020b]. In Section 2.5 we propose a classification of the recent computational advances on the BMEP, including the aspects related to its computational complexity as well as the exact and the approximate solution approaches developed so far [Pardi, 2009, Aringhieri et al., 2011, Catanzaro et al., 2012, Fiorini and Joret, 2012, Desper and Gascuel, 2005, Gascuel, 2005]. Finally, in Section **??** we provide a first perspective on future developments related to the BMEP, by highlighting some currently open theoretical and computational challenges.

2.3 A brief history of the BMEP

We chronicle in this section the main stages of the mathematical development that led to *Pauplin's DCF* [Pauplin, 2000], by starting from the parsimony criterion and by proceeding then with the minimum evolution criterion and the balanced minimum evolution criterion. Figure 2.5 graphically summarizes the main stages of this development. Before proceeding, we introduce some notation and definitions that will prove useful throughout the thesis. In particular, we denote $\mathcal{T}$ as the set of the $(2n-5)!!$ phylogenies of Γ. In addition, for each phylogeny $T \in \mathcal{T}$, we denote $V(T)$ and $E(T)$ as the vertex and the edge set, respectively,

of T; p_{ij} as the unique path on a phylogeny $T \in \mathcal{T}$ from the leaf of taxon i to the leaf of taxon j; and $V(p_{ij})$ and $E(p_{ij})$ as the vertex and the edge sets, respectively, belonging to the path p_{ij}.

2.3.1 The parsimony criterion

Lex parsimoniae, also known in the literature as the *Occam's razor principle*, is a form of abduction widely known in the modern scientific approach [Popper, 2002]. This law is based on the falsifiability principle and states that from among a number of competing theories that may explain a given phenomenon under study, one should prefer the theories requiring the smallest number of assumptions, mainly because they are simplest, hence hopefully more easily falsifiable [Popper, 2002]. Inspired by the rationale of this law, Farris [1970] and Fitch [1971] laid in early 1970s the theoretical ground for the *parsimony criterion* of phylogenetic estimation [Albert, 2005], which can be summarized as follows. Consider an initial molecular sequence extracted from an ancestral taxon. In accordance to classical evolutionary theory, this sequence may change over time due to the presence of both random mutations and a selective pressure acting on it (see Figure 2.6 and [Farris, 1970, Fitch, 1971, Beyer et al., 1974, Waterman et al., 1977, Scheiner and Mindell, 2020, Darwin, 1964]). This change may occur in many forms, including point-mutations (e.g., substitution, insertion, or deletion of a nucleotide) or, where applicable, large-scale mutations (e.g., translocations, losses of heterozigosity, amplifications, deletions, crossovers, or inversions of entire molecular fragments) [Klung et al., 2019]. Speciation after speciation, these changes may give rise to a number of distinct molecular sequences (taxa) observed at time t. Now, classical evolutionary theory states that the equilibrium between adaptation and selective pressure causes that taxa

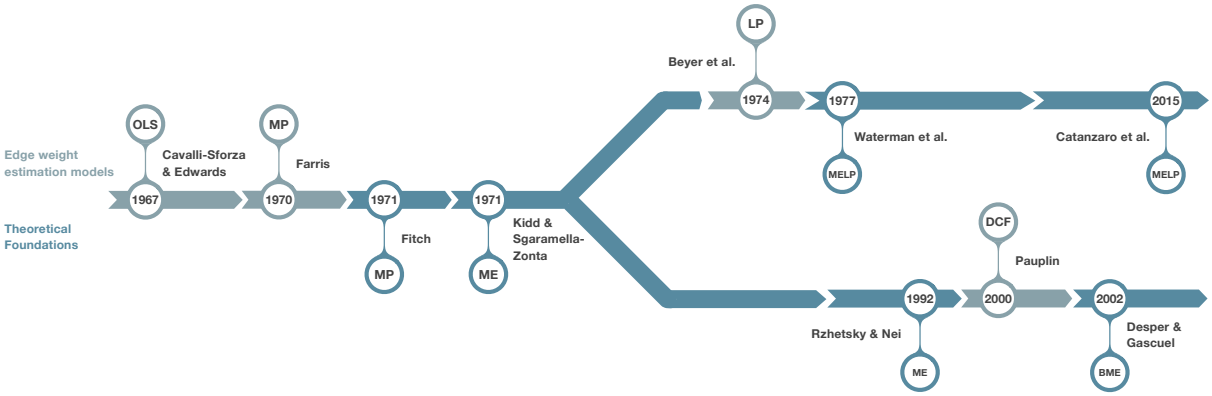


Fig. 2.5. Graphical representation of the main stages of the mathematical development that led to Pauplin's Direct Calculation Formulae of branch-lengths (DCF). The light gray stages represent articles introducing specific edge weight estimation models. We can distinguish between the Ordinary Least-Squares (OLS), the Maximum-Parsimony (MP), Pauplin's DCF, and the Linear Programming (LP). The blue stages represent articles introducing the theoretical foundations of phylogenetic estimation criteria based on specific edge weight estimation models. We can distinguish between the maximum-parsimony criterion (MP), the Minimum Evolution (ME) criterion, the Balanced Minimum Evolution (BME) criterion, and the more general *Minimum Evolution criterion under Linear Programming* (MELP) (see [Catanzaro et al., 2009, Catanzaro, 2009, 2011, Catanzaro et al., 2015]).

evolve from speciation events to speciation events by means of *local minimum changes* [Beyer et al., 1974, Albert, 2005, Page and Holmes, 1998]. Locally rather than globally minimum mostly because the selective pressure that acts on a taxon may not be constant throughout its evolution [Beyer et al., 1974, Albert, 2005]. Although these local minimum changes in general do not combine to form a global evolutionary process characterized by a minimum sum of changes, the hypothesis of a minimum chain of evolutionary events for the observed taxa appears to be, according to lex parsimoniae, more plausible than the hypothesis of more complex chains. Hence, Farris [1970] and Fitch [1971]'s parsimony criterion adduces the phylogeny whose overall change (i.e., the sum of the edge weights) is the *most parsimonious* (i.e., the smallest possible) as the one that properly approximates the overall evolutionary process of the observed taxa [Beyer et al., 1974, Waterman et al., 1977, Albert, 2005, Page and Holmes, 1998, Felsenstein, 2004].

Using the parsimony criterion to estimate the phylogeny of a given set Γ of taxa involves defining a model to estimate the edge weights associated to each plausible phylogeny $T \in \mathcal{T}$. These weights, in fact, estimate (possibly macro) speciation events that occurred throughout evolution of taxa and their sum defines the *length* of T that must be minimized. The earliest version of the parsimony criterion described in the literature on phylogenetics proposed to model the edge weights as Hamming distances between the molecular sequences associated to pairs of adjacent vertices in a given phylogeny $T \in \mathcal{T}$ [Felsenstein, 2004]. Specifically, this criterion assumes that the taxa in Γ are given as molecular sequences of length l over an alphabet Σ and it requires to estimate not only a phylogeny T, but also the molecular sequences of length l over Σ associated to the $n-2$ internal vertices of T. In this way, the weight associated to each edge $(u, v) \in E(T)$ can be expressed as the Hamming distance $d_H(s_u, s_v)$ between the molecular sequences s_u and s_v associated to the vertices u and v in $V(T)$, respectively. The criterion can be stated as follows:

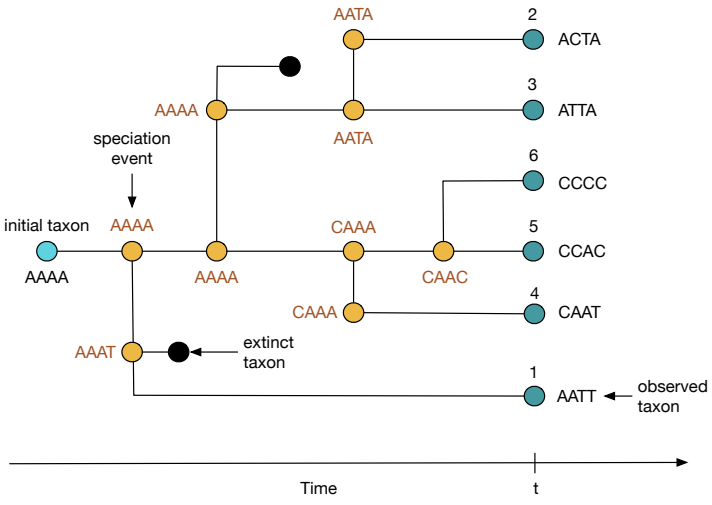


Fig. 2.6. An example of a hypothetical speciation process that starting from an ancestral taxon AAAA (light blue vertex) generates six distinct observed taxa at time t (dark blue vertices), namely AATT, ACTA, ATTA, CAAT, CCAC and CCCC. Random mutations as well as mutation caused by the selective pressure may change the original sequence, by giving rise to the bifurcations represented by the yellow vertices.

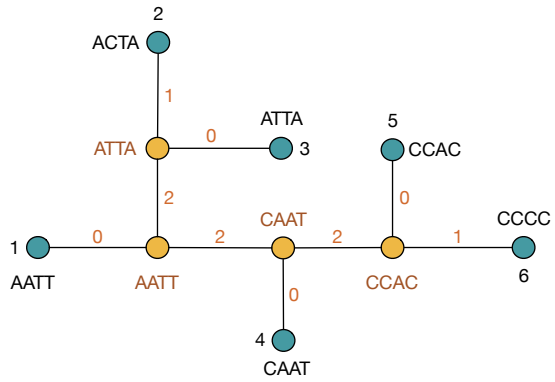


Fig. 2.7. An example of the most parsimonious phylogeny approximating the speciation process shown in Figure 2.6. The sequences associated to the observed taxa (dark blue vertices) are given and their Hamming distances are reported on Figure 2.8. The sequences associated to the internal vertices of the phylogeny (marked in yellow) and the edge weights of the phylogeny are marked in orange.

Problem 2 *(The phylogenetic estimation problem under the maximum parsimony criterion) Given a set Γ of n taxa described as molecular sequences of length l over an alphabet Σ, find (i) a phylogeny $T \in \mathcal{T}$ having Γ as leaf set and (ii) a set of $n-2$ molecular sequences of length l over Σ associated to the internal vertices of T, so as to minimize the following length function*

$$L_{Parsimony}(T) = \sum_{(u,v)\in E(T)} d_H(s_u, s_v).$$

As an example, Figure 2.7 shows a phylogeny T and a set of molecular sequences associated to the internal vertices of T (marked in yellow) that satisfies the above constraints and whose length is provably minimum under the maximum parsimony criterion for the six given taxa. Notice that the length of this phylogeny is 8 as opposed to 11 for the true speciation process of the same set of taxa in Figure 2.6. Also note how the edges of a phylogeny can be used to represent (possibly macro) speciation events over time and that the weights associated to these edges may be used to estimate the changes related to these events.

It is easy to see that Problem 2 is a particular case of Problem 1 in which $\Lambda_\Gamma = L_{Parsimony}$ and no additional constraint is present.

$$\begin{array}{c} 1\ 2\ 3\ 4\ 5\ 6 \\ \begin{pmatrix} 0 & 2 & 2 & 2 & 4 & 4 \\ 2 & 0 & 1 & 4 & 3 & 3 \\ 2 & 1 & 0 & 4 & 4 & 4 \\ 2 & 4 & 4 & 0 & 2 & 3 \\ 4 & 3 & 4 & 2 & 0 & 1 \\ 4 & 3 & 4 & 3 & 1 & 0 \end{pmatrix} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

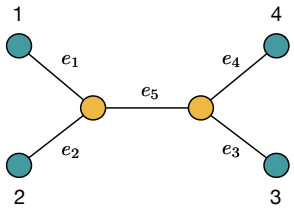
Fig. 2.8. Hamming distances associated to the six taxa shown in Figures 2.6 and 2.7.

2.3.2 The minimum evolution criterion

In 1978, Felsenstein [1978] showed that the edge weight estimation models based on Hamming distances were in general *statistically inconsistent*, i.e., unable to recover the true evolutionary process of the observed taxa under specific theoretical conditions (see [Felsenstein, 2004] for supplementary details and [Schulmeister, 2004] for recent advances on this topic). Because evolution is in general an unobservable process, using models that are provably statistical consistent, i.e., able to recover the true evolutionary process of taxa, becomes highly desirable. Hence, several authors – including, among others, Rzhetsky and Nei [1993, 1994, 1992], Gascuel [2005], Gascuel et al. [2001], Gascuel [2000], Denis and Gascuel [2003], and Gascuel and McKenzie [2004] – started to investigate alternative models, enjoying the statistically consistency of their predictions but still keeping some sort of relation with lex parsimoniae. These authors considered and extended the *Minimum Evolution* (ME) framework proposed by Kidd and Sgaramella-Zonta [1971] in the early 1970s, in which the evolutionary relationships between the taxa in Γ were expressed in terms of measures of their dissimilarity instead of their molecular sequences [Felsenstein, 2004, Gascuel, 2005, Saitou and Imanishi, 1989]. In particular, Kidd and Sgaramella-Zonta introduced the concept of *true evolutionary distances* (in their words *actual evolutionary distances*) between pairs of taxa $i, j \in \Gamma$ (i.e., the distances that one would obtain if all the molecular data from taxa were available). Moreover, they linked the true evolutionary distances with the estimated evolutionary distances d_{ij} from taxa (i.e., the entries of $\mathbf{D}$ defined in the beginning of this chapter). To this end, the authors introduced the conditions that "the amount of evolution observed to separate two populations is equal to the sum of the amounts of evolution from those two populations to their common ancestral population" and "Since [the edge] lengths are in a Euclidean space, the minimum possible value of each is 0." Then, they concluded: "Even though a complete demonstration is lacking, it does seem reasonable that under these conditions the best estimate of the actual evolutionary relationships is the one which requires the least evolution." To assess which phylogeny offers "the least evolution" they noted that "[...] we obtain quantitative 'solutions' for given tree structures and then choose that structure that allows the 'best solution' [...]" and "the tree with the lowest ME length among those evaluated is the 'best.' " We can reformulate their statements into a conjecture, which was proved true by Rzhetsky and Nei [1993], under Kidd and Sgaramella-Zonta's assumption that the "evolution occurred independently in each ancestral population":

Conjecture 1 *(Kidd and Sgaramella-Zonta [1971])*
If the evolutionary distances are unbiased estimates of the corresponding true evolutionary distances, then the true phylogeny of taxa *has an expected length shorter than any other possible phylogeny $T \in \mathcal{T}$ satisfying the following condition: for each pair of taxa i and j in Γ, the sum of the (positive) edge weights belonging to the path p_{ij} in T from i to j is greater than or equal to their estimated evolutionary distance d_{ij}.*

Farris [1970] and Fitch [1971]'s parsimony criterion as well as Kidd and Sgaramella-Zonta [1971]'s minimum evolution share the idea of searching for the phylogeny having minimum



(a)

	Edges				
Paths	e_1	e_2	e_3	e_4	e_5
p_{12}	1	1	0	0	0
p_{13}	1	0	1	0	1
p_{14}	1	0	0	1	1
p_{23}	0	1	1	0	1
p_{24}	0	1	0	1	1
p_{34}	0	0	1	1	0

(b)

Fig. 2.9. (a) An example of a phylogeny of four taxa and the associated edge-path incidence matrix of a tree (b).

length. However, they mainly differ in a subtle yet fundamental aspect: while the parsimony criterion is based on an abductive heuristic, i.e., the plausibility of the simplest hypothesis with respect to the more complex ones, ME lays its theoretical foundations on Kidd and Sgaramella-Zonta [1971] and Rzhetsky and Nei [1993]' result that ensures the statistical consistency of the minimum length phylogeny, provided that the evolutionary distances are unbiased estimates of the true evolutionary distances from taxa.

Estimating phylogenies by ME implies determining both the evolutionary distances from taxa and the edge weights associated to each phylogeny $T \in \mathcal{T}$. Estimation of evolutionary distances is usually carried out by means of probabilistic models of sequence evolution, such as those extensively discussed in [Felsenstein, 2004, Yang, 2014, Gascuel, 2005]. These models proved to be particularly effective in capturing the dynamics of substitution events affecting molecular sequences over time as well as successful in quantifying the amount of change that may have separated a pair of target taxa throughout their evolution [Yang, 2014]. We refer the reader interested in deepening the fundamental theory on models and algorithms for computing these distances to [Felsenstein, 2004, Yang, 2014, Gascuel, 2005]. Concerning estimation of the edge weights, the earliest approaches described in the literature proposed the use of Cavalli-Sforza and Edwards's *Ordinary Least-Squares* (OLS) model [Rzhetsky and Nei, 1993]. This model encodes a phylogeny $T \in \mathcal{T}$ by means of an *Edge–Path incidence matrix of a Tree* (EPT) (see [Nemhauser and Wolsey, 1999] for a general introduction and [Catanzaro et al., 2009] for detailed properties) i.e., a network matrix $\mathbf{Y}$ having a row for each (unique) path in T connecting a given pair of distinct taxa and a column for each edge. The generic entry y_{ij}^e of $\mathbf{Y}$ is equal to 1 if edge e belongs to the path p_{ij} from taxon i to taxon j, and 0 otherwise. As an example, Figure 2.9(b) shows the EPT matrix corresponding to the phylogeny shown in Figure 2.9(a). The OLS model assumes that the *superposition principle* holds at any instant of the evolutionary process of taxa, i.e., that the generic distance d_{ij} can be considered as the sum of the estimated mutation events (i.e., the weights) accumulated along the edges belonging to the path p_{ij}. In other words, fixed both a phylogeny $T \in \mathcal{T}$ and its associated EPT matrix $\mathbf{Y}$, and defined w_e as the weight associated to edge e in T, the

OLS model asserts that

$$\sum_{e \in E(p_{ij})} y_{ij}^{e} w_e = d_{ij} \qquad \forall\ i,j \in \Gamma : i \neq j. \tag{2.7}$$

In general, (2.7) may not admit solutions, hence Cavalli-Sforza and Edwards proposed to estimate $\mathbf{w}$ by means of the least-squares method, i.e., they suggested that the values $\rho_{ij} = \sum_{e \in E(p_{ij})} y_{ij}^{e} w_e$, $i,j \in \Gamma$, $i \neq j$, should minimize the function

$$\sum_{\substack{i,j \in \Gamma \\ i \neq j}} \left(d_{ij} - \rho_{ij}\right)^2 = \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \left(d_{ij} - \sum_{e \in E(p_{ij})} y_{ij}^{e} w_e \right)^2$$

encoding the error related to the approximation of the whole evolutionary process of taxa by means of a sum of mutation events accumulated along the paths of a given phylogeny. This condition holds when

$$\mathbf{w} = (\mathbf{Y}^t \mathbf{Y})^{-1} \mathbf{Y}^t \mathbf{D}^{\triangle} = \mathbf{Y}^{\dagger} \mathbf{D}^{\triangle} \tag{2.8}$$

i.e., when

$$w_e = \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij}, \quad \forall\ e \in E(T), \tag{2.9}$$

where $\mathbf{w} = \{w_e\}$ is the $(2n-3)$-edge weight vector associated to $\mathbf{Y}$; $\mathbf{D}^{\triangle}$ is a $n(n-1)/2$ vector whose components are obtained by taking row by row the entries of the strictly upper triangular matrix part of matrix $\mathbf{D}$; and $\mathbf{Y}^t$ and $\mathbf{Y}^{\dagger} = \{y_{ij}^{e\dagger}\}$ are the transpose and the Moore-Penrose pseudo-inverse of matrix $\mathbf{Y}$, respectively. Thus, when the OLS model is used to estimate the edge weights of a phylogeny, ME can be formalized as follows:

Problem 3 *(The minimum evolution problem under the ordinary least-squares edge weight estimation model)*
Given a set Γ of taxa, find (i) a phylogeny $T \in \mathcal{T}$, described in terms of its EPT matrix $\mathbf{Y}(T)$, having Γ as leaf set and (ii) the weights of the edges of T, such that the following length function is minimized

$$L_{ME}(T) = \sum_{e \in E(T)} w_e$$

subject to the additional constraints

$$w_e = \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij}, \quad \forall\ e \in E(T). \tag{2.10}$$

It is easy to see that Problem 3 is a particular case of Problem 1 where $\Lambda_{\Gamma} = L_{ME}$, the constraints imposed by Ψ_{Γ} are $w_e - \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij} = 0$, for each edge $e \in E(T)$, and constraints $\Omega_{\Gamma}(T) \geq 0$ are not present.

Rzhetsky and Nei [1993] showed that if the distance matrix $\mathbf{D}$ is an unbiased estimate of the true distance matrix from taxa, then the optimal solution to (2.10) is statistical consistent. This milestone result proved Kidd and Sgaramella-Zonta [1971]'s conjecture and constitutes the main pillar of the theoretical foundations of ME.

2.3.3 The balanced minimum evolution criterion

Rzhetsky and Nei [1993] observed an undesirable aspect related to the optimal solution to (2.10): whenever $\mathbf{D}$ is a biased estimate of the true distance matrix from taxa, the optimal solution to (2.10) may be characterized by negative edge weights, which are void of biological meaning [Gascuel, 2005]. The occurrence of negative edge weights may be interpreted as the presence of (possibly concurrent) wrong assumptions, such as the wrong choice of the probabilistic model used to estimate the evolutionary distances from taxa and/or the non-subsistence of the superposition principle throughout the evolution of taxa. Many authors (see, e.g., [Felsenstein, 1997, Gascuel and Levy, 1996, Hubert and Arabie, 1995, Makarenkov and Leclerc, 1999]) tried to find ways around the occurrence of negative edge weights, including exploring their possible biological interpretations [Felsenstein, 2004] as well as the design of methods and algorithms to include in (2.10) the non-negativity constraint $w_e \geq 0$, for all $e \in E(T)$, i.e., solving the following problem:

Problem 4 *(The minimum evolution problem under the ordinary least-squares edge weight estimation model with positivity constraint)*
Given a set Γ of taxa, find (i) a phylogeny $T \in \mathcal{T}$, described in terms of its EPT matrix $\mathbf{Y}(T)$, having Γ as leaf set and (ii) the non-negative weights of the edges of T, such that the following length function is minimized

$$L_{MEPC}(T) = \sum_{e \in E(T)} w_e$$

subject to the additional constraints

$$w_e \geq \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij}, \quad \forall\ e \in E(T)$$

$$w_e \geq 0, \quad \forall\ e \in E(T).$$

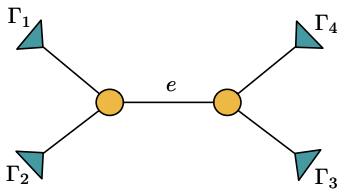


Fig. 2.10. An internal edge e of a phylogeny T and the associated and four subtrees inducing a partition of Γ into the subsets Γ_1, Γ_2, Γ_3, and Γ_4, respectively.

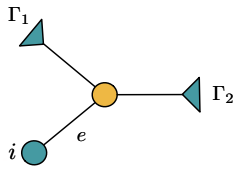


Fig. 2.11. An external edge e of a phylogeny T and the associated and three subtrees inducing a partition of Γ into the subsets $\{i\}$, Γ_1, and Γ_2, respectively.

However, as observed in [Gascuel, 2005], these research efforts did not lead to satisfying conclusions. The publication of Pauplin's article in 2000 marked a turning point on this issue [Pauplin, 2000]. Specifically, the author questioned the biological interpretation of the dependency of the entries of vector $\mathbf{w}$ on the topology encoded by $\mathbf{Y}$ under the OLS model. This dependency can be described by means of the phylogeny shown in Figure 2.10: for each internal edge $e = (u, v) \in E(T)$ of a fixed phylogeny T, consider the four subsets of taxa Γ_1, Γ_2, Γ_3, and Γ_4 belonging to the four subtree rooted in the vertices that are endpoints of the edges incident to u and v; then, it holds that $\bigcup_{i=1}^{4} \Gamma_i = \Gamma$ and $\Gamma_i \cap \Gamma_j = \emptyset$, for all distinct $i, j \in \{1, \ldots, 4\}$; then, under Rzhetsky and Nei's OLS model, the weight associated to edge e can be computed as

$$w_e = \frac{1}{2}\left[\lambda\left(\delta_{\Gamma_1|\Gamma_4} + \delta_{\Gamma_2|\Gamma_3}\right) + (1-\lambda)\left(\delta_{\Gamma_1|\Gamma_3} + \delta_{\Gamma_2|\Gamma_4}\right) - \left(\delta_{\Gamma_1|\Gamma_2} + \delta_{\Gamma_3|\Gamma_4}\right)\right] \tag{2.11}$$

where $|\Gamma_i|$ denotes the cardinality of Γ_i, $i \in \{1, \ldots, 4\}$, λ is defined as

$$\lambda = \frac{|\Gamma_1||\Gamma_3| + |\Gamma_2||\Gamma_4|}{|\Gamma_1 \cup \Gamma_2| \cdot |\Gamma_3 \cup \Gamma_4|} \tag{2.12}$$

and the *average distance* $\delta_{\Gamma_i|\Gamma_j}$ between the subsets of taxa $\Gamma_i, \Gamma_j \subset \Gamma$ is defined as

$$\delta_{\Gamma_i|\Gamma_j} = \frac{1}{|\Gamma_i||\Gamma_j|} \sum_{\substack{r \in \Gamma_i \\ s \in \Gamma_j}} d_{rs}. \tag{2.13}$$

Similarly, in the case of an external edge, such as e in Figure 2.11, the corresponding weight can be computed as

$$w_e = \frac{1}{2}\left(\delta_{\{i\}|\Gamma_1} + \delta_{\{i\}|\Gamma_2} - \delta_{\Gamma_1|\Gamma_2}\right). \tag{2.14}$$

Equations (2.11) and (2.14) tell that the edge weights of a given phylogeny T can be stated exclusively in terms of average (estimated evolutionary) distances between subsets of taxa. The value of λ is correlated to the topology of T encoded by $\mathbf{Y}$ by means of the Moore-Penrose pseudo-inverse and acts as a weight in (2.11). Now, Pauplin [2000] observed that, the average distance (2.13) can be equivalently defined in the following recursive way:

$$\delta_{\Gamma_i|\Gamma_j} = \begin{cases} d_{ij} & \text{if } \Gamma_i = \{i\} \text{ and } \Gamma_j = \{j\}; \\ \frac{|\Gamma_p|}{|\Gamma_j|}\delta_{\Gamma_i|\Gamma_p} + \frac{|\Gamma_q|}{|\Gamma_j|}\delta_{\Gamma_i|\Gamma_q} & \text{otherwise,} \end{cases} \tag{2.15}$$

with $\Gamma_j = \Gamma_p \cup \Gamma_q$. The author observed that this new expression makes it explicit that under the OLS model certain edges of a phylogeny may be weighted more than others and this fact, in general, is not supported by any biological evidence or rationale. Hence, given a phylogeny T of Γ, two disjointed subtrees of T having Γ_i and Γ_j as respective subsets of taxa, and two further subsets of taxa Γ_p and Γ_q such that $\Gamma_j = \Gamma_p \cup \Gamma_q$, Pauplin proposed to replace equation (2.15) by the following recursive formula

$$\delta^T_{\Gamma_i|\Gamma_j} = \begin{cases} d_{ij} & \text{if } \Gamma_i = \{i\} \text{ and } \Gamma_j = \{j\}; \\ \frac{1}{2}\delta^T_{\Gamma_i|\Gamma_p} + \frac{1}{2}\delta^T_{\Gamma_i|\Gamma_q} & \text{otherwise.} \end{cases} \tag{2.16}$$

Pauplin's showed that under this definition of average distance, (2.11) becomes

$$w_e = \frac{1}{4}\left(\delta^T_{\Gamma_1|\Gamma_4} + \delta^T_{\Gamma_2|\Gamma_3} + \delta^T_{\Gamma_1|\Gamma_3} + \delta^T_{\Gamma_2|\Gamma_4}\right) - \frac{1}{2}\left(\delta^T_{\Gamma_1|\Gamma_2} + \delta^T_{\Gamma_3|\Gamma_4}\right), \tag{2.17}$$

(2.14) becomes

$$w_e = \frac{1}{2}\left(\delta^T_{\{i\}|\Gamma_1} + \delta_{\{i\}|\Gamma_2} - \delta^T_{\Gamma_1|\Gamma_2}\right), \tag{2.18}$$

and the length of the given phylogeny T becomes precisely (2.1), i.e., a mere function of both the topology of T and the estimated evolutionary distances. Thus, the problem of finding the minimum length phylogeny under Pauplin's edge weight estimation model, usually referred to as Balanced Minimum Evolution [Gascuel, 2005], can be stated as follows:

Problem 5 *(The balanced minimum evolution problem)*
Given a set Γ of taxa, find an phylogeny $T \in \mathcal{T}$ having Γ as leaf set and minimizing the following length function

$$L(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{d_{ij}}{2^{\tau_{ij}}}. \tag{2.19}$$

It is easy to see that Problem 5 is a particular case of Problem 1 where $\Lambda_\Gamma = L(T)$ and no additional constraint is present apart from the implicit one $T \in \mathcal{T}$. The attribute "balanced" assigned to Problem 5 indicates that the number of times each evolutionary distance appears in the summation of all edge weights is normalized, by reflecting so the fact that all of the edge weights are treated (or "weighted") in the same way. As shown in Desper and Gascuel [2004], the phylogeny having the shortest BME length (i.e., the optimal solution to the BMEP) is statistically consistent and its edge weights are always positive whenever the entries d_{ij} are actually distances, i.e., they satisfy the triangle inequality. Moreover, the topological accuracy of the optimal solution to the BMEP currently constitutes the state-of-the-art, being far better than any other distance method described in the literature [Vinh

and von Haeseler, 2005, Lefort et al., 2015]. The overall good performance of the BME can be possibly explained by the fact that it acts as a *Weighted Least-Squares* (WLS) edge weight estimation model, i.e., as a model in which the vector $\mathbf{w}$ is estimated as

$$\mathbf{w} = (\mathbf{Y}^t\mathbf{V}^{-1}\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{V}^{-1}\mathbf{D}^{\triangle}, \tag{2.20}$$

where $\mathbf{V}$ is a $n(n-1)/2 \times n(n-1)/2$ diagonal matrix whose generic diagonal entry v_{ij} encodes the variance of the evolutionary distance d_{ij}. In particular, when $v_{ij} \propto 2^{\tau_{ij}}$, i.e., when the variance v_{ij} is proportional to the exponential of the path-length τ_{ij}, then the length of the phylogeny encoded by $\mathbf{Y}$ becomes precisely (2.1). Thus, the BME acts de facto as a minimum variance tree length estimator, with variances that are consistent in a biological context when large evolutionary distances have a much higher variance than the shorter ones (see [Gascuel, 2005] for more information).

The need to analyze larger and larger molecular datasets by guaranteeing at the same time a certificate of statistical consistency of the solution or a bound on the error related to the approximation of the optimal solution to the BMEP, motivated over two decades of research efforts on the problem. The next sections will review and classify the outcomes of these efforts, by starting from the combinatorics of the BMEP.

2.4 Combinatorial and optimization aspects of the BMEP

The compact statement of the BMEP hides deep optimization aspects that relate its combinatorics to those of well-known problems, such as the *Traveling Salesman Problem* (TSP) [Lawler et al., 1985] and the *Huffman Coding Problem* (HCP) [Parker and Ram, 1996, Catanzaro et al., 2012, 2020b, 2013a, 2015]. In this section we review these aspects, by starting from Semple and Steel's interpretation of the BMEP length function (2.1) and by subsequently discussing the complexity aspects of the problem as well as the recent advances

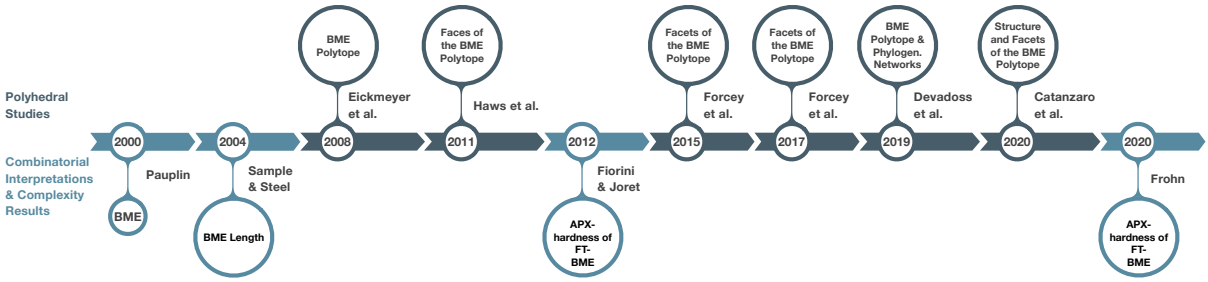


Fig. 2.12. Graphical representation of the development of the BMEP's combinatorics, since Pauplin [2000]'s article. The dark gray stages represent articles studying the polyhedral combinatorics of the BME polytope, including Eickmeyer et al. [2008]'s conjectures, Haws et al. [2011]'s faces, Forcey et al. [2015]'s complete description of the BME polytope for $n = 5$, Forcey et al. [2017]'s further facets of the BME polytope and combinatorial connections with the permutoassociahedron, Devadoss et al. [2019]'s connections of the BME polytope with split-networks, and Catanzaro et al. [2020b]'s recent advances. The blue stages represent articles presenting the combinatorial interpretation of the BME length function [Semple and Steel, 2004] as well as complexity results [Fiorini and Joret, 2012, Frohn, 2020].

on its polyhedral combinatorics. Figure 2.12 graphically summarizes the main stages of this development.

2.4.1 On the combinatorial interpretation of the BMEP length function

As observed in the previous sections, the BME allows to express the length function of a phylogeny T just in terms of estimated distances d_{ij} and path-lengths τ_{ij} between pair of distinct taxa $i, j \in \Gamma$. This aspect, together with the curious form of the term $2^{\tau_{ij}}$, led Semple and Steel [2004] to investigate the combinatorial nature of the BMEP objective function. The authors started from Makarenkov and Leclerc [1997]'s notion of a *circular order* associated to a given phylogeny T of a set Γ of n taxa, i.e., a particular type of permutation π of the taxa in Γ constructed by considering a graphic planar representation of T and by performing the following steps: choose arbitrarily a leaf/taxon of T and set it as the first element of π; then, index the remaining taxa according to a *circular* (i.e., clockwise) scanning of the leaves (taxa) of T [Makarenkov and Leclerc, 1997]. As an example, consider the phylogeny T of six taxa in Γ shown in Figure 2.13(a) and suppose to choose taxon 1. Then, the planar representation of T defines the circular order $\pi_a = (1, 2, 3, 5, 6, 4)$. In the case of Figure 2.13(b), the circular order becomes $\pi_b = (1, 4, 5, 6, 3, 2)$. In both figures these circular orders are highlighted by the black dashed arrows that scan clockwise the leaves of the phylogeny. In contrast, the permutation $\pi_c = (1, 2, 5, 3, 4, 6)$ given by the black dashed arrows depicted in Figure 2.13(c) is not a circular order, e.g., because the positions of taxa 3 and 5 in π should be swapped (taxon 3 precedes in clockwise terms taxon 5 in the given phylogeny). We observe that the notion of a circular order is strictly related to the notion of Eulerian circuit on a phylogeny T. In fact, by scanning clockwise the taxa of T in the planar embedding, each edge of T is traversed exactly twice before coming back to the initial taxon. Hence, given an Eulerian circuit on T (e.g., the one obtained from orange dashed arrows in Figure 2.13(a)), a possible circular order can be

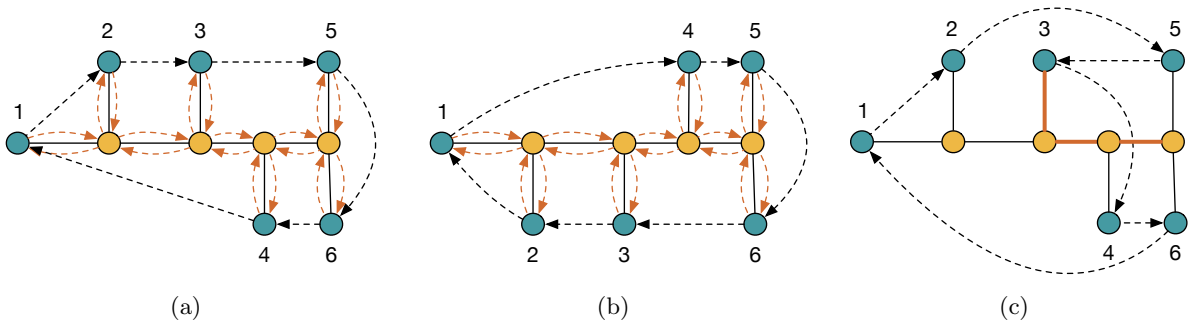


Fig. 2.13. (a) By starting from taxon 1 and by following the black dashed arrows, we obtain the permutation $\pi_a = (1, 2, 3, 5, 6, 4)$, which constitutes an example of a circular order on the given phylogeny T of six taxa. This permutation can be seen as derived from an Eulerian circuit on T (shown in orange dashed arrows) by removing the internal vertices that figures in the circuit. (b) An alternative example of a circular order on the same phylogeny. (c) An example of a permutation π of Γ that is not a circular order on the same phylogeny T: the circuit through T induced by π forces to cross the bold edges more than twice.

obtained by choosing arbitrarily the first taxon and by subsequently removing the internal vertices of the phylogeny T that figures in the Eulerian circuit.

Let $\pi(i)$ denote the i-th element of a given permutation π of Γ. Also assume, by convention, that with $\pi(n+1) = \pi(1)$. Then, Semple and Steel observed (i) that a circular order decomposes T precisely into n paths $p_{\pi(i)\pi(i+1)}$, $i \in \{1, \ldots, n\}$; (ii) that every edge that is incident to a taxon occurs in exactly two of the paths $p_{\pi(1)\pi(2)}, \ldots, p_{\pi(n)\pi(1)}$; and (iii) that every edge that is incident only to internal vertices occurs in a positive and even number γ of paths $p_{\pi(1)\pi(2)}, \ldots, p_{\pi(n)\pi(1)}$. Then, the authors showed that a permutation π of Γ is a circular order on T if and only if $\gamma = 2$. As an example, this property holds true for $\pi_a = (1, 2, 3, 5, 6, 4)$ in Figure 2.13(a) whereas it is not satisfied by $\pi_c = (1, 2, 5, 3, 4, 6)$ in Figure 2.13(c), as the bold internal edges occur more than twice in $p_{\pi(1)\pi(2)}, \ldots, p_{\pi(n)\pi(1)}$. Semple and Steel also observed that for a fixed phylogeny T there exists 2^{n-2} out of $n!$ permutations of the set Γ that are circular orders on T. Based on these results Semple and Steel proved that if

$$d_{ij} = \sum_{e \in E(p_{ij})} w_e \quad \forall\ i, j \in \Gamma \tag{2.21}$$

then

$$2L(T) = 2\sum_{e \in E} w_e = \sum_{i=1}^{|\Gamma|} d_{\pi(i)\pi(i+1)}.$$

In other words, if (2.21) holds true, the length function of the BMEP is equal to half the length of the Hamiltonian cycle encoded by a circular order π of Γ. Moreover, the authors also proved that, because the length of a phylogeny does not depend on the choice of π, also the following equation holds true:

$$L(T) = \frac{1}{|\Pi(T)|} \sum_{\pi \in \Pi(T)} \left(\frac{1}{2} \sum_{k=1}^{n} d_{\pi(k)\pi(k+1)} \right) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{\mu_{ij}}{2} d_{ij} \tag{2.22}$$

where $\Pi(T)$ denotes the set of the circular orders of Γ induced by T and μ_{ij} is the fraction of circular orders in which taxon j immediately follows taxon i. By observing that $(n-2) - |V(p_{ij}) \setminus \{i, j\}| = (n-2) - (\tau_{ij} - 1)$ is the number of internal vertices of T not belonging to the path p_{ij}, we can conclude that the number of circular orders that in which taxon j immediately follows taxon i is equal to $2^{(n-2)-(\tau_{ij}-1)}$. Then, by observing that

$$\frac{\mu_{ij}}{2} = \frac{2^{(n-2)-(\tau_{ij}-1)}}{2|\Pi(T)|} = \frac{2^{(n-2)-(\tau_{ij}-1)}}{2^{n-1}} = 2^{-\tau_{ij}},$$

it follows that (2.22) becomes precisely (2.19).

2.4.2 Complexity results

Deciding the complexity of the BMEP remained an open problem for more than a decade since its introduction. Fiorini and Joret [2012] succeeded in showing that the BMEP is $\mathcal{NP}$-hard and inapproximable within c^n, for some constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$. Their proof

uses a reduction from the *3-Colorability Problem* (3CP) (i.e., the problem of deciding whether the vertex set of a given undirected graph G can be partitioned into three stable sets) [Garey and Johnson, 2003], and consists of showing the existence of a bijection between the set of colors of the 3CP and the vertices of a phylogeny constituted by a *centroid* [Jordan, 1869] and three caterpillars connected to it, so that G is three-colorable if and only if the length of the phylogeny so constructed satisfies specific criteria.

In Chapter 4 we show that the BMEP remains $\mathcal{APX}$-hard even if the topology of the phylogeny is fixed and one wants to assign the taxa to the leaves of this fixed topology so as to minimize the BME length function. This result on this particular restriction of the BMEP, known in the literature as the *Fixed-Tree BMEP* (FT-BMEP) [Aringhieri et al., 2011], seems to indicate the presence of strong ties between the combinatorics of the BMEP and the *Quadratic Assignment Problem* (QAP) [Çela, 1997]. Exploring systematically these ties could provide new approaches to solution of the BMEP.

2.4.3 On the polyhedral combinatorics of the BMEP

The $\mathcal{NP}$-hardness of the BMEP has justified two decades of research efforts aiming to characterize both fundamental properties of phylogenies and the convex hull of the feasible solutions to the BMEP (hereafter referred to as the *BME polytope*) with a view to developing effective exact solution algorithms able to tackle larger and larger instances of the problem. In this section we summarize the major achievements obtained so far on the polyhedral combinatorics of the BMEP, by starting from the review of the fundamental properties that characterize the set Θ of the path-length matrices τ that encode the phylogenies of Γ.

As a first step, we observe that, because any path-length matrix $\tau \in \Theta$ encodes path-lengths on an UBT, its entries τ_{ij} must be integers in the discrete interval $\{2, \ldots, n-1\}$, satisfying the following necessary conditions [Catanzaro et al., 2020b]:

$$\tau_{ii} = 0 \qquad \forall\, i \in \Gamma \qquad \text{(null diagonal path-lengths)} \tag{2.23}$$

$$\tau_{ij} = \tau_{ji} \qquad \forall\, i, j \in \Gamma : i < j \qquad \text{(symmetry equalities on path-lengths)} \tag{2.24}$$

$$\tau_{ij} + \tau_{jk} - \tau_{ik} \geq 2 \qquad \forall\, i, j, k \in \Gamma : i \neq j \neq k \qquad \text{(triangular inequalities on path-lengths).} \tag{2.25}$$

Moreover, because any path-length matrix $\tau \in \Theta$ must encode an unrooted binary tree, the rows of τ must satisfy *Kraft's equalities* [Catanzaro et al., 2012, Parker and Ram, 1996, Catanzaro et al., 2020b]:

$$\sum_{j \in \Gamma \setminus \{i\}} 2^{-\tau_{ij}} = \frac{1}{2} \quad \forall\, i \in \Gamma. \tag{2.26}$$

Kraft's equalities carry a specific combinatorial meaning. Specifically, if Kraft's equality holds on the i-th row of a path-length matrix τ, then the i-th row encodes a binary tree rooted in taxon i (see, e.g., Figure 2.14). In other words, each of the Kraft's equalities allows to capture

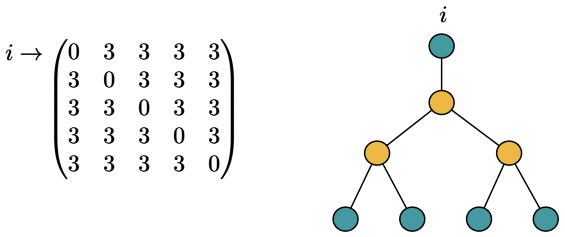


Fig. 2.14. The i-th row of the τ matrix on the left encodes the binary tree rooted in taxon i on the right. Note how the entries of this specific row satisfy (2.26).

in a closed form both the connectivity constraint typical of acyclic graphs and the respect of the degree constraint on the internal vertices of such a graph. If a τ matrix satisfies (2.26) then its rows describe a forest of n binary trees each rooted in the respective taxon $i \in \Gamma$. The trees in this forest, however, do not encode in general the same phylogeny. For example, all of the rows of the path-length matrix of Figure 2.14 satisfy (2.26) but the resulting forest of trees does not encode the same phylogeny: each taxon i wants to have a path-length 3 from all of the remaining taxa and this is not congruent with the fact of having a unique phylogeny. Kraft's equalities are therefore just necessary conditions that must be satisfied by any path-length matrix $\tau \in \Theta$.

A further condition that a path-length matrix $\tau \in \Theta$ must satisfy arises from the relationships between the length function of the BMEP and information entropy (see Chapter 3) and it is usually referred to as the equation of the *phylogenetic manifold* [Catanzaro et al., 2012]:

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij} 2^{-\tau_{ij}} = (2n - 3). \tag{2.27}$$

This equation ties the entries of all of the rows of any path-length matrix $\tau \in \Theta$ and, in the context of information theory, it describes a manifold having cross entropy equal to the number of edges of a phylogeny of n taxa (i.e., $2n - 3$). We discuss this aspect in more detail in Chapter 3.

Equations (2.23)-(2.27) are independent. As an example, the following minimal-dimension matrix

$$\begin{pmatrix} 0 & 3 & 3 & 3 & 3 \\ 3 & 0 & 3 & 3 & 3 \\ 3 & 3 & 0 & 3 & 3 \\ 3 & 3 & 3 & 0 & 3 \\ 3 & 3 & 3 & 3 & 0 \end{pmatrix}$$

satisfies all the considered conditions but (2.27). In fact,

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = \frac{15}{2} \neq 7 = (2n - 3).$$

Similarly, the following matrix

$$\begin{pmatrix} 0 & 5 & 5 & 5 & 4 & 5 \\ 5 & 0 & 2 & 2 & 3 & 4 \\ 5 & 2 & 0 & 2 & 3 & 4 \\ 5 & 2 & 2 & 0 & 3 & 4 \\ 4 & 3 & 3 & 3 & 0 & 4 \\ 5 & 4 & 4 & 4 & 4 & 0 \end{pmatrix}$$

satisfies all the considered conditions but (2.26). In fact, we have that

$$\sum_{j \in \Gamma_1} 2^{-\tau_{1j}} = 0.1875 \neq \frac{1}{2}.$$

Furthermore, the following minimal-dimension matrix

$$\begin{pmatrix} 0 & 3 & 5 & 5 & 2 & 4 \\ 3 & 0 & 3 & 4 & 4 & 3 \\ 5 & 3 & 0 & 2 & 5 & 4 \\ 5 & 4 & 2 & 0 & 5 & 3 \\ 2 & 4 & 5 & 5 & 0 & 3 \\ 4 & 3 & 4 & 3 & 3 & 0 \end{pmatrix}$$

satisfies all of the considered conditions but the triangular inequalities (2.25). In fact, $\tau_{12} + \tau_{12} - \tau_{13} = 1 < 2$.

It is worth noting that because any path-length matrix $\tau \in \Theta$ must encode a tree, the triangular inequalities (2.25) can be further generalized by means of Buneman's *additive property* [Buneman, 1971, Erdös et al., 1999, Waterman et al., 1977], which states that a given graph is a tree if and only if any quartet of its vertices, say i, j, p and q, satisfies exactly one of the following conditions:

$$\begin{cases} \tau_{ij} + \tau_{pq} + 2 \leq \tau_{iq} + \tau_{jp} = \tau_{ip} + \tau_{jq} \\ \tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq} \\ \tau_{ip} + \tau_{jq} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{iq} + \tau_{jp}. \end{cases} \tag{2.28}$$

It is easy to see that (2.25) is a restriction of (2.28) when i, j, p and q belong to Γ and e.g., $p = q$. Now, conditions (2.23)-(2.24) and (2.26)-(2.28) are necessary to characterize Θ, but proving (or disproving) the following conjecture still remains an open problem:

Conjecture 2 *Given a natural number $n \geq 3$, a set Γ of n items and a dissimilarity matrix $\tau \in \mathbb{R}^{n \times n}$ on Γ. If the entries of τ are in the interval $[2, n-1]$ and satisfy conditions* (2.23)-(2.28)*, then there exists an unique weighted unrooted binary tree such that $\sum_{e \in E(p_{ij})} w_e = \tau_{ij}$ for all $i, j \in \Gamma, i \neq j$.*

When studying this conjecture, we can make some observations:

(1) The Kraft equality (2.26) implies that the entries of τ are integers: suppose by contradiction, for $i \in \Gamma$, that the sum (2.26) includes a term 2^{-x} for a rational number x. Then, there exist co-prime integers a, b and c, d respectively, such that $x = a/b$ and $2^{-a/b} = c/d$. This means, $a = (\log_2(d) - \log_2(c)) \cdot b$. Hence, c and d are powers of two. This contradicts that c and d are co-prime. Thus, sum (2.26) only includes terms 2^{-x} for integers x.

(2) Theorem 2 in Buneman [1974] states that, since τ satisfies condition (2.28), there exists a weighted tree T which contains the items of Γ among its vertices such that $\sum_{e \in E(p_{ij})} w_e = \tau_{ij}$ for all $i, j \in \Gamma, i \neq j$.

(3) Γ is the set of leaves of T in observation (2): suppose by contradiction that there exists an item $k \in \Gamma$ which is an internal vertex in T. Then, $\tau_{ik} + \tau_{kj} = \tau_{ij}$. However, we know from condition (2.25) that $\tau_{ik} + \tau_{kj} \geq \tau_{ij} + 2$, leading to a contradiction.

(4) If equation (2.27) would be equal to $\sum_{e \in E(T)} w_e$ for T in observation (2), then the linear system

$$\sum_{e \in E(p_{ij})} w_e = \tau_{ij} \quad \forall\, i, j \in \Gamma, i < j$$
$$\sum_{e \in E(T)} w_e = 2n - 3$$

would yield the desired result of uniform edge weights $w_e = 1$ which would ensure that T is unrooted and binary. However, it is unclear how to incorporate interpretation (2.22) or other conditions to enable this scenario.

If Conjecture 2 should hold, then it would address the missing characterization of sufficient conditions for Θ which would make the development of compact mixed integer programming formulations for the BMEP possible.

The presence of nonlinearities in the above equations is an obstacle for the study of the polyhedral combinatorics of the BMEP and suggests the search for spaces in which at least one between (2.26) and (2.27) can be linearized. Forcey et al. [2015, 2017], and Catanzaro et al. [2020b] proposed to linearize Kraft's equalities, i.e., to study the polyhedral combinatorics of the BME polytope not in the space of the path-length matrices $\tau \in \Theta$, but rather in the space of the $n \times n$ matrices $\mathbf{X}$, whose generic entry

$$x_{ij} = 2^{n-1-\tau_{ij}} \quad \forall\, i, j \in \Gamma. \tag{2.29}$$

Specifically, by (2.29), each matrix $\mathbf{X}$ is symmetric, diagonal dominant and has nonnegative entries such that its rows and columns sum to $3 \cdot 2^{n-2}$. The bijection (2.29) induces the following $n(n+3)/2$ linear independent conditions on the entries of each matrix $\mathbf{X}$, which

are analogous to conditions (2.23), (2.24), and (2.26) for the path-length matrices $\tau \in \Theta$:

$$x_{ii} = 2^{n-1} \qquad \forall\ i \in \Gamma \tag{2.30}$$

$$x_{ij} = x_{ji} \qquad \forall\ i, j \in \Gamma : i \neq j \tag{2.31}$$

$$\sum_{j \in \Gamma_i} x_{ij} = 2^{n-2} \qquad \forall\ i \in \Gamma. \tag{2.32}$$

Note, in particular, how Kraft's equalities becomes linear in (2.32). Catanzaro et al. [2020b] defined $\mathcal{X}$ as the set of the matrices $\mathbf{X}$ encoding the phylogenies of Γ via (2.29) and denoted Space($\mathcal{X}$) as the vector space of minimal dimension that includes $\mathcal{X}$. The authors showed that there exists a particular set of caterpillar phylogenies with n leaves that form a basis for Space($\mathcal{X}$). Thanks to this result it was possible to establish fundamental structural properties of the BME polytope, including, from among others, the characterization of its vertices as well as its dimension, equal to

$$\binom{n}{2} - n. \tag{2.33}$$

From a historical perspective, the first use of Space($\mathcal{X}$) was introduced in the literature by Eickmeyer et al. [2008] in 2008, with the intent to characterize, by means of polyhedral arguments, the approximation ratio of a well-known constructive heuristic for the BMEP known as the *Neighbor-Joining* (NJ) tree (see [Saitou and Nei, 1987, Studier and Keppler, 1988, Gascuel and Steel, 2006] and Section 2.5.2). The authors considered the vectors

$$v_T = (\mu_{12}, \mu_{13}, \ldots, \mu_{(n-1)n})$$

associated to each phylogeny $T \in \mathcal{T}$ and conjectured that the starting point of the NJ tree, i.e., a star graph with n terminal leaves, must lay at the center of the convex hull of these vectors. The authors also conjectured that because the vectors v_T belong to $\mathbb{R}^{\binom{n}{2}}$ and their respective entries $(\mu_{ij})_{j \in \Gamma \setminus \{i\}}$ must satisfy Kraft's equalities (2.26) for each $i \in \Gamma$, then the dimension of the BME polytope should have been equal to (2.33). Three years later, a coauthor of Eickmeyer et al. [2008], namely Ruriko Yoshida, tried a second attack to the BME polytope in Haws et al. [2011], by providing the description of some faces of the BME polytope. The proofs of Eickmeyer et al. [2008]'s conjectures as well as the description of some fundamental facets of the BME polytope came almost a decade later, thanks to the results of Forcey et al. [2015, 2017] and, more recently, of Catanzaro et al. [2020b]. In particular, Forcey et al. [2015] studied the BME polytope in small dimension and they succeeded in characterizing all of its facets for $n = 5$ and some of the facets for $n = 6$. The authors did not use a standard polyhedral approach to assess the facet-defining nature of a given inequality. They rather used topological arguments such as the fact that in a phylogeny the maximum value of a path-length is $(n - 1)$; the fact that there can not exist three taxa having path-length two from another one; and the fact that a fixed (circular) assignment of taxa yields a lower bound on the length of a tour through the tree. These arguments, led to three families of facet-defining inequalities for $n = 5$, two of which appear in Table 2.1. Remarkably, Forcey et al. [2015] showed that some of these inequalities remain facet-defining even for generic n.

Facet Number	Facet-defining inequalities	
1	$x_{ij} \geq 1$	For all distinct $i, j \in \Gamma$
2	$x_{ij} + x_{jk} - x_{ik} \leq 2^{n-3}$	For all distinct $i, j, k \in \Gamma$
3	$\sum_{i,j \in S_1, i<j} x_{ij} \leq (k-1)2^{n-3}$	For any $S_1, S_2 \subset \Gamma$: $S_1 \cap S_2 = \emptyset, \lvert S_1 \rvert = 3$
4	$2^{n-5}x_{i_1i_2} + 2^{n-5}x_{i_1i_3} + x_{i_2i_3} \geq 5 \cdot 2^{n-5}$	For all distinct $i_1, i_2, i_3 \in \Gamma$
5	$2^{n-4}x_{i_1i_2} + 2^{n-4}x_{i_1i_3} + x_{i_2i_3} \geq 2^{n-2}$	
6	$2^{n-4}x_{i_1i_2} - x_{i_3i_4} \geq 0$	For all distinct $i_1, i_2, i_3, i_4 \in \Gamma$
7	$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5} \geq (4+\rho)2^{k-1}$	For all distinct $i_1, i_2, i_3, i_4, i_5 \in \Gamma$ $k = \lfloor \frac{n}{3} \rfloor, \rho = n - 3k$
8	$2x_{i_1i_2} + 2x_{i_3i_4} + x_{i_5i_6} \geq 8$	For all distinct $i_1, i_2, i_3, i_4, i_5, i_6 \in \Gamma$
9	$x_{i_1i_2} + x_{i_3i_4} - 2^{n-4}x_{i_5i_6} \leq 2^{n-4}$	
10	$2x_{i_1i_2} + x_{i_2i_3} + x_{i_3i_4} + x_{i_4i_5} + x_{i_5i_6} + x_{i_1i_6} + x_{i_2i_4} \geq 16$	
11	$-2x_{i_1i_2} + 7x_{i_1i_3} + 7x_{i_1i_4} + 7x_{i_2i_5} + 7x_{i_2i_6} \geq 40$	
12	$\rho x_{i_1i_n} + \sum_{r=1}^{k} x_{i_{2r-1}i_{2r}} \geq 2^{n-k-3}(k+2-\rho) + \rho$	For all distinct $i_1, i_2, \ldots, i_n \in \Gamma$ $k = \lfloor \frac{n}{2} \rfloor, \rho = n - 2k$
	Valid but not facet-defining inequalities	
13	$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} \geq (3+\rho)2^{k-1}$	$k = \lfloor \frac{n}{3} \rfloor, \rho = n - 3k$

Table 2.1. List of the known facet-defining inequalities of the BME polytope when assuming $n = 6$. The references for facets 1 to 3 and 4 to 13 are Forcey et al. [2017] and Catanzaro et al. [2020b], respectively.

Forcey et al. also investigated the combinatorial similarities relating the BME polytope to other polytopes, with a view to identify new facet-defining inequalities for the problem. In Forcey et al. [2015] they studied the connections between the BME polytope and the *Birkhoff polytope*, by showing their combinatorial equivalence for $n = 5$. In Forcey et al. [2017], the authors focused on the connections between the BME polytope and the *permutoassociahedron* $\mathcal{KP}_n$ [Billera et al., 2001, Forcey et al., 2017, Kapranov, 1993, Reiner and Ziegler, 1994] and, by projecting 2^{n-2} vertices of the $\mathcal{KP}_n$ onto the BME polytope, the authors succeeded in establishing a surjection between families of their respective faces that enabled the identification of the so called *split-facets* of the BME polytope (third family of inequalities in Table 2.1). This result, combined with the previous findings discussed in Forcey et al. [2015], allowed the authors to define a polytope, called the *splitohedron*, that contains the BME polytope and fully characterizes it for $n \leq 11$. Catanzaro et al. [2020b] extended Forcey et al. [2015, 2017]'s results, by unveiling new families of facet-defining inequalities for $n \geq 6$, valid inequalities, and a polynomial-time oracle to recognize the vertices of the BME polytope. As in Forcey et al. [2015], the assistance of Polymake [Gawrilow and Joswig, 2000] was fun-

damental to assess the facet-defining nature of given families of inequalities. This reverse engineering approach currently seems to be the only viable way to analyze the polyhedral combinatorics of the problem, mainly due to a lack of general properties to assess the linear independence of a system of matrices in the Space($\mathcal{X}$).

Recently, Devadoss et al. [2019] extended the polyhedral combinatorics of the BME polytope also to the context of phylogenetic networks. In particular, starting from the results discussed in [Forcey et al., 2017], the authors studied the combinatorics of the *level-1 split networks* [Huson et al., 2011], by showing that their convex hull shares families of facets with the BME polytope.

2.5 Computational aspects of the BMEP: A taxonomy of the literature

In this section we present a possible taxonomy of the the solution approaches to the BMEP currently described in the literature. We will start by describing the exact solution approaches, by distinguishing between contributions based on brute-force enumeration and implicit enu-

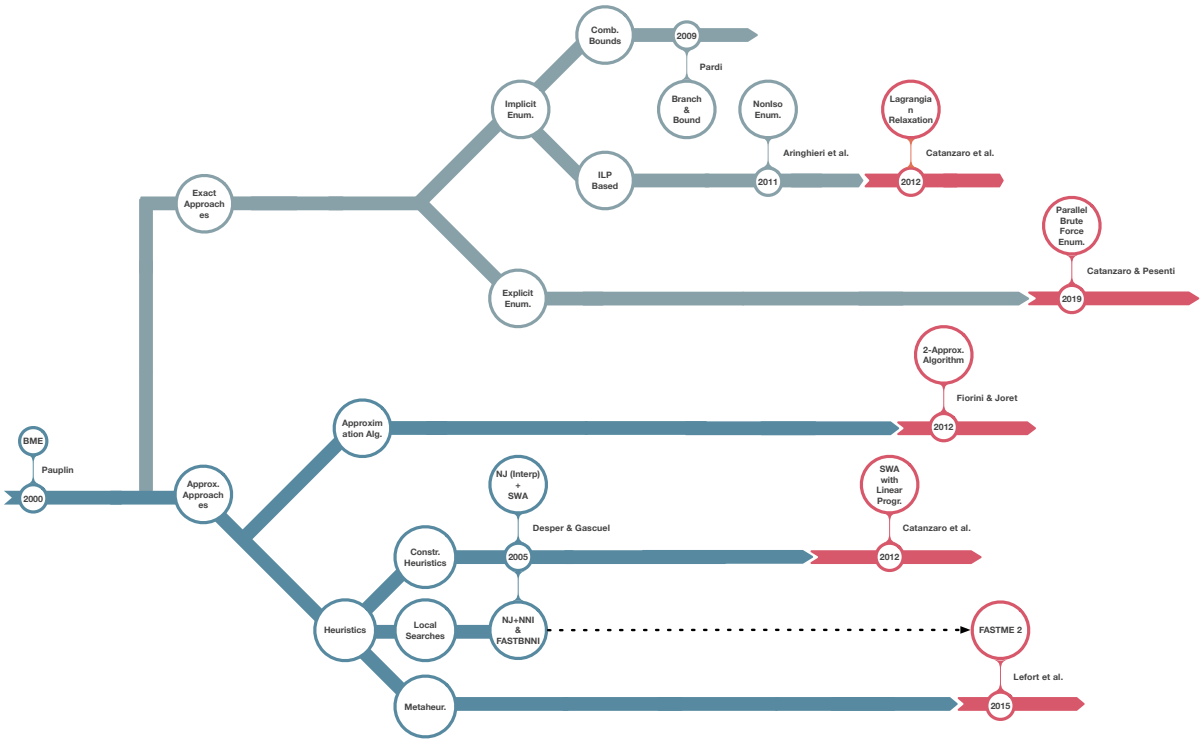


Fig. 2.15. A graphical taxonomy of the solution approaches for the BMEP currently described in the literature. We can distinguish between exact (gray line) and approximate (blue line) solution approaches. The exact approaches can be classified into implicit and explicit enumerations; the implicit enumerations can be further classified in function of the type of lower bound used. The approximate algorithms can be classified into approximation algorithms, heuristics, local searches and metaheuristics. The red lines represent the current state-of-the-art solution approaches.

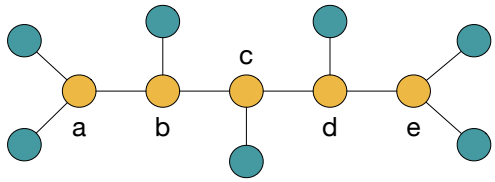


Fig. 2.16. An unlabeled phylogeny with seven leaves.

meration algorithms based either on combinatorial bounds or on mathematical programming. Subsequently, we will discuss the approximate solution approaches, by presenting the approximation algorithms, the heuristics known so far as well as the local searches and the metaheuristics. Figure 2.15 provides a graphical representation of this taxonomy and highlights the current state-of-the-art in the respective areas.

2.5.1 Exact approaches

The earliest exact approach to solution of the BMEP was proposed by Pardi [2009] in 2009. Based on Semple and Steel [2004]'s combinatorial interpretation of the length function, the author derived some combinatorial lower bounds on the value of the optimal solution to the BMEP by determining how much the length of a *partial phylogeny* T' of Γ (i.e., the phylogeny of a subset Γ' of Γ) can increase when *inserting* a taxon $k \in \Gamma \setminus \Gamma'$ and its direct ancestor k' in T' (i.e., when replacing an edge (i, j) of T' by the triplet of edges $\{(i, k'), (k, k'), (k', j)\}$). Subsequently, Pardi [2009] combined these lower bounds with an ingenious implicit enumeration of all of the possible solutions to the problem so as to obtain a branch-&-bound algorithm able to solve instances of the problem containing up to 20 taxa within one hour computing time [Catanzaro et al., 2012].

An alternative exact solution approach for the BMEP was proposed by Aringhieri et al. [2011] in 2011. The authors exploited the concept of *tree-isomorphism* [Kreher and Stinson, 1999] among phylogenies to implicitly explore the solution space of the problem. Specifically, two phylogenies $T_1, T_2 \in \mathcal{T}$ are called *isomorphic* if there exists a graph isomorphism between T_1 and T_2, i.e., a bijection ψ from the vertex set of T_1 to the vertex set of T_2 such that two vertices, say u and v, are adjacent in T_1 if and only if $\psi(u)$ and $\psi(v)$ are adjacent in T_2. Phylogeny isomorphism defines an equivalence relation in $\mathcal{T}$ which is captured by the concept of *unlabeled phylogeny*. For example, the phylogeny in Figure 2.1 belongs to the equivalence class of the unlabeled phylogeny shown in Figure 2.16.

Now, it is worth noting that the phylogenies in $\mathcal{T}$ can be generated by starting from the knowledge of the equivalence classes (i.e., unlabeled phylogenies) in which the set $\mathcal{T}$ is partitioned. In fact, given an unlabeled phylogeny U of Γ and denoting by L the set of its n leaves, any phylogeny $T \in \mathcal{T}$ belonging to the class U can be defined by means of an *assignment* of the taxa in Γ to the leaves in L, i.e., by means of a bijective mapping $f : \Gamma \to L$. Aringhieri et al. [2011] exploited this insight to develop a possible exact solution

approach for the BMEP based on the iteration of the following two operations: (i) choose an unlabeled phylogeny U of Γ, and (ii) assign the taxa in Γ to the leaves of U so as to minimize (2.1). This approach, called *Non-isomorphic Enumerative Approach* (NEA), has two fundamental advantages: first, the number of unlabeled phylogenies for an increasing number n of taxa is a function that, though exponential in n, grows much slower than the corresponding number of phylogenies of Γ (see Table 2.2); second, the (quadratic) assignment subproblem (ii) on each unlabeled phylogeny can be processed independently of the others; hence, it can be parallelized. These characteristics allowed Aringhieri et al. [2011] to solve instances of the BMEP larger than Pardi [2009] within the same time limit.

Catanzaro et al. [2012] introduced in 2012 the first implicit enumeration algorithm for the BMEP based on mathematical programming. The authors inspired to the strong combinatorial connections existing between the BMEP and the *Huffman Coding Problem* (HCP) [Parker and Ram, 1996] to identify some fundamental equalities describing UBTs (namely (2.26) and (2.27)). Then, they combined these equalities with Buneman's conditions (2.28) to derive a (polynomial-sized) *Integer Linear Programming* (ILP) formulation for the BMEP that can be summarized as follows. Denote Γ and V as the set of external and internal vertices of a phylogeny, respectively, and let L denote the set $\{1, 2, 3, \ldots, (n-1)\}$. Moreover, define the following binary decision variables:

$$x_{ij}^k = \begin{cases} 1 & \text{if } \tau_{ij} = k \\ 0 & \text{otherwise} \end{cases} \quad \forall\, i, j \in \Gamma \cup V, i \neq j, \forall\, k \in L$$

and

$$y_{ij}^{qt} = \begin{cases} 1 & \text{if } \tau_{it} + \tau_{jq} \geq \tau_{iq} + \tau_{jt} \\ 0 & \text{otherwise} \end{cases}$$

for all $i, j, q, t \in \Gamma \cup V, i \neq j \neq q \neq t$. Then, Catanzaro et al. [2012]'s ILP formulation of the BMEP reads as follows:

	Phylogenies			Phylogenies	
Number of Taxa	Labeled	Unlabeled	Number of Taxa	Labeled	Unlabeled
3	1	1	14	$3.2 \cdot 10^{11}$	135
4	3	1	15	$7.9 \cdot 10^{12}$	265
5	15	1	16	$2.1 \cdot 10^{14}$	552
6	105	2	17	$6.2 \cdot 10^{15}$	1132
7	945	2	18	$1.9 \cdot 10^{17}$	2410
8	10,395	4	19	$6.3 \cdot 10^{18}$	5098
9	135,135	6	20	$2.2 \cdot 10^{20}$	11 020
10	2,027,025	11	25	$2.5 \cdot 10^{28}$	565 734
11	34,459,425	18	26	$1.19 \cdot 10^{30}$	1 265 579
12	654,729,075	37	...	...	...
13	$1.4 \cdot 10^{10}$	66	n	$(2n-5)!!$	A000672

Table 2.2. Number of phylogenies and unlabeled phylogenies for increasing number of taxa. The general formula for the unlabeled UBTs is quite long and is omitted. The interested reader may however find more information at `https://oeis.org/A000672`, accessible by clicking on the hyperlink A000672 provided in the table.

Formulation 1

$$\min \sum_{i\in\Gamma}\sum_{\substack{j\in\Gamma\\ j\neq i}} d_{ij}\left(\sum_{k\in L\setminus\{1\}} 2^{-k}x_{ij}^k\right)$$

$$s.t. \sum_{k\in L} x_{ij}^k = 1 \qquad \forall\, i\neq j\in\Gamma\cup V \tag{2.34}$$

$$x_{ji}^k = x_{ij}^k \qquad \forall\, i<j\in\Gamma\cup V, k\in L \tag{2.35}$$

$$\sum_{\substack{j\in\Gamma\\ j\neq i}}\sum_{k\in L\setminus\{1\}} 2^{-k}x_{ij}^k = \frac{1}{2} \qquad \forall\, i\in\Gamma \tag{2.36}$$

$$\sum_{\substack{k\in L\\ k\neq 1}} k2^{-k}\sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} x_{ij}^k = (2n-3) \tag{2.37}$$

$$\sum_{k\in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k\in L} k(x_{iq}^k + x_{jt}^k) + (2n-2)y_{ij}^{qt} \qquad \forall\, i\neq j\neq q\neq t\in\Gamma\cup V \tag{2.38}$$

$$\sum_{k\in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k\in L} k(x_{it}^k + x_{jq}^k) + (2n-2)(1-y_{ij}^{qt}) \qquad \forall\, i\neq j\neq q\neq t\in\Gamma\cup V \tag{2.39}$$

$$x_{ij}^1 = 0 \qquad \forall\, i\neq j\in\Gamma \tag{2.40}$$

$$\sum_{j\in V} x_{ij}^1 = 1 \qquad \forall\, i\in\Gamma \tag{2.41}$$

$$\sum_{\substack{j\in\Gamma\cup V\\ j\neq i}} x_{ij}^1 = 3 \qquad \forall\, i\in V \tag{2.42}$$

$$x_{ij}^1 + x_{il}^1 + x_{lj}^1 \leq 2 \qquad \forall\, i\neq j\neq l\in V \tag{2.43}$$

$$x_{ij}^k + 1 \geq x_{il}^{(k-1)} + x_{lj}^1 \qquad \forall\, i\neq j\in\Gamma, l\in V, k\in L\setminus\{1,n-1\} \tag{2.44}$$

$$x_{ij}^k + x_{ij}^{(k-2)} + 1 \geq x_{il}^{(k-1)} + x_{lj}^1 \qquad \forall\, i\neq j\neq l\in\Gamma\cup V, k\in L\setminus\{1,2,n-1\} \tag{2.45}$$

$$x_{ij}^k \in \{0,1\} \qquad \forall\, i,j\in\Gamma\cup V, k\in L \tag{2.46}$$

$$y_{ij}^{qt} \in \{0,1\} \qquad \forall\, i\neq j\neq q\neq t\in\Gamma\cup V. \tag{2.47}$$

The convexity constraints (2.34) ensure that in any feasible solution to the problem the path connecting the pair of taxa i and j is characterized by exactly one length. Constraints (2.35) impose equation (2.24). Constraints (2.36) encode Kraft's equalities (2.26). Constraint (2.37) impose equation (2.27). Constraints (2.38) and (2.39) impose the Buneman's conditions (2.28). The subsequent set of constraints (2.40)-(2.45) impose that the set of edges encoded by variables x_{ij}^1 form an UBT. Finally, the last two constraints impose the integrality of variables x and y. Catanzaro et al. [2012]'s ILP formulation is characterized by very tight linear programming relaxations (generally below 2-3% from the optimum). However, the relative solution times of such formulation prove to be particularly slow, mainly due to the presence of a large number of y variables. Any attempt to project those variables out from the formulation (e.g., by using a Bender's decomposition approach [Martin, 1999]) proved inefficient. Hence, the authors proposed the use of a branch-and-bound approach in which (i) the topology of the phylogeny was explicitly taken into account by means of a step-wise insertion algorithm similar to the one already proposed in [Pardi, 2009] and (ii) the lower bound for a generic node of the search tree was computed by means of a relaxation of the above ILP consisting just of constraints (2.34)-(2.37). This relaxation, in fact, is characterized by values of the linear programming relaxations close to the one provided by the ILP formulation. The

resulting algorithm outperformed Pardi [2009]'s and Aringhieri et al. [2011]'s approaches by proving able to solve instances of the BMEP containing up to 26 taxa within 1h computing time. This performance makes Catanzaro et al. [2012]'s approach the current state-of-the-art exact solution algorithm for the BMEP. However, it is worth noting that many practical instances of the problem include hundreds, or even thousands, of taxa (see, e.g., SARS-CoV-2 genomes available at `https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/`). Because the optimal solution to the BMEP is statistical consistent, developing new exact solution algorithms able to tackle instances of the BMEP much larger than Catanzaro et al. [2012]'s algorithm is highly desirable. In Chapter 5, we present a parallel algorithm that improves upon the benchmarks set by Catanzaro et al. [2012]'s algorithm and solve some previously unsolvable instances of the BMEP. Another promising perspective to solve larger instances would be the use of specific tree-coding schemes mixed to ILP and parallel programming. Specifically, Catanzaro and Pesenti [2019] recently presented a brute-force approach to enumerate all of the vertices of the BME polytope. The authors encoded phylogenies by means of Prüfer codes [Caminiti et al., 2007] and showed techniques to massively parallelize this enumeration both on CPUs and GPUs as well as an optimal algorithm to compute the length of a given Prüfer code encoding a phylogeny $T \in \mathcal{T}$. Investigating ways to compute lower bounds in Catanzaro and Pesenti [2019]'s framework could transform this brute-force algorithm into a massive parallel branch-and-bound algorithm for the BMEP and possibly lead to a tool of practical use in phylogenetic analysis.

2.5.2 Approximate approaches

The absence of exact solution algorithms able to tackle large instances of the BMEP entails the use of approximation algorithms or heuristics to approximate the corresponding optimal solutions. In this section, we will review the approximate solution approaches that are currently described in the literature on the BMEP.

Approximation algorithms

The first – and currently unique – approximation algorithm for the BMEP was proposed by Fiorini and Joret [2012] in 2012. In that article, the authors showed that, despite the BMEP being $\mathcal{APX}$-hard in general, its optimal solution can be 2-approximated in the case $\mathbf{D}$ is a metric matrix. This statement is based on some combinatorial relationships existing between the TSP, the *Minimum Spanning Tree Problem* (MST) [Nemhauser and Wolsey, 1999], and the BMEP. In particular, the authors first noted that the minimum cyclic permutation of Γ (i.e., the optimal solution to the TSP instance encoded by $\mathbf{D}$) is always smaller than or equal to the average of the cyclic permutations of Γ compatible with the optimal tree (i.e., the optimal solution to the BMEP, see Section 2.4.1) [Fiorini and Joret, 2012]. Moreover, the authors also recalled that the optimal solution to the TSP instance encoded by $\mathbf{D}$ is an upper bound on the value of the corresponding MST instance [Garey and Johnson, 2003].

Then, the authors designed an approximation algorithm carrying out the following steps: (i) construct the MST for the given input distance matrix; subsequently, (ii) create an extended semi-metric distance matrix by modifying iteratively the topological structure of the MST so as to obtain a phylogeny T of Γ such that any circular order on T has length exactly twice the optimal length of the MST. Because the error ratio between the optimal solution to the BMEP and the length of T is two, this algorithm 2-approximates the BMEP. The authors conjectured the following fact:

Conjecture 3 *Any approximation algorithm for the BMEP based on a MST scheme cannot have an error ratio smaller than two.*

Proving or disproving the above conjecture still remains an open problem.

Heuristics

There exists two main families of heuristics for the BMEP currently described in the literature on phylogenetics, namely those based on the *Stepwise Addition* (SA) paradigm [Felsenstein, 2004] and those based on the *agglomerative* paradigm [Gascuel, 2005]. The SA paradigm refers to a constructive heuristic that starts from a star-tree $\hat{T}$ connecting any subset $\hat{\Gamma}$ of three taxa in Γ and subsequently iterates the following steps until a complete phylogeny of Γ is obtained: choose a taxon $k \in \Gamma \setminus \hat{\Gamma}$; insert k into an edge e^* of $\hat{T}$ that optimizes a selected criterion; set $\hat{\Gamma} = \hat{\Gamma} \cup \{k\}$ (see Figure 2.17). The heuristics based on the SA paradigm mainly differ from one another by the criterion used to measure the desirability of inserting k on a specific edge of $\hat{T}$. For example, Desper and Gascuel [2002] and Pardi [2009] proposed the use of proxies derived from Semple and Steel [2004]'s combinatorial interpretation of the length function and able to provide an estimation of how much the length of the partial phylogeny $\hat{T}$ including k may increase when inserting k into a specific edge e of $\hat{T}$. Catanzaro et al. [2012], instead, proposed the use of a linear programming relaxation of Formulation 1 constituted by constraints (2.34)-(2.37). The insertion criterion proposed by Catanzaro et al. [2012] is computationally more expensive than those proposed by Desper and Gascuel [2002] and Pardi [2009]; the BME length of the resulting solution, however, may prove shorter than those of Desper and Gascuel [2002] and Pardi [2009].

The heuristics based on the agglomerative paradigm start from an infeasible solution to the problem (i.e., a star-tree with n taxa) and subsequently iterate the following steps until a complete phylogeny of Γ is obtained: select a pair of taxa, say $i, j \in \Gamma$, that optimize a specific selection criterion; *cluster* i and j, i.e., replace i and j by a virtual taxon v that acts as a common ancestor of the selected pair; set $\Gamma = (\Gamma \setminus \{i, j\}) \cup \{v\}$ (see Figure 2.18). The agglomerative approach is well-known in the literature on phylogenetics. It was first introduced by Saitou and Nei [1987] in 1987, under the name of *Neighbor Joining* (NJ) method [Saitou and Nei, 1987, Studier and Keppler, 1988, Gascuel and Steel, 2006], and subsequently declined in a myriad of variants, many of which may not be directly related to

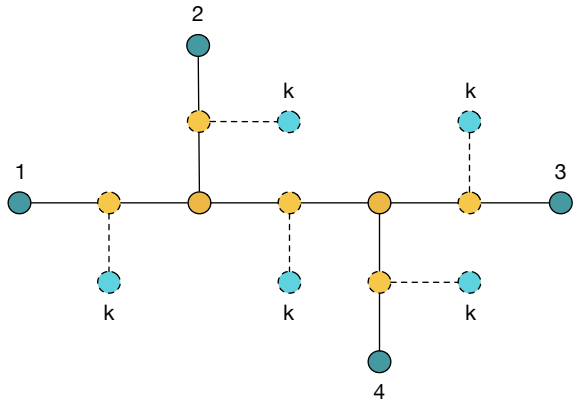


Fig. 2.17. All of the possible insertions of the taxon k into the edges of a partial phylogeny of four taxa. The result of this insertion is a phylogeny of five taxa.

the BMEP [Felsenstein, 2004, Gascuel, 2005]. The validity of the NJ method (see Figure 2.19) as a heuristic for the BMEP was longly investigated, until Desper and Gascuel's breakthrough showed that it is indeed a greedy heuristic for the BMEP (see [Gascuel, 2005]). The reader interested in a historical perspective on this development may refer to Gascuel and Steel [2006]'s article.

The literature also provides a number of improvement heuristics, based on local searches, that try to modify a given initial solution T_0 to the BMEP so as to obtain a new one characterized by a shorter length. These methods usually define a *neighborhood* of T_0, i.e., a set of all of the possible phylogenies of Γ that can be obtained from T_0 by applying e.g., some sort of topological change. Typical topological changes are the *Nearest Neighbor Interchange* (NNI), the *Subtree Pruning and Regrafting* (SPR), and the *Tree Bisection and Reconnection* (TBR) (see [Felsenstein, 2004] and Figure 2.19). The NNI neighborhood is obtained by considering all of the $O(n)$ phylogenies that can be obtained from T_0 by selecting one of its internal edges, say $e = (u, v)$, and by swapping two subtrees connected to u and v [Desper and Gascuel, 2002]. The SPR neighborhood is obtained by considering all of the $O(n^2)$ phylogenies that can be obtained from T_0 by removing a subtree and by reconnecting it to any other edge in the partial phylogeny [Desper and Gascuel, 2002]. The TBR neighborhood is obtained by considering all of the $O(n^3)$ phylogenies that can be obtained from T_0 by removing one of its edges and by reconnecting the two connected components by means of a new edge linking

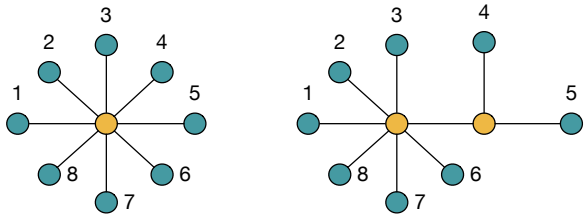


Fig. 2.18. The phylogeny on the right is obtained from the star-tree on the left by "agglomerating" taxa 4 and 5, i.e., by connecting the two taxa to a common ancestor and by joining this ancestor to the center of the remaining star.

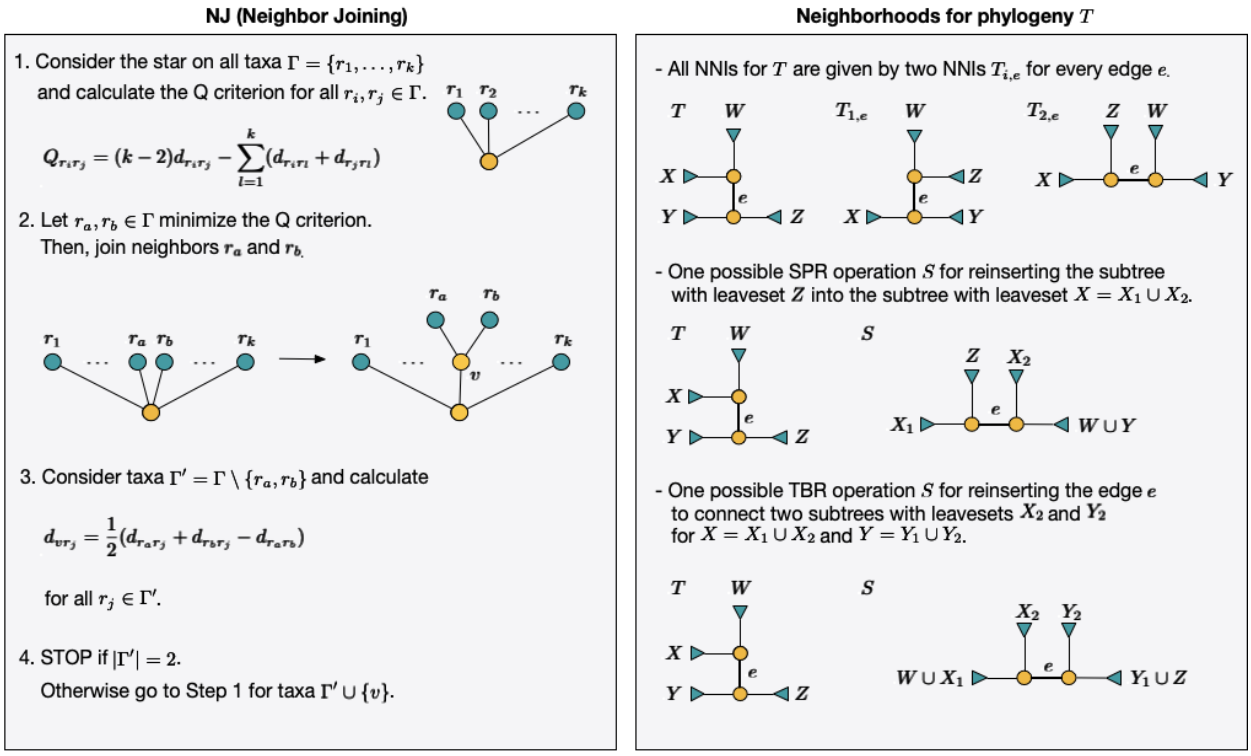


Fig. 2.19. An example of two possible approximate approaches to solution of the BMEP. The figure on the left outlines the so called *Neighbor Joining* (NJ) method [Saitou and Nei, 1987, Studier and Keppler, 1988, Gascuel and Steel, 2006] i.e., a greedy constructive heuristic that starting from an infeasible solution to the problem (i.e., a star-tree) iteratively clusters two terminal leaves according to a specific optimality criterion, until a phylogeny of Γ is obtained. The figure on the right shows three examples of possible neighborhoods of a given phylogeny T. We can distinguish between the *Nearest Neighbor Interchange* (NNI), the *Subtree Pruning and Regrafting* (SPR), and the *Tree Bisection and Reconnection* (TBR), respectively [Felsenstein, 2004, Gascuel, 2005].

some edge in the first subtree with some edge in the second subtree. Many phylogenetic tools and packages implement local searches based on these three families of neighborhoods. A notable tool is FastME [Lefort et al., 2015], which currently constitutes the state-of-the-art for the BMEP. FastME offers several variants of the above starting heuristics as well as all the NNI and SPR neighborhoods. The tool carries out an *Iterated Local Search* [Lourenço et al., 2003] in the solution space of the problem, by fully exploring the considered neighborhoods at each step of the process. The tool then stops as soon as no further improvement on the best-so-far solution is found after a predefined number of iterations.

Atteson's safety radius analysis of the heuristic solutions to the BMEP

The statistical consistency results presented in Desper and Gascuel [2004] concerned just the optimal solution to the BMEP. This issue, however, was left open for the approximate ones. Hence, in the last two decades several authors started to investigate the relationships between suboptimality and statistical consistency, by using Atteson [1999]'s notion of *safety radius* of a solution algorithm A, i.e., a measure of the ability of A to recover the same

phylogeny T when perturbing the input distance matrix of taxa. This measure is formally defined as follows. Let D^{Δ} denote a $n \times n$ symmetric matrix whose generic entry d_{ij}^{Δ} encodes the estimated evolutionary distance between taxa i and j in Γ. Let T denote the phylogeny of Γ computed by means of A when processing the input matrix D^{Δ}. Moreover, let D^{T} denote the *tree metric* estimated by means of A, i.e., a $n \times n$ symmetric matrix whose generic entry d_{ij}^{T} represents the sum of the edge weights belonging to the unique path in T connecting taxon i to taxon j. Finally, let α be a positive scalar and let $w_{\min}$ denote the smallest internal edge weight of T. Then, the *safety radius* of A is the maximum over all of the possible values of α for which if

$$||D^{T} - D^{\Delta}||_{\infty} < \alpha w_{\min} \tag{2.48}$$

holds true, then A still reconstructs T when processing the input matrix D^{Δ}. As shown by Atteson [1999], no exact or approximate algorithm for the BMEP can have a safety radius $\alpha > 1/2$. In other words, the scalar α may vary only within the interval $[0, 1/2]$. Pardi et al. [2010] showed that the safety radius of any exact solution algorithm for the BMEP is precisely $1/2$. Surprisingly enough, this is also the safety radius for the NJ method [Atteson, 1999]. A local search in the SPR neighborhood has a safety radius provably greater than or equal to $1/3$, independently of the starting phylogeny considered [Bordewich et al., 2009a]. Based on empirical evidence, Pardi et al. [2010] conjectured that the safety radius of the local search in the NNI neighborhood should be at least $1/3$, independently of the starting phylogeny considered [Bordewich et al., 2009a]. However, nobody provided a proof of this conjecture so far. Finally, determining the safety radius of the local search in the TBR neighborhood is still an open problem.

In [Gascuel and Steel, 2016] Gascuel and Steel proposed an improvement of Atteson's approach that consists of replacing the infinite norm by the 2-norm. The authors observed, in fact, that in biological terms the infinite norm translates into a sort of worst case approach, while the 2-norm behaves as a sort of average case. The authors also established upper and lower bounds for the resulting safety radius.

2.6 Concluding remarks

The BMEP definitely is one of the most successful estimation models described so far in the literature on phylogenetics. As for the maximum likelihood [Felsenstein, 2004] and Bayesian inference [Huelsenbeck et al., 2001, 2002], it is proven to be statistically consistent and by far the most accurate estimation model from among distance methods [Gascuel, 2005, Gascuel and Steel, 2006]. Moreover, it inherits from distance methods the ability of virtually running on any kind of genetic data for which a measure of dissimilarity exists. This fact enables heuristics for the BMEP (such as NJ or FastME) to tackle very large datasets containing more than 10,000 taxa. The last twenty years saw many advances on numerous aspects related to the statistical consistency, the combinatorics, the computational complexity and the optimization aspects of the BMEP.

In this chapter we have reviewed many of them. Other aspects of the problem, however, still remain open. For example, we know that there exist fundamental equations that must be satisfied by all of the phylogenies of a given set of taxa. However, we do not know if these equations are sufficient to characterize all of them. Similarly, we know some fundamental structural characteristics of the BMEP polytope as well as some of its facets. However, we do not have yet general properties to assess the linear independence of a set of phylogenies. This fact, has a negative impact on the polyhedral analysis of the BMEP, as it entails the use of reverse engineering approaches – based e.g., on Polymake [Gawrilow and Joswig, 2000] – which provide insights only in low dimensions and become quickly intractable. Furthermore, we know about the existence of connections between the BME polytope, Birkhoff's polytope and the permutoassociahedron [Forcey et al., 2015, 2017]. We also know that there exist combinatorial connections between the BMEP and the Quadratic Assignment Problem [Çela, 1997]. However, we do not know how to efficiently exploit this information in the context of implicit enumeration algorithms. Moreover, the exact solution algorithms currently available struggle to solve practical instances of the BMEP. Finally, the last theoretical development on the heuristic approaches for the BMEP leave room to several interesting questions related e.g., to the Attenson's safety radius for the NNI and TBR neighborhoods, and to the identification of the necessary and sufficient conditions that a heuristic for the BMEP must satisfy to ensure the optimality of the solution based on the stochastic variant of the safety radius (see, e.g., [Gascuel and Steel, 2016]). The challenges offered by the BMEP definitely are stimulating theoretical and computational questions per se; but perhaps more importantly, they potentially represent one way in which the Operational Research community can actively contribute in the fight against future pandemics.

3

An Information Theory Perspective on Phylogenetics

The connections between the BMEP and the Huffman Coding Problem (HCP) allowed to unveil a number of fundamental equations at the core of the combinatorics of the BMEP (see Section 2.4.3). These equations are essential to determine the appropriate space in which to study the polyhedral combinatorics of the problem as well as crucial to reduce the gap between the value of the linear programming relaxation of any valid integer linear programming formulation for the problem and the value of the corresponding optimal solution. The connections between the BMEP and the HCP, however, do not limit just to these results. In fact, there exist further shared aspects between the two problems that enable a new original interpretation of the BMEP in terms of information theory. In this chapter, we show that the *Balanced Minimum Evolution Problem* (BMEP) is a cross-entropy minimization problem. This new perspective both extends the previous interpretations of the BMEP length function described in the literature and enables the identification of an efficiently computable family of lower bounds on the value of the optimal solution to the problem. Furthermore, we discuss the possibilities of viewing some phylogenetic estimation problems as cross-entropy minimization problems. In particular, we discuss the minimum evolution criterion (see Sections 2.3.2) and the significance of the information theory framework for the study of statistical inconsistencies under this criterion.

3.1 Background

First, to simplify our notation, fixed a taxon $i \in \Gamma$, we define $\Gamma_i = \Gamma \setminus \{i\}$. And given a phylogeny $T \in \mathcal{T}$ and a taxon $i \in \Gamma$, we define $\sigma_{ij} = \tau_{ij} - 1$, for all $j \in \Gamma_i$, to rewrite the length function of the BMEP as

$$L(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\tau_{ij}}} = \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}}. \tag{3.1}$$

We refer to the vector $\sigma_i = (\sigma_{i1}, \ldots, \sigma_{i(i-1)}, \sigma_{i(i+1)}, \ldots, \sigma_{in})$ as the *rooted path-length sequence of taxon* $i \in \Gamma$ and we observe that its entries encode the lengths of the (unique) paths

in T having origin in the (unique) internal vertex of T adjacent to i and destinations in the remaining $n-1$ taxa in Γ_i. For each taxon $i \in \Gamma$ we denote Θ_i as the set of the path-length sequences σ_i induced by the phylogenies in $\mathcal{T}$. Moreover, recall that Θ denotes the set of path-length matrices encoding phylogenies $T \in \mathcal{T}$, i.e., $\tau \in \Theta$ is composed of entries $\tau_{ij} = \sigma_{ij} + 1$ for all $i, j \in \Gamma$ and suitable $\sigma_i \in \Theta_i$. Therefore, for a phylogeny T, we associate with the corresponding path-length matrix τ a *collection* $\sigma = \{\sigma_i : i \in \Gamma\}$ of path-length sequences associated to T. Then, the problem of determining the phylogeny $T \in \mathcal{T}$ that minimizes (2.1) is equivalent to the problem of determining the corresponding collection σ that minimizes the right-hand side of (3.1).

We recall that the sets Θ_i and Θ are characterized by specific equalities (see Equations (2.26) and (2.27)). In particular, *Kraft's equality* states that for a fixed taxon $i \in \Gamma$, $\sigma_i \in \Theta_i$ if and only if

$$\sum_{j \in \Gamma_i} 2^{-\sigma_{ij}} = 1. \tag{3.2}$$

The second equality constrains each collection σ associated to a path-length matrix $\tau \in \Theta$ to satisfy the following equation:

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \sigma_{ij} 2^{-\sigma_{ij}} = 3n - 6. \tag{3.3}$$

We refer the reader interested in deepening the combinatorial meaning of these equalities to [Catanzaro et al., 2012] and [Catanzaro et al., 2020b].

If not stated otherwise, we shall assume that any generic matrix mentioned in this chapter belongs to $\mathbb{R}^{n \times n}$, with $n = |\Gamma|$. We denote $\mathbf{I}$ as the identity matrix. Moreover, given a square matrix $\mathbf{M}$, we denote: $2^{-\mathbf{M}}$ as the square matrix obtained from $\mathbf{M}$ by setting its generic entry m_{ij} to $2^{-m_{ij}}$, for all $i, j \in \Gamma$; and $\log_2(\mathbf{M})$ as the square matrix obtained from $\mathbf{M}$ (in this case assumed positive) by setting its generic entry to $\log_2(m_{ij})$, for all $i, j \in \Gamma$. Given two matrices $\mathbf{M}$ and $\tilde{\mathbf{M}}$, we say that they have the same *zero pattern* if

$$m_{ij} = 0 \iff \tilde{m}_{ij} = 0 \quad \forall\, i, j \in \Gamma. \tag{3.4}$$

We conclude this section by recalling some definitions from information theory that will be frequently used hereafter. Specifically, given two probability distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$, we define the *information entropy* associated to p, [Sayood, 2017] as

$$\mathcal{H}(p) = -\sum_{j=1}^{n} p_j \log_2(p_j)$$

and the *cross-entropy* associated to p and q [Böcherer and Amjad, 2014, Hirschler and Woess, 2018] as

$$\mathcal{H}(p,q) = -\sum_{j=1}^{n} p_j \log_2(q_j).$$

Finally, we recall that a common way to measure the distance between a pair of probability distributions p and q is given by the *Kullback-Leibler (KL) Divergence* [Sayood, 2017], which is defined as

$$D_{KL}(p\|q) = \mathcal{H}(p,q) - \mathcal{H}(p). \tag{3.5}$$

3.2 The BMEP as a cross-entropy minimization problem

In this section we show that the BMEP is a cross-entropy minimization problem. Because this particular interpretation is enabled by an appropriated scaling of the input distance matrix $\mathbf{D}$, we will start by recalling first some definitions and results from matrix scaling theory. Specifically, we say that a square matrix having nonnegative entries is *doubly-stochastic* if each of its rows and columns sums to one. We also recall that if a square matrix $\mathbf{M}$ has nonnegative entries then there exist — due to Sinkhorn's theorem [Sinkhorn, 1964] — two diagonal matrices, say $\mathit{\Pi}_1$ and $\mathit{\Pi}_2$, such that $\mathit{\Pi}_1\mathbf{M}\mathit{\Pi}_2$ is doubly-stochastic. When $\mathbf{M}$ is also symmetric then $\mathit{\Pi}_1$ and $\mathit{\Pi}_2$ are unique, equal, characterized by positive entries, and computable in polynomial-time [Khachiyan, 1996, Brualdi, 1974, Idel, 2016]. Then, in the light of these definitions and results, the following proposition holds:

Proposition 3.1. *Let* $\mathbf{D}$ *be the input distance matrix of the BMEP. Then, there exists a unique scaling matrix* $\mathit{\Pi} = diag(\pi_1, \ldots, \pi_n)$*, with* $\pi_i > 0$ *for all* $i \in \mathit{\Gamma}$*, such that the matrix* $\mathbf{S} = \mathit{\Pi}(2^{-\mathbf{D}} - \mathbf{I})\mathit{\Pi}$ *is doubly-stochastic and symmetric. Moreover,* $\mathbf{S}$ *has the same* zero pattern *as* $2^{-\mathbf{D}} - \mathbf{I}$*.*

Proof. The matrix $2^{-\mathbf{D}} - \mathbf{I}$ is symmetric has zero-diagonal entries and strictly positive off-diagonal entries. Then, by Brualdi's theorems (see [Brualdi, 1974] and [Idel, 2016, Theorem 5.4]) there exists an unique scaling matrix $\mathit{\Pi}$ having positive diagonal entries and such that $\mathbf{S} = \mathit{\Pi}(2^{-\mathbf{D}} - \mathbf{I})\mathit{\Pi}$ is symmetric, nonnegative and such that the sum of the entries of each row is equal to 1, i.e., such that $\mathbf{S}$ is symmetric and doubly-stochastic. Moreover, Theorem 2 of [Brualdi, 1974] guarantees that $\mathbf{S}$ has the same zero pattern as $2^{-\mathbf{D}} - \mathbf{I}$, so the statement follows.

□

Now, given the matrix $\mathbf{S} = \mathit{\Pi}(2^{-\mathbf{D}} - \mathbf{I})\mathit{\Pi}$, consider the matrix $\mathbf{S} + \mathbf{I}$, having as generic entry

$$s_{ij} = (\mathbf{S} + \mathbf{I})_{ij} = \begin{cases} 1 & \text{if } i = j \\ \pi_i \pi_j 2^{-d_{ij}} & \text{if } i \neq j \end{cases} \quad \forall\ i, j \in \mathit{\Gamma}. \tag{3.6}$$

Moreover, define $\hat{\mathbf{S}} = -\log_2(\mathbf{S} + \mathbf{I})$ as the matrix having as generic entry

$$\hat{s}_{ij} = \hat{\mathbf{S}}_{ij} = \begin{cases} 0 & \text{if } i = j \\ -\log_2(\pi_i) - \log_2(\pi_j) + d_{ij} & \text{if } i \neq j \end{cases} \quad \forall\, i, j \in \Gamma. \tag{3.7}$$

Finally, define the constant

$$K = \sum_{i \in \Gamma} \log_2(\pi_i). \tag{3.8}$$

Then, the following proposition holds:

Proposition 3.2. *The optimal solution to the BMEP does not change if one replaces the input distance matrix* $\mathbf{D}$ *with its scaled version* $\hat{\mathbf{S}}$*, i.e.,*

$$\arg\min_{\tau \in \Theta} \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} = \arg\min_{\tau \in \Theta} \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\hat{s}_{ij}}{2^{\sigma_{ij}}}.$$

Proof. For any fixed collection σ associated to $\tau \in \Theta$ we have that

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\hat{s}_{ij}}{2^{\sigma_{ij}}} = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} - \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\log_2(\pi_i)}{2^{\sigma_{ij}}} - \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\log_2(\pi_j)}{2^{\sigma_{ij}}},$$

which in turn can be rewritten, in the light of (4.1) and (3.8), as

$$\begin{aligned} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\hat{s}_{ij}}{2^{\sigma_{ij}}} &= \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} - \sum_{i \in \Gamma} \log_2(\pi_i) - \sum_{j \in \Gamma} \log_2(\pi_j) \\ &= \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} - 2K. \end{aligned} \tag{3.9}$$

Because K is independent of the choice of the collection σ, the statement follows.

□

Equation (3.9) paves the way for a precise information theory-based interpretation of the BMEP. Specifically, first note that in the light of equation (3.9) we can rewrite (2.1) as follows:

$$L(T) = \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} = \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{\hat{s}_{ij}}{2^{\sigma_{ij}}} + K. \tag{3.10}$$

Now, for each collection σ, define $p_{ij}^{\sigma} = 2^{-\sigma_{ij}}$, for all distinct $i, j \in \Gamma$, and observe that they are strictly positive. Moreover, by Kraft's equality we also have that

$$\sum_{j\in\Gamma_i} 2^{-\sigma_{ij}} = \sum_{j\in\Gamma_i} p_{ij}^{\sigma} = 1,$$

for all $i \in \Gamma$. Hence, each vector $p_i^{\sigma} = \{p_{ij}^{\sigma} : j \in \Gamma_i\}$, $i \in \Gamma$, can be seen as a probability distribution over the underlying set Γ_i. Similarly, for all distinct $i, j \in \Gamma$, define $q_{ij} = 2^{-\hat{s}_{ij}}$, and observe that, by construction, these values are strictly positive and such that

$$\sum_{j\in\Gamma_i} q_{ij} = 1.$$

Hence, also each vector $q_i = \{q_{ij} : j \in \Gamma_i\}$, $i \in \Gamma$, can be seen as a probability distribution over the underlying set Γ_i. Then, given a scalar $\alpha \in [0, 1]$, we can rewrite (3.10) as follows:

$$\begin{aligned} L(T) &= \frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{d_{ij}}{2^{\sigma_{ij}}} && (3.11)\\ &= \frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{\hat{s}_{ij}}{2^{\sigma_{ij}}} + K \\ &= \frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{\hat{s}_{ij} - \alpha\sigma_{ij}}{2^{\sigma_{ij}}} + \alpha\frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{\sigma_{ij}}{2^{\sigma_{ij}}} + K \\ &= \frac{1}{2}\sum_{i\in\Gamma}\left[-\sum_{j\in\Gamma_i}\frac{\log_2(2^{-\hat{s}_{ij}}) - \alpha\log_2(2^{-\sigma_{ij}})}{2^{\sigma_{ij}}}\right] + \alpha\frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{\sigma_{ij}}{2^{\sigma_{ij}}} + K. && (3.12)\end{aligned}$$

Then, in the light of equation (3.3) and by exploiting the definitions of p_{ij}^{σ} and q_{ij}, we can rewrite (3.12) as

$$L(T) = \frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\left(-p_{ij}^{\sigma}\log_2(q_{ij}) + \alpha p_{ij}^{\sigma}\log_2(p_{ij}^{\sigma})\right) + \alpha\frac{3n-6}{2} + K. \tag{3.13}$$

Now, observe that for each taxon $i \in \Gamma$

$$-\sum_{j\in\Gamma_i} p_{ij}^{\sigma}\log_2(p_{ij}^{\sigma})$$

represents the entropy $\mathcal{H}(p_i^{\sigma})$ associated to the probability distributions p_i^{σ}. Similarly,

$$-\sum_{j\in\Gamma_i} p_{ij}^{\sigma}\log_2(q_{ij})$$

represents the cross-entropy $\mathcal{H}(p_i^{\sigma}, q_i)$ associated to the probability distributions p_i^{σ} and q_i over the underlying set Γ_i. Then, (3.13) can be rewritten as

$$L(T) = \frac{1}{2}\sum_{i\in\Gamma}\left(\mathcal{H}(p_i^{\sigma}, q_i) - \alpha\mathcal{H}(p_i^{\sigma})\right) + \alpha\frac{3n-6}{2} + K. \tag{3.14}$$

In particular, observe that if $\alpha = 0$ in (3.14), then solving the BMEP is equivalent to solving the following cross-entropy minimization problem:

$$L(T) = \frac{1}{2}\sum_{i\in\Gamma}\mathcal{H}(p_i^\sigma, q_i) + K. \tag{3.15}$$

Alternatively, if $\alpha = 1$ in (3.14), then solving the BMEP is equivalent to solving the following KL Divergence minimization problem:

$$\begin{aligned} L(T) &= \frac{1}{2}\sum_{i\in\Gamma}(\mathcal{H}(p_i^\sigma, q_i) - \mathcal{H}(p_i^\sigma)) + \frac{3n-6}{2} + K \\ &= \frac{1}{2}\sum_{i\in\Gamma}(D_{KL}(p_i^\sigma \| q_i)) + \frac{3n-6}{2} + K. \end{aligned} \tag{3.16}$$

3.3 On a family of lower bounds for the BMEP

In this section, we show that for any fixed scalar $\alpha \in [0,1]$

$$LB(\alpha) = \frac{1}{2}\sum_{i\in\Gamma}\min_{\sigma_i\in\Theta_i}\sum_{j\in\Gamma_i}\frac{\hat{s}_{ij} - \alpha\sigma_{ij}}{2^{\sigma_{ij}}} + \alpha\frac{3n-6}{2} + K \tag{3.17}$$

provides a lower bound on the value of the optimal solution to the BMEP. Moreover, we also show that, for any fixed scalar $\alpha \in [0,1]$, this lower bound can be efficiently computed.

We start by observing that the lower bound nature of (3.17) can be trivially derived by the following chain of equalities and inequalities:

$$\begin{aligned} \min_{T\in\mathcal{T}} L(T) &= \min_{\tau\in\Theta}\frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{\hat{s}_{ij} - \alpha\sigma_{ij}}{2^{\sigma_{ij}}} + \alpha\frac{3n-6}{2} + K \\ &\geq \frac{1}{2}\sum_{i\in\Gamma}\min_{\sigma_i\in\Theta_i}\sum_{j\in\Gamma_i}\frac{\hat{s}_{ij} - \alpha\sigma_{ij}}{2^{\sigma_{ij}}} + \alpha\frac{3n-6}{2} + K = LB(\alpha). \end{aligned}$$

Deciding the complexity of computing (3.17) for a fixed scalar $\alpha \in [0,1]$ is however a less trivial task. In order to address it, we introduce some notation and definitions that will prove useful throughout the remainder of the chapter. Specifically, given a vector $x \in \mathbb{R}^n_{0^+}$, we define $sort(x)$ as the vector obtained from x by sorting its entries in non-decreasing order. For example, if $x = (8,6,7,9)$, then $sort(x) = (6,7,8,9)$. We also denote $\pi(x)$ as a permutation of x. Given a n-dimensional vector $d_i \in \mathbb{R}^n$ and a scalar $\alpha \in [0,1]$, we define d_i^α as the $(n-1)$-dimensional vector

$$d_i^\alpha = (d_{i1}, \ldots, d_{i(n-2)}, (d_{i(n-1)} + d_{in})/2 - \alpha) \in \mathbb{R}^{n-1}.$$

To indicate that vector $\sigma_i \in \Theta_i$ has dimension m we write $\Theta_i^{[m]}$. Given a path-length sequence $\sigma_i \in \Theta_i^{[n-1]}$, we define

$$\begin{aligned}\sigma^+(\sigma_i) &= (\sigma_{i1}, \ldots, \sigma_{i(n-1)}, \sigma_{in-1} + 1, \sigma_{in-1} + 1),\\ \Theta_i^{s[n]} &= \left\{ sort(\sigma_i) : \sigma_i \in \Theta_i^{[n]} \right\},\\ \Theta_i^{+[n]} &= \left\{ \sigma^+(\sigma_i) : \sigma_i \in \Theta_i^{[n-1]} \right\}.\end{aligned}$$

Finally, given a n-dimensional non-negative vector

$$d_i = (d_{i1}, \ldots, d_{i(i-1)}, 0, d_{i(i+1)}, \ldots, d_{i(n-2)}, d_{i(n-1)}, d_{in}) \in \mathbb{R}^n$$

we consider the function $f_n^\alpha : \Theta_i^{[n]} \times \mathbb{R}^n \to \mathbb{R}$, defined as

$$f_n^\alpha(\sigma_i, d_i) = \sum_{j=1, j\neq i}^{n} (d_{ij} - \alpha\sigma_{ij})/2^{\sigma_{ij}},$$

and we define the following auxiliary problem:

$$\min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, d_i). \tag{3.18}$$

Then, we observe that deciding the complexity of computing (3.17) for a fixed scalar $\alpha \in [0, 1]$ is equivalent to deciding the complexity of the auxiliary problem (3.18), for any $n = |\Gamma|$ and for a given row d_i of matrix (3.7), $i \in \Gamma$. Moreover, in the light of the above notation and definitions, we observe that the following proposition holds:

Proposition 3.3.

1. *For any permutation π and path-length sequence $\sigma_i \in \Theta_i^{[n]}$, it holds that $\pi(\sigma_i) \in \Theta_i^{[n]}$ and $sort(\sigma_i) \in \Theta_i^{[n]}$.*

2. *For any permutation π, path-length sequence $\sigma_i \in \Theta_i^{[n]}$ and vector $d_i \in \mathbb{R}^n$, it holds that $f_n^\alpha(\sigma_i, d_i) = f_n^\alpha(\pi(\sigma_i), \pi(d_i))$.*

3. *For any path-length sequence $\sigma_i \in \Theta_i^{[n-1]}$, it holds that $\sigma^+(\sigma_i) \in \Theta_i^{[n]}$.*

4. $\Theta_i^{[n]} \supset \Theta_i^{+[n]} \supset \Theta_i^{s[n]}$.

Proof.

1. Recall that an integer vector ξ_i belongs to $\Theta_i^{[n]}$ if and only if it satisfies the Kraft's equality (4.1). Because σ_i satisfies (4.1) then, by the commutative property of summation, any of its permutations, hence also $sort(\sigma_i)$, satisfies (4.1).

2. This property trivially follows by observing that $f_n^\alpha(\pi(\sigma_i), \pi(d_i))$ can be obtained by commutating the terms of the summation that defines $f_n^\alpha(\sigma_i, d_i)$.

3. This property follows again from Kraft's equality (4.1). Specifically, as σ_i satisfies (4.1) then also $\sigma^+(\sigma_i)$ satisfies (4.1). The first $n-3$ terms of the two summations are identical; the last term of (4.1), when computed for σ_i, is equal to $2^{-\sigma_{in-1}}$, which in turn is equal to the sum of the last two terms of (4.1), i.e., $2^{-(\sigma_{in-1}+1)} + 2^{-(\sigma_{in-1}+1)}$, when computed for $\sigma^+(\sigma_i)$.

4. The first set inclusion $\Theta_i^{[n]} \supset \Theta_i^{+[n]}$, is implied by Property 3. The second set inclusion $\Theta_i^{+[n]} \supset \Theta_i^{s[n]}$ derives from Theorem 3.1 in Parker and Ram [1996]. Specifically, this theorem guarantees that, given $\sigma_i \in \Theta_i^{[n]}$, the last two entries of $\tilde{\sigma}_i = sort(\sigma_i)$ are equal, i.e., $\tilde{\sigma}_{in-1} = \tilde{\sigma}_{in}$. Hence, we can derive from $\tilde{\sigma}_i$ the integer vector $\xi_i = (\tilde{\sigma}_{i1}, \ldots, \tilde{\sigma}_{in-2}, \tilde{\sigma}_{in-1} - 1) \in \Theta_i^{[n-1]}$. Indeed, ξ_i satisfies Kraft's equality (4.1) as $sort(\sigma_i)$ satisfies (4.1) and $2^{-(\tilde{\sigma}_{in-1}-1)} = 2^{-\tilde{\sigma}_{in-1}} + 2^{-\tilde{\sigma}_{in}}$. Then, the inclusion $\Theta_i^{+[n]} \supset \Theta_i^{s[n]}$ follows as $sort(\sigma_i) = \sigma^+(\xi_i)$.

□

Proposition 3.4. *The auxiliary problem* (3.18) *can be solved in* $O(n\log(n))$ *time.*

Proof. We prove the proposition by recursion. Specifically, for $n = 4$, the proposition holds as the set $\Theta_i^{[4]}$ reduces to $\{(1,2,2)\}$, hence $\sigma_i = (1,2,2)$ trivially is the optimal solution to the auxiliary problem (3.18).

Now, fixed an integer $n > 4$, denote $\tilde{\sigma}_i = sort(\sigma_i)$, $\tilde{d}_i = sort(d_i)$, and observe that, for any given path-length sequence $\sigma_i \in \Theta_i^{[n]}$, the following relationship holds:

$$\begin{aligned} f_n^\alpha(\sigma_i, d_i) &= \sum_{j=1, j\neq i}^{n} \frac{d_{ij}}{2^{\sigma_{ij}}} - \alpha \sum_{j=1, j\neq i}^{n} \frac{\sigma_{ij}}{2^{\sigma_{ij}}} \\ &\geq \sum_{j=1, j\neq i}^{n} \frac{\tilde{d}_{ij}}{2^{\tilde{\sigma}_{ij}}} - \alpha \sum_{j=1, j\neq i}^{n} \frac{\tilde{\sigma}_{ij}}{2^{\tilde{\sigma}_{ij}}} \\ &= f_n^\alpha(sort(\sigma_i), sort(d_i)). \end{aligned} \tag{3.19}$$

Indeed, we have that

$$\sum_{j=1, j\neq i}^{n} \sigma_{ij} 2^{-\sigma_{ij}} = \sum_{j=1, j\neq i}^{n} \tilde{\sigma}_{ij} 2^{-\tilde{\sigma}_{ij}};$$

moreover, we also have that

$$\sum_{j=1,j\neq i}^{n} d_{ij}2^{-\sigma_{ij}} \geq \sum_{j=1,j\neq i}^{n} \tilde{d}_{ij}2^{-\tilde{\sigma}_{ij}}$$

as $d_{ip} < d_{iq}$ and $\sigma_{ip} \geq \sigma_{iq}$, for $1 \leq p < q \leq n$, imply $d_{ip}2^{-\sigma_{ip}} + d_{iq}2^{-\sigma_{iq}} > d_{ip}2^{-\sigma_{iq}} + d_{iq}2^{-\sigma_{ip}}$. We also note that, for any given vectors $d_i \in \mathbb{R}^n$ and $\sigma_i \in \Theta_i^{[n-1]}$, it holds that

$$\begin{aligned} f_n^\alpha(\sigma^+(\sigma_i), d_i) - f_{n-1}^\alpha(\sigma_i, d_i^\alpha) &= \frac{d_{i(n-1)} - \alpha(\sigma_{i(n-1)} + 1)}{2^{\sigma_{i(n-1)}+1}} \\ + \frac{d_{in} - \alpha(\sigma_{i(n-1)} + 1)}{2^{\sigma_{i(n-1)}+1}} &- \frac{\left(\frac{d_{n-1}+d_n}{2} - \alpha\right) - \alpha\sigma_{i(n-1)}}{2^{\sigma_{i(n-1)}}} = 0. \end{aligned} \tag{3.20}$$

Next, we observe that, for a given vector d_i, the following chain of inequalities holds

$$\begin{aligned} \min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, d_i) &= \min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, sort(d_i)) \\ &\leq \min_{\sigma_i^+ \in \Theta_i^{+[n]}} f_n^\alpha(\sigma_i^+, sort(d_i)) \\ &\leq \min_{\tilde{\sigma}_i \in \Theta_i^{s[n]}} f_n^\alpha(\tilde{\sigma}_i, sort(d_i)) \\ &\leq \min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, d_i) \end{aligned} \tag{3.21}$$

where the first equality is implied by Properties 1 and 2 in Proposition 3.3 and the next two inequalities are implied by Property 4 of the same proposition. Finally, the last inequality is implied by (3.19). In turn, inequalities (3.21) and (3.20) imply that

$$\begin{aligned} \min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, d_i) &= \min_{\sigma_i^+ \in \Theta_i^{+[n]}} f_n^\alpha(\sigma_i^+, sort(d_i)) \\ &= \min_{\sigma_i \in \Theta_i^{[n-1]}} f_{n-1}^\alpha(\sigma_i, sort(d_i)^\alpha). \end{aligned} \tag{3.22}$$

Then, solving

$$\min_{\sigma_i \in \Theta_i^{[n]}} f_n^\alpha(\sigma_i, d_i)$$

reduces to solving

$$\min_{\sigma_i \in \Theta_i^{[4]}} f_4^\alpha(\sigma_i, d_i(4))$$

by recursively applying (3.22) and the following definition

$$\begin{aligned} &d_i(k-1) = sort(d_i(k))^\alpha \quad \text{for } k = n, \ldots, 5 \\ &d_i(n) = d_i. \end{aligned} \tag{3.23}$$

The value of

$$\min_{\sigma_i \in \Theta_i^{[4]}} f_4^{\alpha}(\sigma_i, d_i(4))$$

can be determined in $O(1)$ as the set $\Theta_i^{[4]}$ consists of the single element $(1, 2, 2)$. The vector $d_i(4)$ can instead be computed in $O(nlog(n))$ time. Specifically, for $k = n$, computing $d_i(n-1) = sort(d_i(n))^{\alpha}$ requires $O(nlog(n))$ time as we have to sort d_i in $O(nlog(n))$ time and determine $d_{in-1}(n-1)$ in $O(1)$ time. For $k < n$, computing $d_i(k-1) = sort(d_i(k))^{\alpha}$ requires $O(log(n))$ time because we can sort $d_i(k)$ in $O(log(n))$ time as only its last entry $d_{ik}(k)$ is possibly not already ordered. Thus, the statement follows.

□

In the next example we show how the recursion (3.23) works.

Example 3.5. Consider an instance of the auxiliary problem (3.18) associated to the eight-dimensional vector $d_i = (0, 1, 3, 3, 4, 4, 4, 4)$, $i = 1$, and the scalar $\alpha = 1$. By Proposition 3.4, the optimal solution to

$$\min_{\sigma_i \in \Theta_i^{[8]}} f_8^{\alpha}(\sigma_i, d_i)$$

can be computed by applying the recursion (3.23) to $d_i(8) = d_i$ until $k = 3$. This process gives rise to the following sequence of vectors:

$$\begin{aligned}
sort(d_i(8)) &= (0, 1, 3, 3, 4, 4, 4, 4) && \Rightarrow d_i(7) = (0, 1, 3, 3, 4, 4, 3) \\
\Rightarrow sort(d_i(7)) &= (0, 1, 3, 3, 3, 4, 4) && \Rightarrow d_i(6) = (0, 1, 3, 3, 3, 3) \\
\Rightarrow sort(d_i(6)) &= (0, 1, 3, 3, 3, 3) && \Rightarrow d_i(5) = (0, 1, 3, 3, 2) \\
\Rightarrow sort(d_i(5)) &= (0, 1, 2, 3, 3) && \Rightarrow d_i(4) = (0, 1, 2, 2).
\end{aligned}$$

As $(1, 2, 2) = \arg\min \{f_4^{\alpha}(\sigma_i, d_i(4)) : \sigma_i \in \Theta_i^4\}$, the base case of the recursion returns zero. Now, let π_k be the permutation used to map $sort(d_i(k))$ to $d_i(k)$. Then, we know from equations (3.22) that the corresponding optimal solution $\hat{\sigma}(k)$ is given by $\pi_k(\sigma^+(\hat{\sigma}(k-1)))$. Thus, the recursion returns the following sequences:

$$\begin{aligned}
\hat{\sigma}(4) &= (1, 2, 2) && \Rightarrow \sigma^+(\hat{\sigma}(4)) = (1, 2, 3, 3) \\
\Rightarrow \hat{\sigma}(5) &= (1, 3, 3, 2) && \Rightarrow \sigma^+(\hat{\sigma}(5)) = (1, 3, 3, 3, 3) \\
\Rightarrow \hat{\sigma}(6) &= (1, 3, 3, 3, 3) && \Rightarrow \sigma^+(\hat{\sigma}(6)) = (1, 3, 3, 3, 4, 4) \\
\Rightarrow \hat{\sigma}(7) &= (1, 3, 3, 4, 4, 3) && \Rightarrow \sigma^+(\hat{\sigma}(7)) = (1, 3, 3, 4, 4, 4, 4).
\end{aligned}$$

3.4 On the tightness of the $LB(\alpha)$ bounds

In this section we study the tightness of the lower bounds $LB(\alpha)$. Before proceeding, we recall some definitions and results from the literature that will be frequently used hereafter.

Specifically, we say that $\mathbf{D}$ is *uniform* if all of its non-diagonal entries are equal. We also recall that if the matrix $\mathbf{D}$ is *additive* then there exists a phylogeny $\hat{T} \in \mathcal{T}$ that *fits* $\mathbf{D}$ [Felsenstein, 2004], or equivalently, there exists a collection $\hat{\sigma}$ associated to a path-length matrix $\tau \in \Theta$ such that $d_{ij} = \hat{\sigma}_{ij}$, for all distinct $i, j \in \Gamma$. Finally, with a little abuse of notation, we say that $\hat{\mathbf{S}}$ is uniform or additive whenever the corresponding matrix $\mathbf{D}$ is uniform or additive, respectively, and we write $\hat{\mathbf{S}}^U$ or $\hat{\mathbf{S}}^A$ to distinguish the two respective cases. Then, the following proposition holds:

Proposition 3.6. *Let* $\mathbf{D}$ *be the input distance matrix of the BMEP. If* $\mathbf{D}$ *is uniform, then the lower bound* $LB(0)$ *is tight. Alternatively, if* $\mathbf{D}$ *is additive, then the lower bound* $LB(1)$ *is tight.*

Proof. If the input distance matrix $\mathbf{D}$ is uniform, then $d_{ij} = d$ for all distinct $i, j \in \Gamma$. This fact implies: (i) that the non-diagonal entries of the matrix $\hat{\mathbf{S}}^U$ are equal to $\log_2(n-1)$; (ii) that

$$K = \frac{n(d - \log_2(n-1))}{2};$$

and (iii) that

$$\min_{T \in \mathcal{T}} \; L(T) = \min_{\tau \in \Theta} \frac{1}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\sigma_{ij}}} = \frac{d}{2} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{1}{2^{\sigma_{ij}}} = \frac{nd}{2}.$$

Now, observe that (3.17) attains the value $nd/2$ for $\alpha = 0$. Specifically, in this situation we get

$$LB(0) = \sum_{i \in \Gamma} \min_{\sigma_i \in \Theta_i} \frac{1}{2} \sum_{j \in \Gamma_i} \frac{\log_2(n-1)}{2^{\sigma_{ij}}} + K = \frac{nd}{2},$$

thus the statement follows. Now suppose that the input distance matrix $\mathbf{D}$ is additive and let $\hat{\sigma} = \{\hat{\sigma}_i, i \in \Gamma\}$ be the collection induced by the phylogeny $\hat{T} \in \mathcal{T}$ that fits $\mathbf{D}$. Because the additivity hypothesis is equivalent to saying that $\hat{\mathbf{s}}^A_{ij} = \hat{\sigma}_{ij}$, for all distinct $i, j \in \Gamma$, we have that

$$\begin{aligned} \min_{\tau \in \Theta} \sum_{i \in \Gamma} D_{KL}(p_i^\sigma \| q_i) &= \sum_{i \in \Gamma} \min_{\sigma_i \in \Theta_i} D_{KL}(p_i^\sigma || q_i) \\ &= \sum_{i \in \Gamma} D_{KL}(p_i^{\hat{\sigma}} \| q_i) = 0. \end{aligned}$$

Then, the statement trivially follows once again by observing that

	$\hat{\mathbf{S}}^U$		$\hat{\mathbf{S}}^A$	
n	$LB(0)/n$	$LB(1)/n$	$LB(0)/n$	$LB(1)/n$
10	1.5850	1.2225	1.0918	1.2000
11	1.6610	1.2632	1.1158	1.2273
12	1.7297	1.2922	1.1250	1.2500
13	1.7925	1.3117	1.1440	1.2692
14	1.8502	1.3234	1.1374	1.2857
15	1.9037	1.3287	1.1607	1.3000
25	2.2925	1.4225	1.1356	1.3800
50	2.8074	1.4817	1.1576	1.4400
100	3.3147	1.5112	1.1342	1.4700

Table 3.1. A comparison of the (normalized values of the) lower bounds $LB(0)$ and $LB(1)$ for random uniform or additive matrices $\hat{\mathbf{S}}$ having increasing values of n.

$$\begin{aligned}\min_{\tau\in\Theta}\frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i}\frac{d_{ij}}{2^{\sigma_{ij}}} &= \min_{\tau\in\Theta}\frac{1}{2}\sum_{i\in\Gamma}D_{KL}(p_i^\sigma\|q_i)+\frac{3n-6}{2}+K\\ &= \frac{1}{2}\left(\sum_{i\in\Gamma}\min_{\sigma_i\in\Theta_i}D_{KL}(p_i^\sigma||q_i)\right)+\frac{3n-6}{2}+K\\ &= \frac{3n-6}{2}+K=LB(1).\end{aligned}$$

□

Table 3.1 provides a computational confirmation of the above proposition, by showing a comparison of the normalized values of the lower bounds $LB(0)$ and $LB(1)$ for random uniform or additive matrices $\hat{\mathbf{S}}$ having increasing values of n. The table shows that if $\hat{\mathbf{S}}$ is uniform then $LB(0) \geq LB(1)$. Vice versa, if $\hat{\mathbf{S}}$ is additive then $LB(0) \leq LB(1)$.

Notice that it is a strong assumption to assert $\hat{\mathbf{S}}$ being uniform or additive because we set $K = 0$ for the constant K associated with the doubly-stochastic rescaling required to obtain $\hat{\mathbf{S}}$ from a distance matrix $\mathbf{D}$ (see Proposition 3.2 and Equation (3.10)). In practice such evolutionary distances $\hat{\mathbf{S}}$ are usually not derived from molecular data. For example,

	Input distance matrix $\mathbf{D}$						Input distance matrix $\hat{\mathbf{S}}$					
Data set	Opt.	LB(0)	LB(0.1)	LB(0.3)	LB(0.6)	LB(1)	Opt.	LB(0)	LB(0.1)	LB(0.3)	LB(0.6)	LB(1)
Primates12	803	-82,691	-1,033	-5,334	-11,785	-20,387	84,637	74,351	82,777	78,477	72,025	63,424
M17	542	-83,859	-1,551	-5,852	-12,303	-20,905	84,937	74,389	82,851	78,550	72,099	63,497

Table 3.2. An example of some lower bounds LB(α) for input distance matrices $\mathbf{D}$ and $\hat{\mathbf{S}}$ on 12 taxa. The rescaled input $\hat{\mathbf{S}}$ was obtained from $\mathbf{D}$ by the first method described in [Parlett and Landis, 1982].

consider in Table 3.2 two real aligned DNA data sets of evolutionary distances built by using the general time-reversible model of DNA sequence evolution Catanzaro et al. [2006b]. We can observe that the values for the rescaling constant K reflect a vast change in the input data, e.g., in the case of LB(1) for the data set "Primates12" Equation (3.10) tells us that

$$-20,387 = 63,424 + K \quad \Leftrightarrow \quad K = -83,811.$$

This shift makes most lower bounds LB(α) redundant because $L(T) \geq 0$. However, for the same data set we can find that (up to an accuracy of 10^{-6}) the optimal lower bound with input distance matrix $\mathbf{D}$ and $\hat{\mathbf{S}}$ is achieved for $\alpha = 0.011595$ at LB(α) $= 744$ and LB(α) $= 84,580$, respectively. This means, even though on average the quality of lower bounds decreases for input $\mathbf{D}$ compared to $\hat{\mathbf{S}}$, finding a good approximation of the optimal choice for α provides some utility to bound the BMEP length function. We will discuss the practical implications of this observation in Chapter 5.

3.5 Information theory and the minimum evolution criterion

In this section we want to illustrate the connections between phylogenetic estimation problems under minimum evolution and information theory more broadly. As we have seen in Section 2.3.2, one of the earliest approaches to formulate a phylogenetic estimation problem under the minimum evolution criterion was Rzhetsky and Nei [1993]'s OLS model in which edge weights are estimated by solving the following least squares equations (see Problem 3):

$$\mathbf{w} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{D}^{\Delta}$$

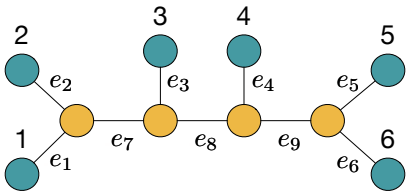


Fig. 3.1. An example of a phylogeny T of a set of six taxa. We define the columns of the EPT matrix $\mathbf{Y}$ corresponding to T in the order of edges $e_1, \ldots, e_9$. Moreover, we consider paths p_{ij} to define rows of $\mathbf{Y}$ before paths p_{kl} for taxa $i < k$ and $j < l$. The resulting least squares estimation looks as follows:

$$(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t = \begin{pmatrix}
\frac{1}{2} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & 0 & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{4} & \frac{1}{4} & 0 \\
0 & 0 & 0 & \frac{1}{8} & -\frac{1}{8} & 0 & 0 & \frac{1}{8} & -\frac{1}{8} & 0 & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{2} \\
0 & 0 & 0 & -\frac{1}{8} & \frac{1}{8} & 0 & 0 & -\frac{1}{8} & \frac{1}{8} & 0 & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & \frac{1}{2} \\
-\frac{1}{2} & \frac{1}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & 0 & 0 & 0 \\
0 & -\frac{1}{4} & \frac{5}{36} & \frac{1}{18} & \frac{1}{18} & -\frac{1}{4} & \frac{5}{36} & \frac{1}{18} & \frac{1}{18} & \frac{2}{9} & \frac{5}{36} & \frac{5}{36} & -\frac{1}{4} & -\frac{1}{4} & 0 \\
0 & 0 & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & 0 & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{2}
\end{pmatrix}$$

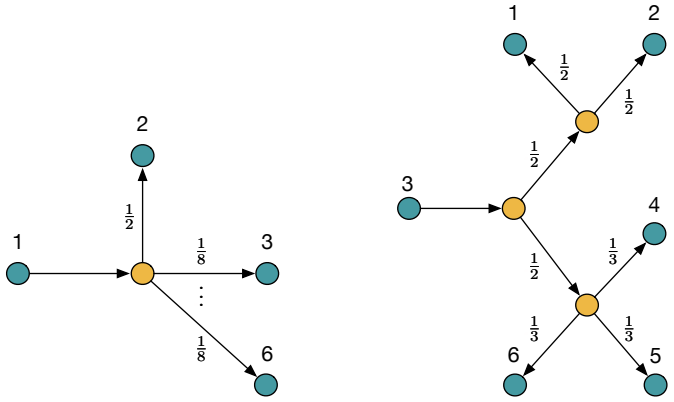


Fig. 3.2. The left tree illustrates the probability distribution $p_1 = (1/2, 1/8, 1/8, 1/8, 1/8)$ obtained from the least squares estimator in Figure 3.1 for edge e_1. The right tree illustrates the probability distribution $p_3 = (1/4, 1/4, 1/6, 1/6, 1/6)$ obtained from the least squares estimator in Figure 3.1 for edge e_3.

An example for this estimation is shown in Figure 3.1. However, as a phylogeny T can be rooted on each taxon $i \in \Gamma$, we can translate some submatrices of the least squares estimator $(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$ into probability distributions. They encode the topological information of the evolutionary process as "seen" from taxon i (see Figure 3.2), i.e., the evolutionary process in which taxon i is the root from which all of the other taxa have been generated over time. Notice that in the OLS estimation model this evolutionary process is not bifurcating. Then, the least square estimator for internal edges of T establishes a consensus between the probability distributions p_i derived for distinct taxa i. The same observations can be made for the BMEP to arrive at a more complete picture: Each term $\mathcal{H}(p_i^\sigma)$ in Equation (3.14) encodes the information entropy of a bifurcating evolutionary process "seen" from the perspective of taxon i. Hence, the term $\mathcal{H}(p_i^\sigma, q_i)$ encodes the error of approximating the evolutionary process "seen" from the perspective of taxon i with the process represented by the estimated evolutionary distances. This opens up the question if the non-bifurcating evolutionary pro-

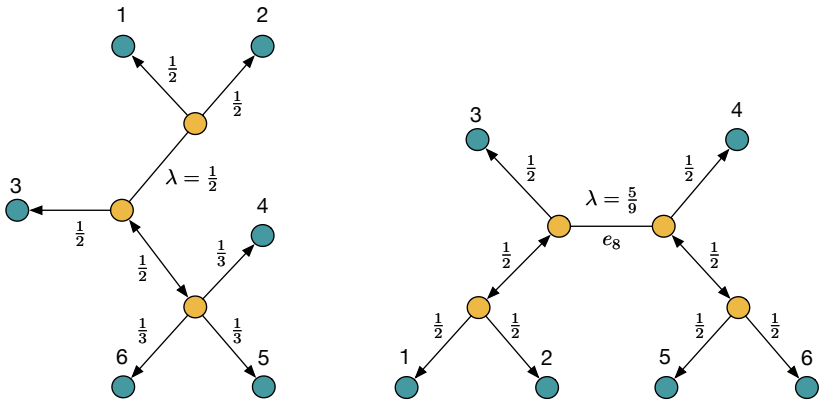


Fig. 3.3. The left tree illustrates the consensus of probability distributions p_i, $i \in \Gamma$, obtained from the least squares estimator in Figure 3.1 for edge e_7. The right tree illustrates the of probability distributions p_i, $i \in \Gamma$, obtained from the least squares estimator in Figure 3.1 for edge e_8.

cesses included in the calculation of the length of a phylogeny under the OLS model induce a cross-entropy with a similar structure as Equation (3.14).

We know that the notion of encoding the topological information of an evolutionary process with respect to a fixed taxon i as a probability distribution p_i^σ is at the core of the proof of statistical consistency for the BMEP Desper and Gascuel [2004]. In fact, not only do vectors p_i^σ, $i \in \Gamma$, allow us to write the average distances (2.16) as inner products between probability distributions, but we can also investigate the edge weight estimation of every internal edge (Equation (2.17)) independently of local topological information. To explain the latter point, consider Figure 3.3. We can observe for edge e_7 that the topological information about the phylogeny is limited. For example, if we look at the weight estimation of edge e_7 as "seen" from the perspective of taxon 1, then taxon 1 evolves into taxa 2 to 6 with probabilities

$$\left(\frac{1}{2}, \lambda \cdot \frac{1}{2}, \lambda \cdot \frac{1}{2} \cdot \frac{1}{3}, \lambda \cdot \frac{1}{2} \cdot \frac{1}{3}, \lambda \cdot \frac{1}{2} \cdot \frac{1}{3}\right) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}\right). \tag{3.24}$$

Clearly, probability distribution (3.24) is an overestimation of the accurate encoding p_1^σ of the phylogeny. We obtain a tighter estimation of p_1^σ for the evolutionary process "seen" from the perspective of taxon 1 when we look at the consensus that our least squares estimator $(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$ finds for edge e_8 in Figure 3.3:

$$\left(\frac{1}{2}, \frac{1}{2} \cdot \frac{1}{2}, \frac{1}{2} \cdot \lambda \cdot \frac{1}{2}, \frac{1}{2} \cdot (1-\lambda) \cdot \frac{1}{2} \cdot \frac{1}{2}, \frac{1}{2} \cdot (1-\lambda) \cdot \frac{1}{2} \cdot \frac{1}{2}\right) = \left(\frac{1}{2}, \frac{1}{4}, \frac{5}{36}, \frac{1}{9}, \frac{1}{9}\right). \tag{3.25}$$

However, the accurate encoding p_1^σ is never fully taken into account for the weight estimation of any internal edge under the OLS model. Since this is the case for the BMEP, the evolutionary process "seen" from the perspective of a taxon i is identical for the weight estimation of every internal edge under balanced minimum evolution. This means, there exists a formula to describe the change in the length function $L(T)$ which is valid for any local topological interchange (e.g. operations shown on the right in Figure 2.19). This explains the elegance of the proof of statistical consistency of the BMEP by Desper and Gascuel in contrast to the one developed by Rzhetsky and Nei for the OLS model and provides evidence for the lack of proofs of statistical consistency for phylogenetic estimation models under the minimum evolution criterion with estimators of evolution different from ones like (3.24) or (3.25). We want to formalize this discussion by unifying Problem 3 and the BMEP from the perspective of information theory.

Let $T \in \mathcal{T}$ be a phylogeny of Γ. Then, for a vertex $v \in V(T)$ or $v \in \Gamma$ and the tripartition $A_0 \cup A_1 \cup A_2 = \Gamma$ or the subset $A_0 = \Gamma_v$, respectively, induced by the removal of v, we consider probability distributions

$$\rho_{v,A_0\cup A_1} = \{\rho_{vw}\}_{w\in A_0\cup A_1}, \quad \rho_{v,A_1\cup A_2} = \{\rho_{vw}\}_{w\in A_1\cup A_2}, \quad \text{and} \quad \rho_{v,A_0\cup A_2} = \{\rho_{vw}\}_{w\in A_0\cup A_2}$$

which encode the topological information of an evolutionary process as "seen" from vertex v to all taxa in $A_0 \cup A_1$, $A_1 \cup A_2$ and $A_0 \cup A_2$, respectively. Moreover, for a distance matrix $\mathbf{D}$, subsets of taxa $A, B \subset \Gamma$ and the roots r_A and r_B of rooted subtrees of T having leaveset A and B, respectively, we define the quantity

$$\beta^T_{A|B} = \sum_{a \in A} \sum_{b \in B} \rho_{r_A a} \rho_{r_B b} d_{ab}. \tag{3.26}$$

We can see that quantity (3.26) is a generalization of the notion of average distance we introduced in Section 2.3.3. For example, if we choose

$$\rho_{v, A_i \cup A_{i+1 \bmod 3}} = \left(\frac{1}{|A_i \cup A_{i+1 \bmod 3}|}, \cdots, \frac{1}{|A_i \cup A_{i+1 \bmod 3}|} \right), \quad i = 0, 1, 2 \tag{3.27}$$

for all $v \in \Gamma \cup V(T)$ with corresponding tripartition $A_0 \cup A_1 \cup A_2 = \Gamma$, then $\beta^T_{A|B} = \delta_{A|B}$ for $A, B \subset \Gamma$. Or if we choose

$$\rho_{v, A_i \cup A_{i+1 \bmod 3}} = \left\{ 2^{-\tau_{vw}} \right\}_{w \in A_i \cup A_{i+1 \bmod 3}}, \quad i = 0, 1, 2 \tag{3.28}$$

for all $v \in \Gamma \cup V(T)$ with corresponding tripartition $A_0 \cup A_1 \cup A_2 = \Gamma$, then $\beta^T_{A|B} = \delta^T_{A|B}$ for $A, B \subset \Gamma$. Furthermore, recall that there always exist a quadripartition $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ and a tripartition $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ associated with an internal edge $e \in E_i(T)$ and an external edge $e \in E_e(T)$, respectively (see Figures 2.10 and 2.11). Then, Problem 3 and the BMEP are special cases of the following problem:

Problem 6 *(The minimum evolution problem under general information)*
Given a set Γ of taxa, a constant $\eta_e \in (0, 1)$ for every quadripartition of Γ and an algorithm $\mathcal{A}$ which computes, for any input phylogeny $T \in \mathcal{T}$, probability distributions

$$\rho_{v, A_i \cup A_{i+1\, mod\, 3}} \in \mathbb{R}^{|A_i \cup A_{i+1\, mod\, 3}|}, \quad i = 0, 1, 2 \tag{3.29}$$

for all $v \in \Gamma \cup V(T)$ with corresponding tripartition $A_0 \cup A_1 \cup A_2 = \Gamma$, find (i) a phylogeny $T \in \mathcal{T}$ having Γ as leaf set, (ii) the weights of the edges of T, such that the following length function is minimized

$$L_{MEG}(T) = \sum_{e \in E(T)} w_e$$

subject to the additional constraints

$$w_e = \frac{1}{2} \left[\eta_e \left(\beta^T_{\Gamma_1|\Gamma_4} + \beta^T_{\Gamma_2|\Gamma_3} \right) + (1 - \eta_e) \left(\beta^T_{\Gamma_1|\Gamma_3} + \beta^T_{\Gamma_2|\Gamma_4} \right) - \left(\beta^T_{\Gamma_1|\Gamma_2} + \beta^T_{\Gamma_3|\Gamma_4} \right) \right] \qquad \forall\, e \in E_i(T) \tag{3.30}$$

$$w_e = \frac{1}{2} \left(\beta^T_{\Gamma_1|\Gamma_2} + \beta^T_{\Gamma_1|\Gamma_3} - \beta^T_{\Gamma_2|\Gamma_3} \right) \qquad \forall\, e \in E_e(T) \tag{3.31}$$

Hence, we can restate the statistical consistency results discussed in this section as follows:

Proposition 3.7. *Problem 6 is statistically consistent if*

(i) $\eta_e = 0.5$ *for every quadripartition of* Γ *and* $\beta^T_{A|B} = \delta^T_{A|B}$.

(ii)

$$\eta_e = \frac{|\Gamma_1||\Gamma_3| + |\Gamma_2||\Gamma_4|}{|\Gamma_1 \cup \Gamma_2| \cdot |\Gamma_3 \cup \Gamma_4|}$$

for every quadripartition $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4 = \Gamma$ *and* $\beta^T_{A|B} = \delta_{A|B}$.

Equation (3.14) and Proposition 3.7 indicate that there might exist a family of phylogenetic estimation problems under the minimum evolution criterion without positivity constraints (Problem 6) which define a class of cross-entropy minimization problems. In this class, Problem 3 and the BMEP use minimal and maximal information about the topology of a phylogeny to determine an optimal solution, respectively (see Equations (3.27) and (3.28)). However, only the latter problem remains statistical consistent if we impose the desirable constraint of positive edge weights to the minimum evolution criterion [Gascuel, 2005]. For Problem 6 with positive edge weights constraints, this poses the following question: under which conditions is a diminished amount of (local) information about the topology of a phylogeny, i.e., a heightened amount of entropy, responsible for statistical inconsistencies? For example, for the phylogeny in Figure 3.1 and probability distribution p_3 in Figure 3.2 associated to Problem 3, we have

$$\mathcal{H}(p_3) \approx 2.292.$$

while the entropy of the same phylogeny and taxon is lower when considering the BMEP:

$$\mathcal{H}(p_3^\sigma) = 2.25.$$

Further research into this question warrants an extension of the notion of cross-entropy. However, from a practical point of view, Equations (3.27) and (3.28) themselves extend the notion of statistical inconsistency, too. Even though the BMEP is statistically consistent and provides an optimal solution with positive edge weights, some entries in the probability distributions of type (3.28) can grow very small because $2^{-\tau_{ij}} \geq 2^{1-n}$, i.e., they can potentially vanish if they are too small to be represented by floating-point arithmetics on a machine and therefore may lead to an underestimation of p_i^σ, $i \in \Gamma$. We will discuss this practical aspect in more detail in Chapter 5.

3.6 Concluding remarks

The key aspect of the interpretation of the BMEP as a cross-entropy minimization problem is a change of perspective on the information encoded by the matrix τ associated to a given phylogeny of Γ and the input distance matrix $\mathbf{D}$. Specifically, because the rows and the columns of τ must satisfy Kraft's equalities, they can be interpreted, up to a constant factor of two, as probability distributions. Propositions 3.1 and 3.2 show that an analogous interpretation is possible for the input matrix $\mathbf{D}$. Then, the optimal solution to the BMEP can be interpreted as the *consensus tree* [Felsenstein, 2004] between the phylogenies rooted on each taxon $i \in \Gamma$, i.e., as the Pareto-optimal solution between the concurrent minimum entropy processes encoded by the differences $\mathcal{H}(p_i^\sigma, q_i) - \mathcal{H}(p_i^\sigma)$ (see Equation (3.14)). This reinterpretation establishes new lower bounds on the BMEP length function that can be efficiently calculated as well as a new framework for the analysis of the statistical consistency of phylogenetic estimation problems under minimum evolution different from the BMEP. However, it remains unclear which theoretical aspects of the minimum evolution criterion are responsible for statistical inconsistencies. Moreover, we do not know how to exploit these insight in order to extend estimation models under minimum evolution to phylogenetic networks who do not form trees. This extension, e.g., could accurately capture how viruses mutate in time, space, and in presence of multiple spreading events. In the next chapter we want to discuss the structure of phylogenies that could aid the development of estimation models for phylogenetic networks in the future.

4

Computational Complexity Aspects of the Balanced Minimum Evolution Problem

The *Fixed-Tree BMEP (FT-BMEP)* is a special case of the *Balanced Minimum Evolution Problem* (BMEP) that consists of finding the assignment of a set of n taxa to the n leaves of a given unrooted binary tree so as to minimize the BMEP objective function. Deciding the computational complexity of the FT-BMEP has been an open problem for almost a decade. Here, we show that a few modifications to Fiorini and Joret's proof of the $\mathcal{NP}$-hardness of the BMEP suffice to prove the general $\mathcal{NP}$-hardness of the FT-BMEP as well as its strong inapproximability. Furthermore, we discuss the differences in complexity between the FT-BMEP and the BMEP. Since the set of feasible solutions to the FT-BMEP is a proper subset of the feasible solutions to the BMEP, there might exist a polynomial time algorithm to explore the set of phylogenies $\mathcal{T}$ and reduce the solvability of the BMEP to the solvability of a series of instances to the FT-BMEP. We want to weaken this idea in the later sections of this chapter.

4.1 On the complexity of the taxa assignment problem on a fixed unlabeled phylogeny

We denote $\Pi(n)$ as the set of permutations of a given set of n elements and $\mathbb{R}_{0^+}$ as the set of the non-negative real numbers. Recall that a phylogeny T of Γ is *unlabeled* when the input set of taxa has not yet been assigned to its leaves. In other words, an unlabeled phylogeny of Γ is just an UBT with n leaves (see e.g., Figure 4.1). Also recall that two unlabeled phylogenies T_1 and T_2 are *isomorphic* if there exists a graph isomorphism between T_1 and T_2, i.e., a bijection ρ from the vertex set of T_1 to the vertex set of T_2 such that two vertices, say

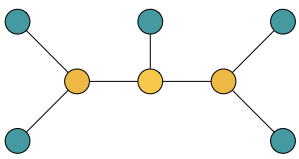


Fig. 4.1. An example of an unlabeled phylogeny of five taxa.

u and v, are adjacent in T_1 if and only if $\rho(u)$ and $\rho(v)$ are adjacent in T_2. We denote $\mathcal{T}_U$ as the set of the possible unlabeled non-isomorphic phylogenies of Γ [Felsenstein, 2004]. Finally we recall that the phylogenies of Γ satisfy the following *Kraft equality* (Equation 2.26):

$$\sum_{j \in \Gamma_i} \frac{1}{2^{\tau_{ij}}} = \frac{1}{2} \quad \forall\, i \in \Gamma. \tag{4.1}$$

In the light of the above notation and definitions, in this section we investigate the computational complexity of the following problem:

Problem 7 *(The Fixed-Tree Balanced Minimum Evolution Problem [FT-BMEP])*
Given a positive integer $n \geq 3$, a set $\Gamma = \{1, 2, \ldots, n\}$ of n taxa, a symmetric distance matrix $\mathbf{D} \in \mathbb{R}_{0^+}^{n \times n}$, and a fixed unlabeled phylogeny $T \in \mathcal{T}_U$, find a permutation $\pi^* \in \Pi(n)$ such that

$$\pi^* = \underset{\pi \in \Pi(n)}{\arg\min} \left\{ L_\pi(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} d_{ij} 2^{-\tau_{\pi(i)\pi(j)}} \right\}.$$

To determine the computational complexity of the FT-BMEP we consider the following decision problem, hereafter referred to as the *Fixed-Tree Balanced Assignment Problem* (FT-BAP):

Problem 8 *(Fixed-Tree Balanced Assignment Problem [FT-BAP])*
Given a positive constant B, a positive integer $n \geq 3$, a set $\Gamma = \{1, 2, \ldots, n\}$ of n taxa, a symmetric distance matrix $\mathbf{D} \in \mathbb{R}_{0^+}^{n \times n}$, and a fixed unlabeled phylogeny $T \in \mathcal{T}_U$, is there a permutation $\hat{\pi} \in \Pi(n)$ such that $L_{\hat{\pi}}(T) \leq B$?

We will show that the FT-BAP is $\mathcal{NP}$-complete and inapproximable within a constant factor unless $\mathcal{P} = \mathcal{NP}$. The $\mathcal{NP}$-hardness of the FT-BMEP will then follow as a direct consequence of this result. Similarly to Fiorini and Joret [2012], in our proof we will use a reduction from the following $\mathcal{NP}$-Complete decision problem [Garey and Johnson, 2003]:

Problem 9 *(The 3-Colorability Problem [3CP])*
Given an undirected graph $G = (V, E)$, can V be partitioned into three stable sets, i.e. sets in which no two vertices are adjacent in G?

Denoting $(\Gamma, \mathbf{D}, T, B)_{FT-BAP}$ and $G = (V, E)_{3CP}$ as an instance of FT-BAP and an instance of 3CP, respectively, the following proposition holds:

Proposition 4.1. *The FT-BAP is $\mathcal{NP}$-Complete.*

Proof. Given any instance $G = (V, E)_{3CP}$ of the 3CP we show that we can construct in polynomial-time an instance $(\Gamma, \mathbf{D}, T, B)_{FT-BAP}$ of the FT-BAP such that the following claim holds true:

$$\begin{aligned} &\text{The graph } G = (V, E)_{3CP} \text{ is 3-colorable} \iff \\ &\qquad \exists\ \hat{\pi} \in \Pi(n) : L_{\hat{\pi}}(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\tau_{\hat{\pi}(i)\hat{\pi}(j)}}} \leq B. \end{aligned} \tag{4.2}$$

where $n = |\Gamma|$. To this end, we set $p := |V|$ and $m := |E|$. Let λ be an arbitrary constant such that $\frac{2}{3} - \frac{\log_2(p(p-1)/2)}{p} < \lambda < \frac{2}{3}$. We remark that, for $p \geq 218$, we have $\frac{3}{5} < \lambda < \frac{2}{3}$ and $m \leq p(p-1)/2 \leq 2^{(2/3-\lambda)}$. This assumption is without loss of generality as we are interested in proving condition (4.2) asymptotically. We also denote k as the smallest positive integer satisfying $k \geq 3p/(2\lambda - 1)$ and $k \equiv 0 \pmod 3$. Finally, let $V = \{v_1, v_2, \ldots, v_p\}$. Now, we define an instance of the FT-BAP as follows. We set $n := 3p + k$, $\Gamma := \{1, 2, \ldots, n\}$, and we associate the first p taxa in Γ with the corresponding vertices of $G = (V, E)_{3CP}$ so that whenever we consider taxon $i \leq p$ in Γ we also refer implicitly to the corresponding vertex v_i in $G = (V, E)_{3CP}$ and vice versa. This means that any permutation that assigns the taxa in Γ to the leaves of T is defined by the following two properties: (i) assign taxa $\{1, \ldots, p\}$ to p leaves of T; (ii) assign taxa in Γ not corresponding to vertices in G to the remaining $n - p$ leaves of T. We also define the generic entry d_{ij} of the distance matrix $\mathbf{D}$ as

$$d_{ij} := \begin{cases} 1 & \text{if } \max\{i, j\} \leq p \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \qquad \forall\ i, j \in \Gamma.$$

Finally, we set the constant $B := 2^{1-\lambda k-(2/3-\lambda)(7-8\lambda)n/3}$ and we construct an unlabeled phylogeny $T \in \mathcal{T}_U$ as in Figure 4.2, i.e., we join an internal vertex v, hereafter referred to as *centroid* [Jordan, 1869], with three rooted subtrees, referred to as *subcaterpillars* T_1,T_2 and T_3, each containing $p + k/3$ leaves. It is easy to see that the construction process of $(\Gamma, \mathbf{D}, T, B)_{FT-BAP}$ can be carried out in polynomial-time and that it is valid because $k \equiv 0 \pmod 3$.

In order to prove claim (4.2) we first show the following intermediate result:

$$\begin{aligned} &\text{If } \exists\ \hat{\pi} \in \Pi(n) : \tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k \text{ for every edge } (v_i, v_j) \in E, \\ &\text{then } G = (V, E)_{3CP} \text{ is 3-colorable.} \end{aligned} \tag{4.3}$$

To prove (4.3), consider any edge $(v_i, v_j) \in E$ such that $\tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k$. Then, the considered permutation $\hat{\pi}$ must assign i and j to leaves located in distinct subcaterpillars T_l, $l \in \{1, 2, 3\}$, otherwise one has

$$\tau_{\hat{\pi}(i)\hat{\pi}(j)} \leq p + \frac{k}{3} \leq \frac{2}{3}\lambda k < \lambda k,$$

where the first inequality is derived from the maximal distance between two leaves in a subcaterpillar T_l, $l \in \{1, 2, 3\}$, and the second inequality is a reformulation of our assumption $k \geq 3p/(2\lambda - 1)$. This fact, however, contradicts the hypothesis of having $\tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k$. Therefore, the sets S_l of taxa assigned to the leaves of the subcaterpillars T_l, $l \in \{1, 2, 3\}$

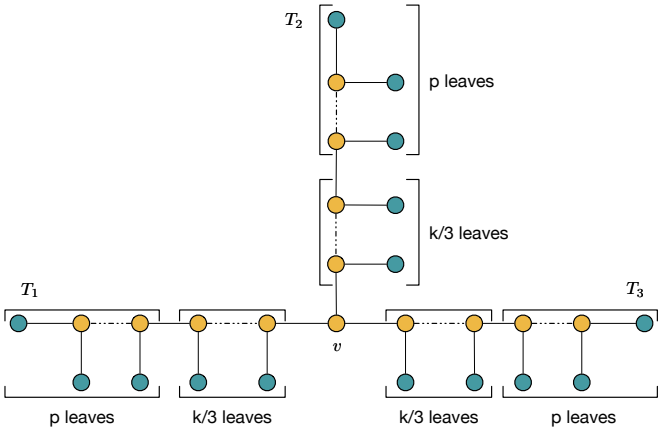


Fig. 4.2. An example of an unlabeled phylogeny with a centroid v and three subcaterpillars.

induce a partition of the vertices of $G = (V, E)_{3CP}$ into three stable sets. Thus, $G = (V, E)_{3CP}$ is 3-colorable and the statement of claim (4.3) follows.

It is worth noting that if $G = (V, E)_{3CP}$ is not 3-colorable then, by contraposition, for any permutation $\pi \in \Pi(n)$, there exists at least one pair of adjacent vertices in $G = (V, E)_{3CP}$, say v_s and v_l, such that $\tau_{\pi(s)\pi(l)} \leq \lambda k$. As a consequence, for any permutation $\pi \in \Pi(n)$,

$$L_\pi(T) = 2 \cdot \sum_{(v_i,v_j)\in E} \frac{1}{2^{\tau_{\pi(i)\pi(j)}}} \geq 2 \cdot 2^{-\tau_{sl}} \geq 2^{1-\lambda k}.$$

Since $B < 2^{1-\lambda k}$ by our choice of B, we can deduce that if there exists a permutation $\pi \in \Pi(n)$ with $L_\pi(T) \leq B$ then $G = (V, E)_{3CP}$ is certainly 3-colorable.

Conversely, assume that $G = (V, E)$ is 3-colorable. Then, let S_1, S_2 and S_3 denote the sets constituting the tripartition of V induced by the 3-coloration. Moreover, consider the following permutation $\pi \in \Pi(n)$ that assigns the taxa in Γ to the leaves of T:

1. for each $l \in \{1, 2, 3\}$, assign arbitrarily the taxa corresponding to the vertices in S_l to the leaves of the l-th subcaterpillar of T that are farthest from the centroid (see Figure 4.2);

2. assign the remaining $2p + k$ taxa arbitrarily to the remaining leaves.

By construction, for any $(v_i, v_j) \in E$, it holds that $\tau_{\pi(i)\pi(j)} > 2k/3$ (see, again, Figure 4.2). Moreover, by recalling that $m = |E|$, it also holds that

$$L_\pi(T) = 2 \cdot \sum_{(v_i,v_j)\in E} \frac{1}{2^{\tau_{\pi(i)\pi(j)}}} < \sum_{(v_i,v_j)\in E} 2^{1-2k/3} = 2^{1-2k/3} m.$$

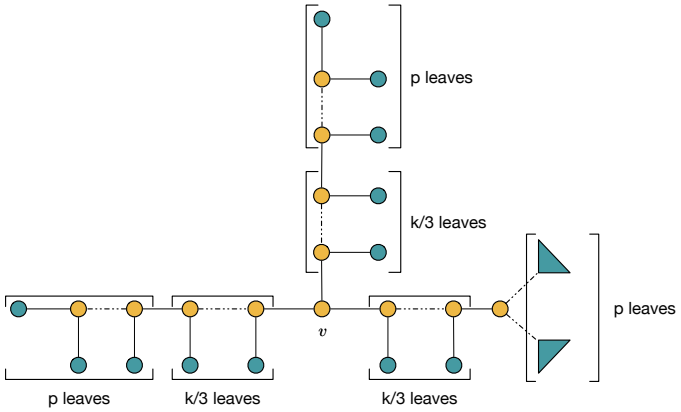


Fig. 4.3. An example of an unlabeled phylogeny with a centroid v and two subcaterpillars (up and left) and a generic subtree (right). The gray triangles represent two generic rooted binary trees.

Now, observe that because $n = k+3p$, $m \leq 2^{(2/3-\lambda)p}$ and $|V| = p \leq (2\lambda-1)k/3 \leq (2\lambda-1)n/3$, it holds that

$$\begin{aligned} \frac{2^{1-\lambda k}}{m \cdot 2^{1-2k/3}} &= \frac{2^{(2/3-\lambda)k}}{m} \geq 2^{(2/3-\lambda)(k-p)} \\ &= 2^{(2/3-\lambda)(n-4p)} \geq 2^{(2/3-\lambda)(7-8\lambda)n/3} \end{aligned} \tag{4.4}$$

which is equivalent to $2^{1-2k/3}m \leq 2^{1-\lambda k-(2/3-\lambda)(7-8\lambda)n/3} = B$. In other words, the length function satisfies $L_\pi(T) \leq B$. This completes the proof of claim (4.2), and the statement of the proposition follows.

□

Proposition 4.1 shows that there exists at least one specific unlabeled phylogeny $T \in \mathcal{T}_U$ for which the FT-BAP is $\mathcal{NP}$-complete. Indeed, there exist exponentially many other unlabeled phylogenies in $\mathcal{T}_U$ for which a similar result holds. For example, all of the unlabeled phylogenies that can be obtained from the tree shown in Figure 4.2 by arbitrarily rearranging the topology connecting the p leaves in each target subcaterpillar T_l (see, e.g., Figure 4.3). Moreover, it is easy to realize that by adjusting appropriately the values of k and λ in the above proof, the $\mathcal{NP}$-completeness still persists also for unlabeled phylogenies characterized by a *bicentroid* instead of a centroid, i.e., a unique edge whose removal decomposes T into two subtrees containing roughly $n/2$ taxa each. Note that, according to the following result, a tree always contains either a centroid or a bicentroid:

Proposition 4.2. *[Jordan, 1869] Let T be a tree with n vertices.*

1. *If $n = 2k+1$ for some $k \in \mathbb{N}$, then there exists a unique vertex c in T, called the centroid, such that all (two or more) subtrees obtained by removing c contain at most k vertices.*

2. *If $n = 2k$ for some $k \in \mathbb{N}$, then there exists in T either*

a) a unique vertex c, called the centroid, such that all (three or more) subtrees obtained by removing c contain less than k vertices, or

b) a unique edge b, called the bicentroid, such that the two subtrees obtained by removing b contain exactly k vertices.

The following result completes our study:

Proposition 4.3. *There exists a constant $c > 1$ such that FT-BAP has no c^n-approximation algorithm unless $\mathcal{P} = \mathcal{NP}$.*

Proof. Consider again a 3CP instance $(V, E)_{3CP}$ and reduce this instance to $(\Gamma, \mathbf{D}, T, B)_{FT-BAP}$ as described in the previous proof. In addition, set

$$c := 2^{(2/3-\lambda)(7-8\lambda)/3} > 1$$

Then, it follows from inequality (4.4) that a c^n-approximation algorithm for the FT-BAP could be used to decide whether G is 3-colorable or not.

□

4.2 Exploration of the set of phylogenies

The $\mathcal{NP}$-hardness of the FT-BMEP suggests that the major difficulty in solving the BMEP rests in the determination of the bijection between leaves and taxa of an optimal phylogeny.

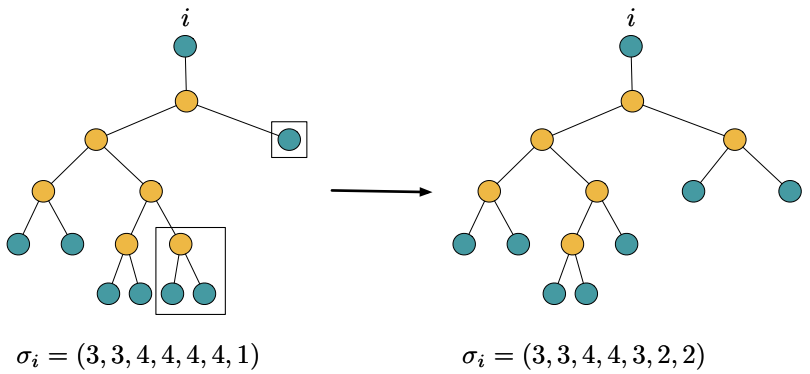


Fig. 4.4. Two phylogenies on eight taxa where the following *ternary exchange* transforms the phylogeny on the left into the one on the right: Choose three distinct taxa, two of which are adjacent to the same internal node. Then, remove the shared internal node and the singleton taxon from the phylogeny, swap them and reconnect both to the tree. The corresponding rooted path-length sequences σ_i encode the path-lengths between i and taxa Γ_i who are assigned to the leaves from left to right.

Therefore, exploiting the structure of the set $\mathcal{T}_U$ of unlabeled non-isomorphic phylogenies might prove beneficial to develop new search strategies for approximation algorithms for the BMEP. In this chapter we discuss why viewing the FT-BMEP solely as a nested subproblem of the BMEP does not simplify the analysis of the problem. Indeed, we will present evidence that points to the contrary conclusion.

4.2.1 Order relations on the set of phylogenies

Again, we recall that Θ_i denotes the set of the rooted path-length sequences of taxon $i \in \Gamma$. Then, an example for a natural ordering relation between a pair of rooted path-length sequences in Θ_i is shown in Figure 4.4. Indeed, [Parker and Ram, 1996] showed that the ternary exchange between two rooted path-length sequences forms a partial order $\trianglelefteq$ on the set Θ_i where the antisymmetry, for $s_i, t_i \in \Theta_i$, is defined by

$$s_i \trianglelefteq t_i \quad \text{only if} \quad \sum_{j \in \Gamma_i} (s_i)_j \leq \sum_{j \in \Gamma_i} (t_i)_j. \tag{4.5}$$

Furthermore, the authors proved that ternary exchanges define a lattice on the set Θ_i, i.e., an algebra $\langle \Theta_i, \trianglelefteq, \wedge, \vee \rangle$ where, for all $s_i, t_i \in \Theta_i$, there exists an unique *greatest lower bound* $s_i \wedge t_i$ and an unique *least upper bound* $s_i \vee t_i$ with respect to the partial order $\trianglelefteq$.

These insights on how to partially order rooted path-length sequences pose the question if a similar structure exists for the set of unlabeled phylogenies $\mathcal{T}_U$. More specifically, does any sort of topological change give rise to a lattice on $\mathcal{T}_U$? Alternatively, we could search for

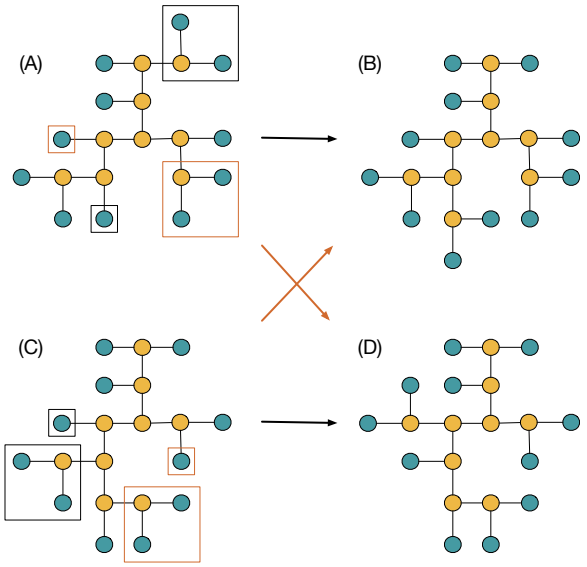


Fig. 4.5. Four phylogenies on eleven taxa where phylogenies (A) and (C) can be transformed by a ternary exchange into phylogenies (B) and (D), respectively. In addition, there exists no ternary exchange to transform phylogeny (A) into (C) and (B) into (D), respectively.

a useful binary relation on $\mathcal{T}_U$ without imposing antisymmetry. However, in our pursuit of approximation schemes for the FT-BMEP and BMEP the lack of a measure of quality to discriminate between distinct phylogenies is not beneficial. Hence, to address the question, observe that any operation to transform one phylogeny into another is based on a composition of ternary exchanges. This means, we are tasked to define a property that establishes the antisymmetry between pairs of unlabeled phylogenies.

For a first approach to this task, recall that any phylogeny can be encoded by a path-length matrix $\tau \in \Theta$ where each row of τ induces a rooted path-length sequence. Hence, we can extend property (4.5) to define a binary relation $\trianglelefteq_U$ as a ternary exchange for pairs of phylogenies corresponding to $\tau, \tau' \in \Theta$ with

$$\tau \trianglelefteq_U \tau' \quad \text{only if} \quad \sum_{i,j \in \Gamma} \tau_{ij} \leq \sum_{i,j \in \Gamma} \tau'_{ij}. \tag{4.6}$$

Then, it is easy to see that $\trianglelefteq_U$ defines a partial order on Θ. When we consider the example of ternary exchanges in Figure 4.5, we notice that calculating $\sum_{i,j \in \Gamma} \tau_{ij}$ for phylogenies (A) to (D) yields values $292, 288, 290$ and 288, respectively. This means, phylogenies (B) and (D) are greatest lower bounds for (A) and (C), respectively. Thus, the partial order $\trianglelefteq_U$ based on property (4.6) does not define a lattice on unlabeled phylogenies. In the next section we will see that using measures different from (4.6) that may give rise to a lattice on $\mathcal{T}_U$ remain insufficient for the analysis of approximation schemes for the FT-BMEP or BMEP.

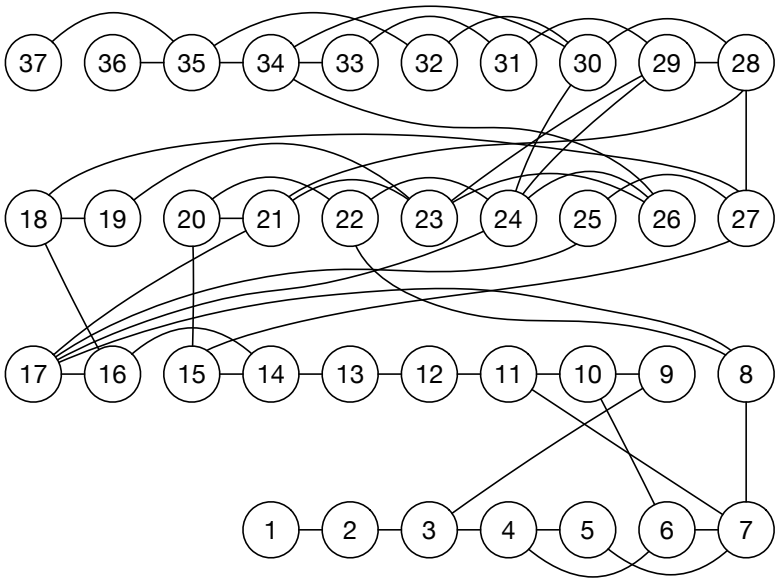


Fig. 4.6. The vertices represent the 37 unlabeled phylogenies on twelve taxa enumerated increasing in $||2^{-\tau}||_F$. The edges form the transitive closure induced by ternary exchanges with respect to the measure $||2^{-\tau}||_F$ on $\mathcal{T}_U$.

4.2.2 Alternative order relations based on path-length matrices

A different way to introduce antisymmetry to the ordering of phylogenies can be identified from the length function of the BMEP by using the Cauchy-Schwarz inequality [Cauchy, 1821, Mitrinović et al., 1993] that allows us to write:

$$L(T) = 2\sum_{i<j} \frac{d_{ij}}{2^{\tau_{ij}}} = \frac{\cos^{-1}(\text{angle}(\text{vec}(D), \text{vec}(2^{-\tau})))}{||D||_F||2^{-\tau} - I||_F}$$

where matrix $2^{-\tau} = \{2^{-\tau_{ij}}\}_{i,j\in\Gamma}$, $||\cdot||_F$ denotes the Frobenius norm of a matrix ([Golub and van Loan, 1996]), $\text{vec}(X)$ denotes the vector we obtain from matrix X by stacking up all non-diagonal entries of the upper triangular matrix of X row by row and $\text{angle}(x, y)$ the angle between vectors x and y. Hence, for $\tau \in \Theta$ corresponding to $T \in \mathcal{T}_U$, the Frobenius norm of $2^{-\tau}$ could offer an useful partial order of unlabeled phylogenies to analyze the behaviour of function $L(T)$ on the set $\mathcal{T}_U$. For example, Figure 4.6 shows this ordering where no two phylogenies are identical in the measure $||2^{-\tau}||_F$, i.e., it defines a lattice on $\mathcal{T}_U$ for $n = 12$. However, this insight is not sufficient to improve our analysis so far as the next proposition illustrates:

Proposition 4.4. *Let $T \in \mathcal{T}_U$ be the unlabeled phylogeny corresponding to path-length matrix*

$$\tau = \begin{pmatrix} 0\,2\,3\,4 & \cdots & (n-2) & (n-1) & (n-1) \\ 0\,3\,4 & \cdots & (n-2) & (n-1) & (n-1) \\ 0\,3 & \cdots & (n-3) & (n-2) & (n-2) \\ 0 & & \cdots & (n-3) & (n-3) \\ & & \ddots & \vdots & \vdots \\ & & 0 & 3 & 3 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}.$$

and consider the first row of τ by $\hat{\sigma}_1 = (\tau_{12} - 1, \ldots, \tau_{1n} - 1) \in \Theta_1$. Then,

(i) $||2^{-\tau}||_F^2 = (75n + 14 + 4^{4-n})/72$,

(ii) $||p_1^{\hat{\sigma}}||_2^2 = (1 - 4^{2-n})/3 + 2^{4-2n} = \max\{||p_1^{\sigma}||_2^2 : \sigma_1 \in \Theta_1\}$.

Proof. First, we have

$$\begin{aligned}\frac{1}{2}||2^{-\tau} - I||_F^2 &= \sum_{i<j} 2^{-2\tau_{ij}} \\ &= 2 \cdot 2^{-2\cdot 2} + \sum_{i=3}^{n-2}(n+2-i) \cdot 2^{-2i} + 4 \cdot 2^{-2(n-1)} \\ &= \frac{1}{8} + \frac{1}{9} \cdot 2^{-2n-4}(3 \cdot 2^{2n} \cdot n - 2^{2n+2} - 2048) + 2^{4-2n} = \frac{3n + 14 + 4^{4-n}}{144}.\end{aligned}$$

Hence, claim (i) holds. Furthermore, we get

$$\begin{aligned}||p_1^{\hat{\sigma}}||_2^2 &= \sum_{i=1}^{n-2} 2^{-2i} + 2^{-2(n-2)} \\ &= \frac{1}{4}\sum_{i=0}^{n-3}\left(\frac{1}{4}\right)^i + 2^{4-2n} = \frac{1 - 4^{2-n}}{3} + 2^{4-2n}.\end{aligned}$$

Therefore the first equation of claim (ii) holds. Now, consider the rooted path-length sequence σ_i' that we obtain from $\hat{\sigma}_i$ through a ternary exchange between taxa j_1 and j_2 with $\hat{\sigma}_{ij_1} = \hat{\sigma}_{ij_2} = n-2$ and taxon j_3 with $\hat{\sigma}_{ij_3} = k < n-3$. Let $l = n-2-k \geq 2$. Then, we get

$$\begin{aligned}||p_i^{\hat{\sigma}}||_2^2 - ||p_i^{\sigma'}||_2^2 &= 2^{-2k} + 2 \cdot 2^{-2(n-2)} - 2 \cdot 2^{-2(k+1)} - 2^{-2(n-3)} \\ &= (1 + 2^{1-2l} - 2^{-1} - 2^{2-2l}) \cdot 2^{-2k} \\ &= \left(2^{-1} - 2^{1-2l}\right) \cdot 2^{-2k}\end{aligned}$$

We conclude that $||p_i^{\hat{\sigma}}||_2^2 \geq ||p_i^{\sigma'}||_2^2$ because $l \geq 2$. Furthermore, observe that any rooted path-length sequence can be obtained from $\hat{\sigma}$ by applying a series of ternary exchanges that have the same properties as the ternary exchange between $\hat{\sigma}$ and σ'. Thus, the second equation of claim (ii) follows.

□

Notice that rooted path-length sequences $\sigma_i \in \Theta_i$ can be ordered by sorting their entries [Parker and Ram, 1996]. This technique is applicable independently of the condition used to ensure antisymmetry, as illustrated in the proof of Proposition 4.4. The lack of sufficient conditions for path-length matrices $\tau \in \Theta$ prevents an analysis for τ similar to the one shown for $\sigma_i \in \Theta_i$. This means, solely defining an antisymmetry condition is not sufficient to be able to study partial orders on the set $\mathcal{T}_U$.

A third approach that produces the same counterexample is the determinant of $2^{-\tau}$:

$$\det(2^{-\tau}) = \sum_{\pi \in \Pi(n)} \operatorname{sgn}(\pi) \prod_{i \in I} 2^{-\tau_{i\pi(i)}} = \sum_{\pi \in \Pi(n)} \operatorname{sgn}(\pi) 2^{-\sum_{i \in I} \tau_{i\pi(i)}}$$

where $\operatorname{sgn}(\pi)$ denotes the sign of π. In this case, we do not have further insights into vectors

$$\left(\tau_{1\pi(1)}, \ldots, \tau_{n\pi(n)}\right)$$

for different $\pi \in \Pi(n)$ to analyze ternary exchanges based on measure

$$\sum_{i \in \Gamma} \tau_{i\pi(i)}$$

in more detail. However, the determinant of $2^{-\tau}$ remains of interest in the context of semidefinite programming because of the following proposition:

Proposition 1 *The matrix $2^{-\tau}$ is positive definite and the set*

$$C = \{(X, s) \in \mathbb{R}^{n \times n} \times \mathbb{R} : X \text{ positive semidefinite}, s \geq 0, \det(X)^{1/n} \geq s\}$$

is a proper convex cone.

Proof. First, for a matrix X, we consider the n Gerschgorin discs D_i with center x_{ii} and radius $\sum_{j \in \Gamma_i} |x_{ij}|$. Then, all eigenvalues of X lie inside the union of the discs D_i, $i \in \Gamma$ [Gerschgorin, 1931]. For $X = 2^{-\tau}$, every disc D_i is centered in 1 with radius $\sum_{j \in \Gamma} 2^{-\tau_{ij}} = 1/2$ because of the Kraft equality. Since $2^{-\tau}$ is symmetric, all eigenvalues are real and therefore lie in the interval $[0.5, 1.5]$. Thus, $2^{-\tau}$ is positive definite. Our second claim follows from the following observations:

C is a convex cone: Let $(X, s), (Y, t) \in C$ and $\alpha, \beta \geq 0$. Then,

$$\det(\alpha X + \beta Y)^{1/n} \geq \det(\alpha X)^{1/n} + \det(\beta Y)^{1/n} = \alpha \det(X)^{1/n} + \beta \det(Y)^{1/n} \geq \alpha s + \beta t$$

where the first inequality holds due to Minkowski's determinant inequality [Mirsky, 1955].

C is pointed: Suppose $(X, s), (-X, -s) \in C$. Then, $X = 0$ and $s = 0$. Otherwise X would not be positive semidefinite and s would not be non-negative. Thus, $C \cap (-C) = \{0\}$.

C is non-empty: Consider X as the identity matrix and $s = 1$. Then, an open neighbourhood of (X, s) lies in the interior of C.

C is closed: Follows from the fact that the determinant is a continuous function.

□

Now, since the measures for ternary exchanges of phyogenies presented so far do not explain the structure of $\mathcal{T}_U$ well, a semidefinite formulation of the BMEP could present an opportunity to explore matrices in Θ, encoding the set $\mathcal{T}_U$, independently of ternary exchanges. However, the description of the necessary conditions (2.27) and (2.28) for the feasible solutions Θ to

the BMEP both requires an approximation of the logarithm to turn both types of conditions into linear matrix inequalities. This poses a major obstacle to any semidefinite programming formulation because the quality of the approximation of the logarithm can not have a positive impact on the computation time to solve a semidefinite program and the quality of the solution to this program simultaneously. Hence, instead of developing a semidefinite program to circumvent the use of ternary exchanges but imposing constraints involving the logarithm, one could use the constraints themselves instead as a measure for ordering phylogenies. We discuss this possibility in the next section.

4.2.3 Majorization

A more general notion of exchange than ternary exchanges was introduced by Hardy-Littlewood-Pólya's theory of majorization Hardy et al. [1978]. Specifically, the authors defined the following partial order to characterize an exchange between two vectors $x, y \in \mathbb{R}^n$:

$$x \preceq y \quad \text{if and only if} \quad \begin{cases} \sum_{j=1}^{k} \operatorname{sort}\downarrow (x)_j \leq \sum_{j=1}^{k} \operatorname{sort}\downarrow (y)_j & \text{for } k = 1, \ldots, n-1 \\ \sum_{j=1}^{k} x_j = \sum_{j=1}^{k} y_j & \text{for } k = n \end{cases}$$

where the operation $\operatorname{sort}\downarrow (x)$ rearranges the entries of a vector x in non-increasing order. When $x \preceq y$, x is said to be *majorized* by y (y majorizes x). One of the many applications of this definition of exchange was done by Parker and Ram [1996] who showed that Huffman coding is a submodular optimization problem. In particular, the authors proved that ternary exchanges between rooted path-length sequences form a bijection to a majorization order: For $s_i, t_i \in \Theta_i$,

$$s_i \trianglelefteq t_i \quad \text{if and only if} \quad \begin{cases} \sum_{j=1}^{k} \operatorname{sort}\downarrow (2^{-s_i})_j \leq \sum_{j=1}^{k} \operatorname{sort}\downarrow (2^{-t_i})_j & \text{for } k = 1, \ldots, n-1 \\ \sum_{j=1}^{k} 2^{-(s_i)_j} = \sum_{j=1}^{k} 2^{-(t_i)_j} & \text{for } k = n \end{cases}$$

This result rests on the fact that the equality condition of the majorization ordering holds due to Kraft's equality (2.26). It inspires the use of the necessary condition (2.27) analogously for set Θ to establish an equality condition for a majorization ordering on Θ as follows: For $\tau, \tau' \in \Theta$, define

$$\tau \preceq_U \tau' \quad \text{if and only if} \quad \begin{cases} \sum_{i=1}^{k} \sum_{j=1}^{l} \tau_{ij} 2^{-\tau_{ij}} \leq \sum_{i=1}^{k} \sum_{j=1}^{l} \tau'_{ij} 2^{-\tau'_{ij}} & \text{for } k < n \text{ and } l < n \\ \sum_{i=1}^{k} \sum_{j=1}^{l} \tau_{ij} 2^{-\tau_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \tau'_{ij} 2^{-\tau'_{ij}} & \text{for } k = n \text{ and } l = n \end{cases}$$

As we have already seen in Proposition 4.4, a simple ordering of entries of a rooted path-length sequence makes the analysis of topological properties of a phylogeny possible. In the case of path-length matrices $\tau, \tau' \in \Theta$ it is unclear which ordering of rows and columns of τ and τ' makes the majorization ordering $\preceq_U$ useful, if any. In addition, only a subset of the inequalities defining this majorization order might be suitable to make unlabeled phylogenies comparable with respect to $\preceq_U$.

4.3 Concluding remarks

First note that Proposition 4.3 does not make any assumption on the input distance matrix for the FT-BMEP. Hence, an interesting open question is whether there exist instances of the FT-BMEP that can be ε-approximated, for some $\varepsilon > 0$, when dealing with particular types of input distances matrices $\mathbf{D}$. Furthermore, all the obstacles illustrated in this chapter indicate that the sets of labeled phylogenies $\mathcal{T}$ and unlabeled phylogenies $\mathcal{T}_U$ do not differ much in complexity when concerned with the ordering of phylogenies. On the one hand, the geometry of ternary exchanges seems to obstruct an useful structure of the set $\mathcal{T}_U$. On the other hand, when we abandon ternary exchanges, then the potential differences in complexity between the BMEP and the FT-BMEP become less apparent. Thus, connecting the existence of notable partial orders on phylogenies to the $\mathcal{NP}$-hardness of the BMEP or FT-BMEP requires further research. However, the notion of majorization is also interesting outside of an analysis of the structure of $\mathcal{T}_U$. We have seen in Chapter 3 that the choice of probability distributions (3.29) and *consensus parameters* η_e in Problem 6 determine the statistical consistency of a phylogenetic estimation problem under minimum evolution. Moreover, we can observe that probability distribution (3.28) majorizes probability distribution (3.27). This raises the question if there exist sets of probability distribution such that a majorization order is able to discriminate between choices for (3.29) which lead to statistical consistency and ones that produce inconsistencies in a theoretical (Chapter 3) or practical (Chapter 5) sense.

5

A Massively Parallel Exact Solution Algorithm for the Balanced Minimum Evolution Problem

In practice, instances of the BMEP may include hundreds or even thousands taxa (see Chapter 2 and [Lefort et al., 2015, Gascuel and Steel, 2006, Pardi, 2009, Desper and Gascuel, 2008, Auch et al., 2006, Yang, 2014, Bordewich et al., 2009a, Pardi and Gascuel, 2012, Pardi et al., 2010, Pardi and Gascuel, 2016, Criscuolo and Michel, 2009, Jung, 2012, Bordewich et al., 2009b]). Hence, it is highly desirable to design exact solution algorithms able to tackle and solve instances much larger than Catanzaro et al. [2012]'s state-of-the-art algorithm. In this chapter we make a further step in this direction, by building upon the results described in previous chapters and in [Catanzaro et al., 2012, Semple and Steel, 2004, Catanzaro et al., 2020b, 2015, Aringhieri et al., 2011, Gascuel and Steel, 2006, 2016]. We will show in this chapter that combining massive parallelism with new theoretical results on the polyhedral combinatorics of the BMEP [Catanzaro et al., 2020b] leads to a new massively parallel exact solution algorithm for the problem that defines a new reference in terms of performance: up to an order of magnitude faster than the current state-of-the-art exact sequential algorithm on benchmark instances and up to 25% more taxa solved within the same time limit. We will also see that there exists a theoretical limit on the performance achievable by an exact algorithm for the BMEP, which materializes whenever the generic addend $d_{ij}2^{-\tau_{ij}}$ in the length function approaches (or get smaller than) the machine epsilon. Despite the study of the numerical stability aspect remained marginal in literature on the BMEP, it has deep implications on the statistical consistency of the optimal solution T^* as we have hinted at in Chapter 3. In particular, we will see that no rescaling technique may guarantee at the same time the numerical stability of the problem and the statistical consistency of the optimal solution. In Section 5.1 we recall the state-of-the-art exact solution algorithm for the BMEP which exploits the combinatorics of the problem (see Section 2.5.1). Furthermore, we present lower bounds and upper bounds on the optimal solution to the problem that prove useful in the development of a massively parallel exact solution algorithm for the BMEP in Section 5.2. We close with a discussion of the practical performance of our parallel method and numerical stability as well as statistical consistency issues connected to it in Section 5.3.

Before proceeding, we extend the notation used so far slightly. For a given subset $S \subseteq \Gamma$, we define a *sub-phylogeny* T_S of Γ as a phylogeny that involves just the taxa in S, i.e., as a pair constituted by an UBT with $|S|$ leaves and an assignment of the taxa in S to the leaves

of the UBT. We denote $V(T_S)$ and $E(T_S)$ as the sets of internal vertices and edges of T_S, respectively. Moreover, whenever the context will be sufficiently clear for the sake of notation we will drop the index S from T. For example, we shall write just T whenever $S = \Gamma$.

5.1 On the state-of-the-art exact solution algorithm for the BMEP

In this section we recall the foundations of Catanzaro et al.'s exact solution algorithm for the BMEP. To model the BMEP the authors introduced a mixed-integer program which uses binary decision variables to formulate topological distances:

$$x_{ij}^{\ell} = \begin{cases} 1 & \text{if } \tau_{ij} = \ell \\ 0 & \text{otherwise} \end{cases} \quad \forall\, i, j \in \Gamma \cup V, i \neq j, \ell \in \mathcal{L}$$

where $\mathcal{L} = \{1, 2, \ldots, (n-1)\}$. In particular, they distinguished between edge variables $(x_{ij}^{\ell}, \ell = 1)$ and path variables $(x_{ij}^{\ell}, \ell \geq 2)$ and linked them with the following conditions:

$$\begin{aligned} x_{ij}^{\ell} + 1 &\geq x_{il}^{(\ell-1)} + x_{lj}^{1} && \forall\, i \neq j \in \Gamma, l \in V, \ell \in \mathcal{L} \setminus \{1, n-1\}, \\ x_{ij}^{\ell} + x_{ij}^{(\ell-2)} + 1 &\geq x_{il}^{(\ell-1)} + x_{lj}^{1} && \forall\, i \neq j \neq l \in \Gamma \cup V, \ell \in \mathcal{L} \setminus \{1, 2, n-1\}. \end{aligned}$$

The use of edge variables allowed them to model the four-point condition (2.28) on all quartets of vertices of a phylogeny. Since it is unclear whether a restriction to the set Γ is sufficient to encode a phylogeny, modeling the four-point condition inequalities with edge variables ensures the acyclicity of the graphs encoded by the MIP. To this end, they linearized condition (2.28) through binary decision variables:

$$y_{ijqt} = \begin{cases} 1 & \text{if } \tau_{it} + \tau_{jq} \geq \tau_{iq} + \tau_{jt} \\ 0 & \text{otherwise} \end{cases} \quad \forall\, i, j, q, t \in \Gamma \cup V, i \neq j \neq q \neq t.$$

Then, the following mixed-integer program relies on the necessity and sufficency of conditions (2.23)-(2.24) and (2.26)-(2.28) to characterize Θ to model the BMEP:

Formulation 2

$$\min \quad z = \sum_{i\in\Gamma}\sum_{j\in\Gamma_i} d_{ij}\left(\sum_{\ell\in\mathcal{L}\setminus\{1\}} 2^{-\ell}x_{ij}^{\ell}\right) \tag{5.1}$$

$$s.t. \quad \sum_{\ell\in\mathcal{L}} x_{ij}^{\ell} = 1 \qquad \forall\, i,j \in \Gamma\cup V, i\neq j, \tag{5.2}$$

$$x_{ij}^{\ell} = x_{ji}^{\ell} \qquad \forall\, i,j \in \Gamma\cup V, i<j, \ \forall\, \ell\in\mathcal{L}, \tag{5.3}$$

$$\sum_{j\in\Gamma_i}\sum_{\ell\in\mathcal{L}\setminus\{1\}} 2^{-\ell}x_{ij}^{\ell} = \frac{1}{2} \qquad \forall\, i\in\Gamma, \tag{5.4}$$

$$\sum_{\ell\in\mathcal{L}\setminus\{1\}} \ell 2^{-\ell}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i} x_{ij}^{\ell} = (2n-3), \tag{5.5}$$

$$\sum_{\ell\in\mathcal{L}} \ell(x_{ij}^{\ell}+x_{qt}^{\ell}) - \sum_{\ell\in\mathcal{L}} \ell(x_{iq}^{\ell}+x_{jt}^{\ell}) \leq (2n-2)y_{ijqt} \qquad \forall\, i,j,q,t \in \Gamma\cup V, \ i\neq j\neq q\neq t, \tag{5.6}$$

$$\sum_{\ell\in\mathcal{L}} \ell(x_{ij}^{\ell}+x_{qt}^{\ell}) - \sum_{\ell\in\mathcal{L}} \ell(x_{it}^{\ell}+x_{jq}^{\ell}) \leq (2n-2)(1-y_{ijqt}) \qquad \forall\, i,j,q,t \in \Gamma\cup V, \ i\neq j\neq q\neq t \tag{5.7}$$

$$x_{ij}^{1} = 0 \qquad \forall\, i,j\in\Gamma, i\neq j, \tag{5.8}$$

$$\sum_{i,j\in\Gamma\cup V, i\neq j} x_{ij}^{1} = (2n-3), \tag{5.9}$$

$$\sum_{j\in V} x_{ij}^{1} = 1 \qquad \forall\, i\in\Gamma, \tag{5.10}$$

$$\sum_{j\in\Gamma\cup V, i\neq j} x_{ij}^{1} = 3 \qquad \forall\, i\in V, \tag{5.11}$$

$$x_{ij}^{1} + x_{il}^{1} + x_{lj}^{1} \leq 2 \qquad \forall\, i,j,l\in V, i\neq j\neq l, \tag{5.12}$$

$$x_{ij}^{\ell} + 1 \geq x_{il}^{(\ell-1)} + x_{lj}^{1} \qquad \forall\, i,j\in\Gamma, i\neq j, l\in V, \ \ell\in\mathcal{L}\setminus\{1,n-1\}, \tag{5.13}$$

$$x_{ij}^{\ell} + x_{ij}^{(\ell-2)} + 1 \geq x_{il}^{(\ell-1)} + x_{lj}^{1} \qquad \forall\, i,j,l\in\Gamma\cup V, i\neq j\neq l, \ \ell\in\mathcal{L}\setminus\{1,2,n-1\}, \tag{5.14}$$

$$x_{ij}^{\ell} \in \{0,1\} \qquad \forall\, i,j\in\Gamma\cup V, \ell\in\mathcal{L}, \tag{5.15}$$

$$y_{ijqt} \in \{0,1\} \qquad \forall\, i,j,q,t\in\Gamma\cup V, \ i\neq j\neq q\neq t. \tag{5.16}$$

Constraints (5.2) impose that variables τ_{ij} assume exactly one value in L. Constraints (5.3) impose the symmetry equalities (2.24). Constraints (5.4) impose Kraft's equalities (2.26). Constraints (5.6) and (5.7) impose the four-point inequalities (2.28). Constraints (5.8)-(5.14) describe the structure of a phylogeny. Specifically, constraint (5.8) imposes that no edge exists between taxa in Γ. Constraint (5.9) imposes that exactly $(2n-3)$ edges be present in a phylogeny. Constraints (5.10) and (5.11) impose the degree constraint on vertices of a phylogeny. Constraints (5.12) prevent triangles. Finally, constraint (5.5) imposes the phylogenetic manifold (2.27). Furthermore, Catanzaro et al. presented multiple valid modifications of Formulation 2 which enable exact solution algorithms different from the state-of-the-art. Hence, we omit their details.

Now, to evaluate the quality of Formulation 2, Catanzaro et al. used prominent real aligned DNA data sets of evolutionary distances built by using the general time-reversible model of DNA sequence evolution Catanzaro et al. [2006b]. The authors first tested the performance of Formulation 2 by using a Branch-&-Cut approach in which constraints associated with the four-point condition and further valid inequalities were added dynamically. This led to the observation that Formulation 2 is a tight formulation of the BMEP which converged to the optimal solution with significantly less branches than previously known methods. However, apart from the unicity (5.2), Kraft (5.4) and phylogenetic manifold (5.5) constraints, the computation time required to respect all constraints was non-decreasing in consecutive branching steps to such an extend that most tested instances could not be solved for more than a dozen taxa. This motivated Catanzaro et al. to employ a different branching scheme known in the literature as the *Stepwise Addition Strategy* (SAS) that was introduced in the 1960s [Felsenstein, 2004] and formalized for phylogenetic estimation models by Hendy and Penny [1982] in 1982 (see Section 2.5.2). We recall a brief overview of this strategy and highlight relaxations of Formulation 2 that exploit the structural advantages of a SAS in comparison to more generic branching rules.

Given a sub-phylogeny T_S of Γ, taxon $t \in \Gamma \setminus S$, internal vertex $v \notin V(T_S)$ and edge $(a,b) \in E(T_S)$, define a *branching of* T_S *on* t *and* (a,b) as the sub-phylogeny $T_{S\cup\{t\}}$ with

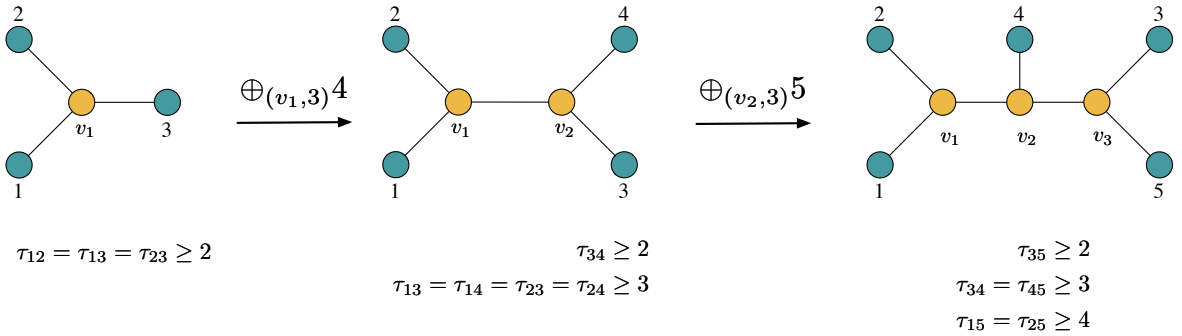


Fig. 5.1. Two feasible branching steps for the SAS. For $n \geq 4$ taxa, the sub-phylogeny in the center is a branching of the sub-phylogeny on the left on taxon 4 and edge $(v_1, 3)$. For $n \geq 5$ taxa, the sub-phylogeny on the right is a branching of the sub-phylogeny in the center on taxon 5 and edge $(v_2, 3)$.

$$E(T_{S\cup\{t\}}) = E(T_S) \setminus \{(a,b)\} \cup \{(a,v),(v,b),(v,t)\}.$$

To simplify the notation for branchings we define the following operation for any branching $T_{S\cup\{t\}}$ of T_S on t and (a,b):

$$T_{S\cup\{t\}} = T_S \oplus_{(a,b)} t.$$

Moreover, to denote the set of taxa underlying the sub-phylogeny T_S we write

$$S = \text{taxa}(T_S).$$

An example of two consecutive branchings is shown in Figure 5.1. First, taxon 4 is inserted into the star graph on three taxa by introducing a new internal vertex v_2, removing the edge $(v_1,3)$ and adding edges (v_1,v_2), $(v_2,3)$ and $(v_2,4)$. Similarly, taxon 5 is inserted into the resulting tree on four taxa by adding internal vertex v_3, edges (v_2,v_3), $(v_3,3)$, $(v_3,5)$ and

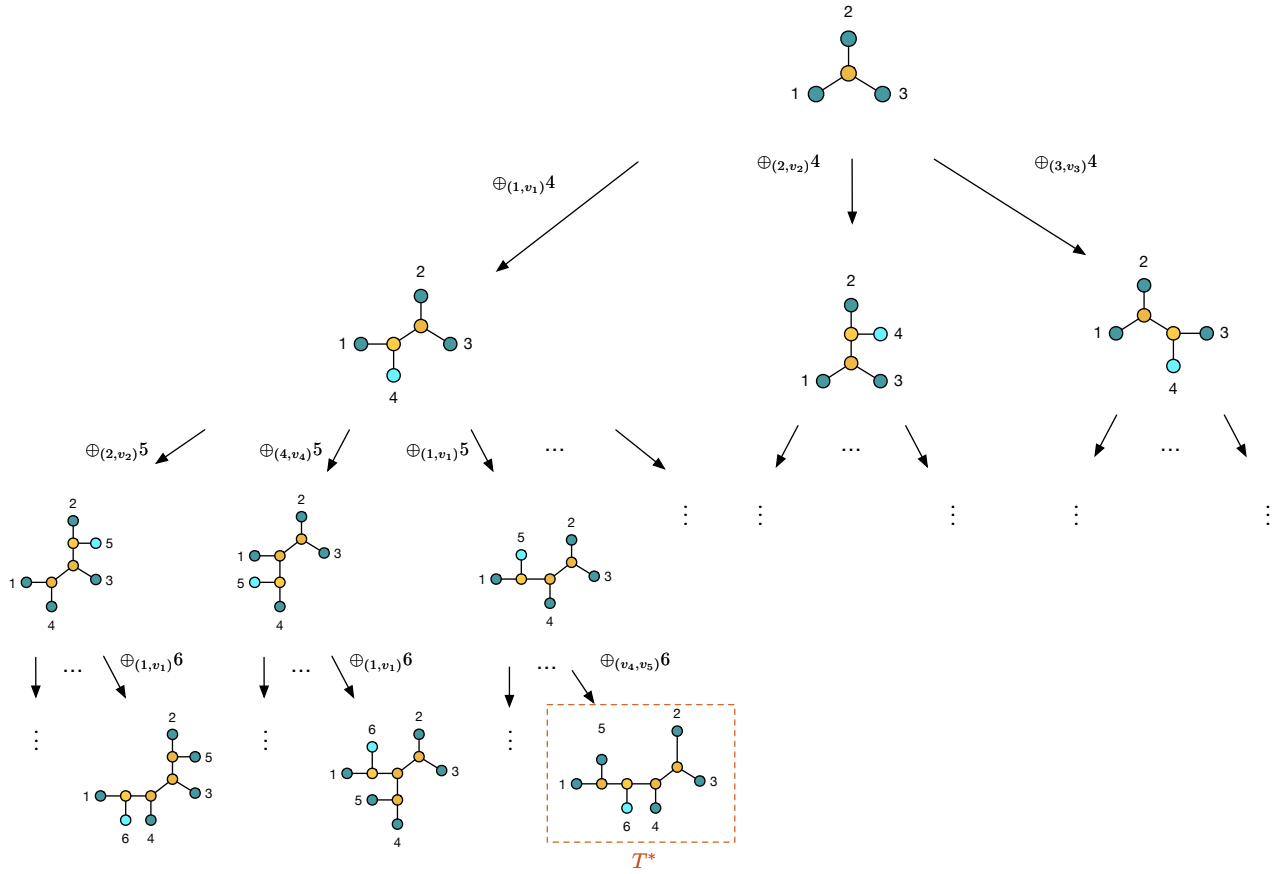


Fig. 5.2. The search tree of a stepwise addition strategy to solve the BMEP for the instance "Primates12" of 6 taxa used in [Catanzaro et al., 2009, Aringhieri et al., 2011, Catanzaro et al., 2012, 2013a, 2015]. The recursion of Algorithm 2 runs from left to right. All labels for internal vertices are omitted and the use of internal vertices in branchings is clear from the context. Moreover, all subphylogenies T_S are omitted for which either lowerbound$(T_S) \geq L(T^*)$ or lowerbound(T_S) is not minimal among all branchings $T_S = T \oplus_e t$ of a fixed sub-phylogeny T. Then, the final phylogeny T^* is optimal for the BMEP by enumeration.

Algorithm 1: Stepwise addition strategy - Initialization and termination

Input: Ordered set of taxa $\Gamma = \{1, \ldots, n\}$, distance matrix D for Γ
Output: A phylogeny T^* of Γ
$T^* \leftarrow$ NULL;
$L(T^*) \leftarrow \infty$;
$S \leftarrow \{1, 2, 3\}$;
$T_S \leftarrow$ The only sub-phylogeny of Γ for S;
Search(Γ, D, T_S, T^*);
return T^*;

removing edge $(v_2, 3)$. Then, for the BMEP, the SAS can be outlined as in Algorithms 1 and 2 where lowerbound(T_S) denotes a specific lower bound on the optimal objective function value of the BMEP. This bound is

(i) derived from the sub-phylogeny T_S,

(ii) valid for the objective function value $L(T)$ of all sub-phylogenies T on taxa(T_S),

(iii) non-decreasing in the number of taxa.

Equipped with properties (i)-(iii), lowerbound(T_S) is a strict lower bound for $L(T^*)$ if and only if it cannot be precluded that a series of branchings of T_S yield an optimal solution to the BMEP. This means, the intermediate candidate for an optimal solution T^* dynamically prunes the search-tree built by the recursion of Algorithm 2. Thus, to specify the implementation of a SAS, one has to define

Algorithm 2: Stepwise addition strategy - Implicit enumeration of all phylogenies (Search)

Input: Ordered set of taxa $\Gamma = \{1, \ldots, n\}$, distance matrix D for Γ, sub-phylogeny T_S of Γ, phylogeny T^* of Γ
if lowerbound$(T_S) < L(T^*)$ **then**
  if taxa$(T_S) = \Gamma$ **then**
    $T^* \leftarrow T_S$;
  else
    $t \leftarrow |\text{taxa}(T_S)| + 1$;
    for $e \in E(T_S)$ **do**
      Search$(\Gamma, D, T_S \oplus_e t, T^*)$;

(A) an order of taxa Γ,

(B) an order in which edges $e \in E(T_S)$ in the for-loop of Algorithm 2 are processed,

(C) lowerbound(T_S) satisfying properties (i)-(iii).

An example for one implementation of a SAS is shown in Figure 5.2. Here, the set of taxa $\Gamma = \{1, \dots, n\}$ is ordered according to the labels of taxa, the ordered set of edges $E(T_S)$ is shown in the choice of the non-omitted branchings, e.g.,

$$E\left(T_{\{1,2,3\}}\right) = \{(1, v_1), (2, v_2), (3, v_3)\} \quad \text{or} \quad E\left(T_{\{1,2,3,4\}}\right) = \{(2, v_2), (4, v_4), (1, v_1), \dots\},$$

and the valid lower bound on sub-phylogenies T_S is given by

$$\text{lowerbound}(T_S) = L(T_S) + \sum_{t \in \Gamma \setminus S} \min_{\substack{i,j \in S \\ i<j<t}} \frac{1}{2} \left(d_{it} + d_{jt} - d_{ij}\right). \tag{5.17}$$

The latter was, among other more intricate bounds, introduced in the context of the SAS by Pardi [2009]. Since Algorithm 2 is an implicit enumeration of all phylogenies of Γ, the phylogeny T^* is optimal for the BMEP when Algorithm 1 terminates. Note that in this example lowerbound(T_S) is never evaluated twice for the same sub-phylogeny T_S. Indeed, in general, the fact that the set of taxa Γ is a totally ordered set ensures that no sub-phylogeny is evaluated more than once. In a SAS every phylogeny is determined by a series of branchings starting with an initial sub-phylogeny. Therefore, every phylogeny we enumerate is unique if the choice of taxa introduced by consecutive branchings is fixed.

However, the choice for an order of sets $V(T_S)$ and $E(T_S)$ as well as the quality of lower bounds on sub-phylogenies can have a major negative impact on the runtime of the SAS. To improve the latter compared to previously known methods, like the SAS of Pardi [2009], Catanzaro et al. [2012] established the following integer programming model, for a given sub-phylogeny T_S, which is closely related to Formulation 2:

Formulation 3

$$\min \quad z(T_S) = \sum_{i\in\Gamma}\sum_{j\in\Gamma_i} d_{ij}\left(\sum_{\ell\in P_{ij}(T_S)} 2^{-\ell}x_{ij}^{\ell}\right) \tag{5.18}$$

$$s.t. \quad \sum_{\ell\in P_{ij}(T_S)} x_{ij}^{\ell} = 1 \qquad \forall\, i,j\in\Gamma, i\neq j \tag{5.19}$$

$$x_{ij}^{\ell} = x_{ji}^{\ell} \qquad \forall\, i,j\in\Gamma, i\neq j, \forall\,\ell\in P_{ij}(T_S) \tag{5.20}$$

$$\sum_{j\in\Gamma_i}\sum_{\ell\in P_{ij}(T_S)} 2^{-\ell}x_{ij}^{\ell} = \frac{1}{2} \qquad \forall\, i\in\Gamma \tag{5.21}$$

$$\sum_{\ell\in P_{ij}(T_S)} \ell 2^{-\ell}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i} x_{ij}^{\ell} = 2n-3 \tag{5.22}$$

$$x_{ij}^{\ell}\in\{0,1\} \qquad \forall\, i,j\in\Gamma, i\neq j, \forall\,\ell\in P_{ij}(T_S)$$

where $P_{ij}(T_S)$ are subsets of $\mathcal{L}$ such that, for path-length matrix τ of T_S,

$$P_{ij}(T_S) = \begin{cases} \{\ell\in\mathcal{L} : \tau_{ij}\leq\ell\leq\tau_{ij}+|\Gamma\setminus S|\} & \text{if } i,j\in S,\\ \{\ell\in\mathcal{L} : 2\leq\ell\leq\max\limits_{s\in S}\tau_{is}+|\Gamma\setminus S|\} & \text{if } i\in S, j\in\Gamma\setminus S,\\ \{\ell\in\mathcal{L} : 2\leq\ell\leq\max\limits_{s,t\in S}\tau_{st}+|\Gamma\setminus S|\} & \text{if } i,j\in\Gamma\setminus S. \end{cases}$$

First, observe that the four-point condition constraints are not present in Formulation 3 and therefore, compared to Formulation 2, the binary variables y_{ijqt} as well as the binary variables x_{ij}^{ℓ} for $i,j\in V$ and all associated constraints are missing from the model. Therefore, Formulation 3 is a relaxation of Formulation 2 for a fixed sub-phylogeny T_S. Furthermore, Catanzaro et al. [2012] made the crucial observation that every branching T_S of a sub-phylogeny introduces lower bounds on the topological distances that are valid for all branchings of T_S. For example, the second branching in Figure 5.1 sets $x_{13}^3 = x_{23}^3 = x_{34}^2 = 0$ and introduces $x_{15}^{\ell} = x_{25}^{\ell} = 0$ for $\ell = 2,3$ and $x_{45}^2 = 0$. The authors proved that the sets $P_{ij}(T_S)$ fully characterize the bounds on topological distances induced by T_S which in turn enables the use of decision variables x_{ij}^{ℓ} only for $\ell\in P_{ij}(T_S)\subset\mathcal{L}$. Thus, solving Formulation 3 provides a worse lower bound on the optimal solution to the BMEP than Formulation 2, but the drastically smaller number of variables makes the solution faster accessible.

Since Catanzaro et al. [2012] showed that the optimal solution $z^*(T_S)$ of Formulation 3 provides a valid choice for lowerbound(T_S) in a SAS, we call the optimal objective function value of its LP relaxation $z^*_{\text{lin}}(T_S)$ the *Linear Programming Lower Bound* (LPLB). Notice that we can reduce the computational cost to obtain $z^*_{\text{lin}}(T_S)$ by removing equalities (5.20) and restricting constraints (5.21) and (5.22) as well as the objective function to $i<j$ for taxa $i,j\in\Gamma$. However, solving the LPLB remains computationally expensive. Therefore, Catanzaro et al. [2012] also introduced a Lagrangian relaxation (LagLB) of the LPLB by using Lagrangian multipliers for constraints (5.21) and (5.22) which yields the following

model:

$$\begin{aligned}
\min \quad & z_{\text{lag}}(T_S, \mu, \lambda) = \sum_{i<j\in\Gamma} \left(\sum_{\ell\in P_{ij}(T_S)} (2d_{ij} - \mu_i - \mu_j - 2\ell\lambda)\, 2^{-\ell} x_{ij}^{\ell} \right) + (2n-3)\lambda + \sum_{i\in\Gamma} \frac{\mu_i}{2} \\
\text{s.t.} \quad & \sum_{\ell\in P_{ij}(T_S)} x_{ij}^{\ell} = 1 \quad \forall\, i, j \in \Gamma, i \neq j \\
& x_{ij}^{\ell} \in [0,1] \quad \forall\, i, j \in \Gamma, i \neq j, \forall\, \ell \in P_{ij}(T_S)
\end{aligned}$$

This approach makes the decision variables x_{ij}^{ℓ} independent of one another. Therefore, for given multipliers μ and λ, LagLB can be solved analytically a lot faster than the LPLB to obtain a valid lower bound for a SAS:

$$\text{lowerbound}(T_S) = z_{\text{lag}}^{*}(T_S, \mu, \lambda). \tag{5.23}$$

A suitable choice for μ and λ are the shadow-prices of constraints (5.21) and (5.22), respectively, that we obtain if we solve the LPLB before employing lower bound (5.23) in the SAS. Thus, Algorithm 3 depicts a valid implementation of Algorithm 2 which combines the ideas behind Formulation 2 to obtain high quality solutions with the advantages of the SAS and LagLB to speed-up overall computations [Catanzaro et al., 2012]. This concludes our discus-

Algorithm 3: Stepwise addition strategy - Implicit enumeration of all phylogenies (ImprovedSearch)

Input: Ordered set of taxa $\Gamma = \{1, \ldots, n\}$, distance matrix D for Γ, sub-phylogeny T_S of Γ, phylogeny T^* of Γ

```
continue ← FALSE;
if |taxa(T_S)| = 3 then
    if z*_lin(T_S) < L(T*) then
        continue ← TRUE;
else
    if z*_lag(T_S) < L(T*) then
        if z*_lin(T_S) < L(T*) then
            continue ← TRUE;
if continue then
    if taxa(T_S) = Γ then
        T* ← T_S;
    else
        t ← |taxa(T_S)| + 1;
        for e ∈ E(T_S) do
            ImprovedSearch(Γ, D, T_S ⊕_e t, T*);
```

sion of the state-of-the-art solver for the BMEP. Hence, in the next section we build on this method to develop a parallel exact solution algorithm.

5.2 A massively parallel exact solution algorithm for the BMEP

In this section we show how the exact solution algorithm of Catanzaro et al. [2012] presented in the last section can be parallelized to improve the method. Moreover, we discuss lower bounds and upper bounds on the optimal solution to the BMEP. To this end, we first investigate how a parallel version of the stepwise addition strategy can operate and why it is well suited for parallel exact solution algorithms for the BMEP.

5.2.1 Branching on partial solutions to the BMEP in parallel

A parallel Branch-&-Bound scheme for the BMEP could be based on a parallelized SAS to make updates of the global intermediate optimal solution T^* less frequent and therefore prune more nodes of the search tree. A naive idea to achieve this objective would be to start, for some fixed $m \in \mathbb{N}$, by enumerating all sub-phylogenies with less than m taxa iteratively. Then, assign every available branching on m taxa to a fixed core of a multi core processor and run a SAS initialized with $S = \{1, \ldots, m\}$ on each core. The issue with this idea is twofold. On the one hand, for any choice of m there are between $((2m-7)!! + 1)$ and $(2m-5)!!$ sub-phylogenies for which a SAS needs to be completed. Hence, the number of SAS's to be completed can be a vast mismatch with the number of cores available, leading to many idle cores. On the other hand, even if all or most of the cores are employed to enumerate candidates for an optimal solution to the BMEP, the workload of each individual core can be very unequal. Looking at the example in Figure 5.2 we can see that if we would assign all branchings on four taxa to three different cores, then two of the cores would turn idle long before the optimal phylogeny is found. To mitigate both problems, we introduce a queue to (i) manage the distribution of the workload for all cores by storing a sufficient among of sub-phylogenies in the queue for idle cores to pop and process, and (ii) rebalance the workload for each individual core by pushing sets of unprocessed sub-phylogenies from some working cores to the queue.

To introduce this procedure formally, we define a *job* as a triple (Γ, T_S, F) where T_S is a sub-phylogeny of Γ and $F \subseteq E(T_S)$ a subset of edges in T_S. Then, we establish a queue managing our jobs as shown in Algorithm 4. First, the star graph on three taxa T_S is constructed and for all three branchings of T_S on taxon 4 we push a job into the queue Q. All calculations until the end of the for-loop can be processed on one core. Then, the while-loop distributes jobs in the queue Q to all available cores of the multi core processor. Furthermore, we process a job $J = (\Gamma, T_S, F)$ from Q on one core C by following the same stepwise addition search strategy as Algorithm 3 while also rebalancing the number of branchings that core C has to

Algorithm 4: SAS - Parallel initialization and termination

Input: Distance matrix D for Γ, queue Q of jobs, ordered set of taxa $\Gamma = \{1, \ldots, n\}$
Output: A phylogeny T^* of Γ
$T^* \leftarrow$ NULL, $L(T^*) \leftarrow \infty$, $S \leftarrow \{1, 2, 3\}$, $T_S \leftarrow$ The only sub-phylogeny of Γ for S;
for $e \in E(T_S)$ **do**
    $T \leftarrow T_S \oplus_e 4$;
    Push $(\Gamma, T, E(T))$ into Q;
while Q not empty **do**
    Execute ProcessJob$(D, Q, \text{pop}(Q), T^*)$ on an idle core;
return T^*;

Algorithm 5: SAS - Implicit enumeration of phylogenies on one core (ProcessJob)

Input: Distance matrix D for Γ, queue Q of jobs, job $J = (\Gamma, T_S, F)$, phylogeny T^*
continue $\leftarrow$ FALSE;
if $|\text{taxa}(T_S)| = 3$ **then**
    if $z^*_{lin}(T_S) < L(T^*)$ **then**
        continue $\leftarrow$ TRUE;
else
    if $z^*_{lag}(T_S) < L(T^*)$ **then**
        if $z^*_{lin}(T_S) < L(T^*)$ **then**
            continue $\leftarrow$ TRUE;
if continue **then**
    if $\text{taxa}(T_S) = \Gamma$ **then**
        $T^* \leftarrow T_S$;
    else
        $t \leftarrow |\text{taxa}(T_S)| + 1$;
        for $e \in F$ **do**
            $T \leftarrow T_S \oplus_e t$;
            if There exist idle cores **or** queue Q contains less jobs than there are overall cores **then**
                Partition $E(T) = F_1 \cup F_2$ such that $||F_1| - |F_2|| \leq 1$;
                if queue Q contains less jobs than there are overall cores **then**
                    Push (Γ, T, F_2) into Q;
                    ProcessJob$(D, Q, (\Gamma, T, F_1), T^*)$;
                else
                    ProcessJob$(D, Q, (\Gamma, T, F), T^*)$;
            ProcessJob$(D, Q, (\Gamma, T, F), T^*)$;

build and evaluate before turning idle as shown in Algorithm 5. After the construction of an unpruned branching T we either partition $E(T)$ into two disjoint subsets F_1 and F_2 whose cardinalities differ at most by one or process the new job (Γ, T, F). This decision depends on the length of queue Q because the creation of new jobs for subsets F_1 and F_2 instead of F can not decrease the total idle time of all cores if every core can already process a job from Q. We verify this state of the queue again once job (Γ, T, F_2) is pushed into Q and the recursion continues on core C. This is necessary due to the fact that either idle cores can decrease the length of Q or working cores different from C can increase the length of Q. Both cases are checked by linking two if-conditions as shown in Algorithm 5.

Proposition 5.1. *Algorithm 4 terminates and returns an optimal solution for the BMEP.*

Proof. Suppose by contradiction that Algorithm 4 does not terminate. This means, queue Q is never empty. Hence, an infinite number of jobs is created in Algorithm 5 because one job is removed from Q in every iteration of the while-loop. However, for the job (Γ, T, F_2) that can be pushed into queue Q and the given job $J = (\Gamma, T_S, F)$ in Algorithm 5, $|F_2|$ is not larger than $|F|$. Therefore, the number of jobs that Algorithms 4 and 5 create is finite, leading to a contradiction. Thus, Algorithm 4 terminates.

To see that Algorithm 4 returns an optimal solution to the BMEP observe that the enumeration order of Algorithms 4 and 5 coincides with the enumeration order of all phylogenies by Algorithms 1 and 3 if the edges in F_1 and F_2 of Algorithm 5 follow the same order as the edges in F. Notice that no additional pruning of the search-tree is introduced throughout the recursion of Algorithm 5. Thus, T^* is an optimal solution for the BMEP by enumeration. □

Proposition 5.1 shows that reorganizing the SAS in parallel with a queue of jobs does extent single-core solvers like Catanzaro et al.'s method. Any implementation of Algorithms 4 and 5 then requires the specification of

(A) the order of taxa Γ,

(B) the order of edges F,

(C) the order of jobs in the queue Q,

(D) the partition of edges $E(T)$.

We can see in Algorithm 5 that the partitioning aspect (D) can be generalized to not only include bipartitions, but any partition $F_1 \cup \cdots \cup F_p$ composed of ordered edge sets $F_i \subseteq E(T), 1 \leq i \leq p$. Since we can only process one job at a time on one core we push jobs (Γ, T, F_i) for $i = 2, \ldots, p$ into the queue Q. Moreover, we keep the assumption $||F_i| - |F_j|| \leq 1$

for $i, j = 1, \ldots, p$ because then every job $(\Gamma, T, F_i), i = 2, \ldots, p$ offers a roughly equal number of branchings to be build and evaluated. This extends aspect (D) of Algorithm 5 to include

(D.1) the number $p \leq |E(T)|$ of partition sets F_i, $i = 1, \ldots, p$,

(D.2) for some positive constant integer c, the minimum cardinality

$$c^*(p) = \min \left\{ \max \left\{ c, \left\lfloor \frac{|E(T)|}{p} \right\rfloor \right\}, |E(T)| \right\}$$

of partition sets F_i, $i = 1, \ldots, p$.

The latter definition arises from the restriction $p \leq |E(T)|$. Since $|E(T)|$ grows linearly in the number of taxa, large values for p may produce an overabundance of jobs for sub-phylogenies defined on few edges. The lower bound (D.2) on the number of edges included in one job ensures that the rebalancing of workloads can be dynamically tailored to the size of sub-phylogenies. Overall, any partition of edges $E(T)$ aims at making at least as many jobs available in the queue as there are cores in the multi core processor. However, as a sufficiently large queue is the bottleneck for any parallel exact algorithm following a stepwise addition strategy, it is unlikely a partitioning process minimizing the total idle time of all cores could be defined a-priori.

Furthermore, the quality of the partition is not only impacted by the order of computed branchings, but also the choice of lower bounds and upper bounds on the optimal solution to the problem. We discuss different choices for such bounds in the next section. In particular, the lower bounds $\mathrm{LB}(\alpha)$ we derived in Chapter 3.

To conclude the discussion on branchings, we consider a popular alternative to the SAS to enumerate all phylogenies called the *agglomerative approach* (see Section 2.5.2). These are

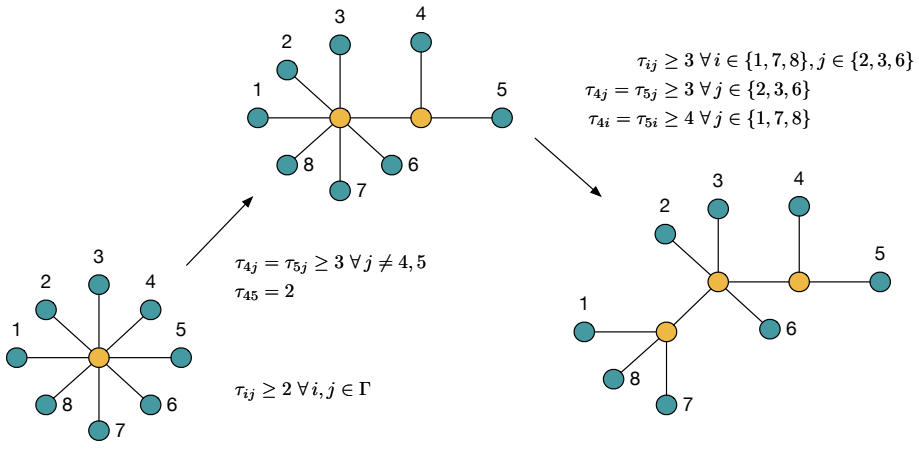


Fig. 5.3. An example of two steps of the agglomerative approach in the case of eight taxa.

constructive methods which start from an infeasible solution to the BMEP and iteratively cluster taxa according to a specific optimality criterion, until when a phylogeny of Γ is obtained. An example of two consecutive steps of the agglomerative approach are shown in Figure 5.3. First, an internal edge is inserted between the subtrees on taxa $\{1, 2, 3, 6, 7, 8\}$ and $\{4, 5\}$ into the star graph on eight taxa. Then, an internal edge is inserted between the subtrees on taxa $\{1, 7, 8\}$ and $\{2, 3, 4, 5, 6\}$ in the resulting tree. Like in the SAS, every step of the agglomerative approach corresponds to branchings on specific topological distances. However, since the insertion of internal edges is fully determined by the incident internal vertices and not the taxa of the tree, this technique does not break symmetries in the search-tree. We can see this by continuing our example: reversing the order in which we insert the internal edges produces a tree not different from the one on the right in Figure 5.3. Hence, we can not expect to obtain a Branch-&-Bound scheme through the agglomeration of taxa which is competitive with the SAS in terms of computation time.

As the two branching approaches presented so far capture all topological information that could be retrieved by inserting external or internal edges, we conclude that the SAS is a suitable choice to compute branchings in our parallel exact solution algorithm for the BMEP. Therefore, we assess the bounding aspect of the SAS in the next section.

5.2.2 Lower bounds and upper bounds on the optimal solution to the BMEP

We have seen in Section 5.1 that $z^*_{\text{lin}}(T_S)$ is a valid lower bound for a SAS on the BMEP derived from some characteristics of the set Θ. Moreover, its Lagrangian relaxation was employed to speed-up the pruning of the search-tree of the SAS. Specifically, sub-phylogenies are pruned and discarded if they do not improve upon the shortest length of all previously constructed phylogenies. Instead of discarding all of them, one might observe that a non-optimal phylogeny shares a majority of topological distances with an optimal solution to the BMEP. For example, the SAS shown in Figure 5.2 constructs, among others, the three

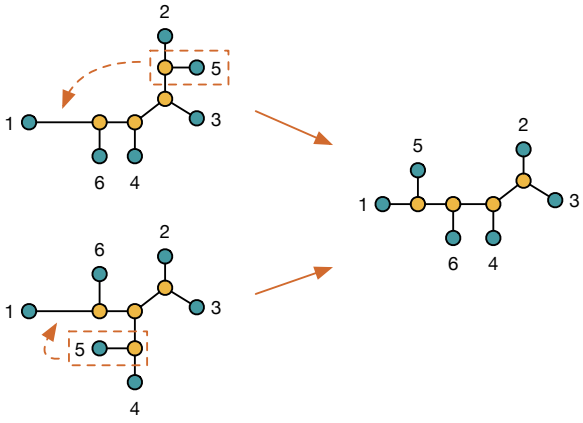


Fig. 5.4. Two examples where the phylogeny on the right is obtained through a *Subtree Pruning and Regrafting* (SPR) operation [Felsenstein, 2004] on either one of the phylogenies on the left.

phylogenies shown in Figure 5.4. The phylogeny on the right is an optimal solution to the BMEP and the two phylogenies on the left are not optimal. However, we can make the small modification of removing taxon 5, its adjacent internal vertex and the external edge connecting them and reinserting both vertices and the external edge on the path between taxon 1 and the internal vertex connected to it. Then, we obtain an optimal solution to the

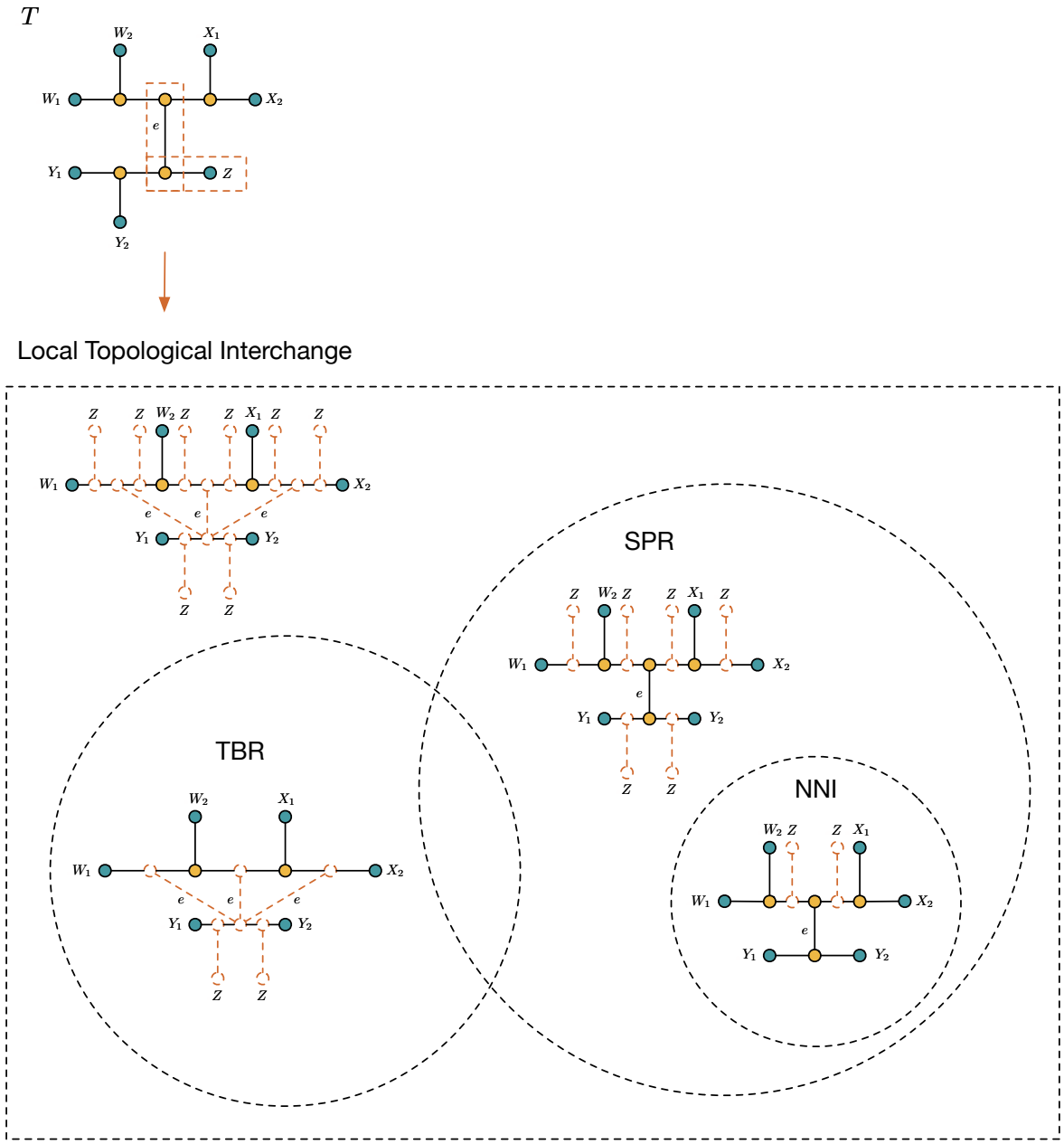


Fig. 5.5. Shows a phylogeny as a graph T in which each external vertex represents an arbitrary but fixed binary tree rooted in the vertices $W_1, W_2, X_1, X_2, Y_1, Y_2$ and Z, respectively. Moreover, all topological interchanges shown below can be obtained through the reconnection of the edge e and/or subtree Z in T. Among all possible local topological interchanges, we can distinguish between the *Nearest Neighbor Interchange* (NNI), the *Subtree Pruning and Regrafting* (SPR), and the *Tree Bisection and Reconnection* (TBR), respectively (see Section 2.5.2).

	$n = 10$				$n = 11$				$n = 12$			
Data set	Opt.	γ-Bound	LPLB	LB(α^*)	Opt.	γ-Bound	LPLB	LB(α^*)	Opt.	γ-Bound	LPLB	LB(α^*)
Primates12	124.9	117.5	**121.7**	109.7	332.8	308.5	**324.2**	299.9	802.5	740.1	**782.7**	745.3
M17	105.1	100.1	104	**104.5**	261.2	246.4	258.1	**260.4**	541.6	508.4	**535**	526.5
M18	190.3	164.7	**186.2**	177.5	396.9	330.5	**386.7**	349.4	805.9	667.7	**784**	695.6
SeedPlant25	91.4	77.4	**86.6**	68	209.8	173.2	**195.9**	141.9	427.8	356.9	**407**	294.1
M43	105	99	**103.6**	97.2	209.3	196.5	**204.2**	193.7	434.3	401	**426.5**	401.2
RbcL55	152.4	140.7	149.7	**149.8**	328.6	303.6	**323.3**	316.1	685.9	628.7	**670.8**	653.1
M62	163.8	158.8	**162.1**	156.5	350.4	338.4	**345.1**	329	753.8	725.2	**738.2**	699.3
Rana64	41.2	**39**	38.8	37.3	87.1	81.5	**82.8**	81.6	183	166.2	**174.5**	171.2
M82	53.1	45.6	**50.8**	50.4	106.4	85.7	**101.6**	101.3	225.8	181.8	215.8	**216.7**

Table 5.1. Different lower bounds on the optimal solution to the BMEP calculated at the root node of the SAS search of Algorithm 4 for instances built from real aligned DNA data sets Catanzaro et al. [2006b] on $n = 10, 11, 12$ taxa.

BMEP in constant time after building either one of the two phylogenies on the left. Hence, we can improve the lower bounds evaluated in the SAS by local topological interchanges of taxa, subtrees of taxa and/or internal edges on the phylogenies we construct. The classification of these operations in the literature are shown in Figure 5.5. To keep the computational cost of the local search for local topological interchanges inexpensive, i.e., keep the neighbourhood of all phylogenies that can be obtained through one topological interchange small, we restrict ourselves to NNIs.

Also consider that we have already seen in Section 3.4 that the family of bounds LB(α) can provide useful lower bounds on the optimal solution to the BMEP. Since we do not know how to maximize LB(α) with respect to α, we calculate, for a given input distance matrix $\mathbf{D}$, LB(α) for all $\alpha \in [0, 1]$ up to a numerical accuracy of 10^{-6} by using Proposition 3.4. We call the maximum among the bounds we obtain in this way LB(α^*). Preliminary experiments have shown that a higher accuracy for α does improve the lower bound LB(α^*) only very

	$n = 19$				$n = 20$			
Data set	Opt.	γ-Bound	LPLB	LB(α^*)	Opt.	γ-Bound	LPLB	LB(α^*)
SeedPlant25	75,261	52,409	**70,014**	43,409	164,507	109,511	**149,517**	81,056
M43	71,769	63,521	**70,743**	67,951	157,472	138,290	**155,142**	146,920
RbcL55	114,623	88,318	**111,223**	106,122	236,424	171,925	**225,473**	214,269
M62	145,642	136,472	**142,447**	132,464	298,561	277,545	**290,179**	273,866
Rana64	36,052	30,943	**34,437**	32,087	81,105	69,694	**77,259**	73,619
M82	43,997	30,240	**41,198**	15,916	89,431	59,127	**82,092**	27,599

Table 5.2. Different lower bounds on the optimal solution to the BMEP calculated at the root node of the SAS search of Algorithm 4 for instances built from real aligned DNA data sets Catanzaro et al. [2006b] on $n = 19, 20$ taxa.

marginally while requiring a lot more computation time. To assess the quality of bound $\mathrm{LB}(\alpha^*)$ we compare it to the optimal solution of the LPLB at the root node of our SAS search (Algorithm 4). Moreover, we consider a strengthened version of lower bound (5.17). To this end, we define, for a sub-phylogeny T_S,

$$\gamma_t^{ij} = \min_{\substack{i,j \in S \\ i<j<t}} \frac{1}{2} \left(d_{it} + d_{jt} - d_{ij} \right)$$

and denote $(\gamma_t^{(1)}, \gamma_t^{(2)}, \dots)$ as the vector obtained by sorting $(\gamma_t^{ij})_{i<j<t}$ in ascending order. Then, we call

$$L(T_S) + \sum_{t \in \Gamma \setminus S} \left(\sum_{i=1}^{t-3} 2^{-i} \gamma_t^{(i)} + 2^{3-t} \gamma_t^{(t-2)} \right) \tag{5.24}$$

the γ-*Bound* Pardi [2009]. Tables 5.1 and 5.2 show the results of this experiment for $n = 10, 11, 12, 19, 20$ for prominent instances used in the next section. We can see that overall the new bounds $\mathrm{LB}(\alpha^*)$ do not provide an improvement compared to previously known lower bounds and do not show clear patterns, e.g., $\mathrm{LB}(\alpha^*)$ is the best considered lower bound for instance M82 on 12 taxa, but by a large margin the worst bound on 20 taxa for the same instance.

Outside of a Branch-&-Bound-framework, both the SAS and the agglomerative approach can be used as a basis for the greedy method to approximate the BMEP, too. For the SAS, this means we evaluate consecutive branchings of the initial sub-phylogeny on three taxa by only continuing our search on branchings T which are minimal with respect to a selection criterion $C(T)$ (see Algorithm 6). For the agglomerative approach, we specify the greedy method by starting with the star graph and inserting new internal edges (u_1, u_2), replacing internal vertices u. Such an internal vertex u is determined by the minimization of a selection criterion on all bipartitions $P = (N_1, N_2)$ of the neighbourhood of u. Then, the neighbourhood of u_1 and u_2 are set to be $N_1 \cup \{u_2\}$ and $N_2 \cup \{u_1\}$, respectively. We continue this process of internal edge insertions until we obtain a phylogeny, i.e., until we have agglomerated $(2n-3)$

Algorithm 6: Stepwise addition strategy - Greedy search

Input: Ordered set of taxa $\Gamma = \{1, \dots, n\}$, distance matrix D for Γ
Output: A phylogeny of Γ
$t \leftarrow 3$;
$T \leftarrow$ The only sub-phylogeny of Γ for $\{1, 2, 3\}$;
while $t < n$ **do**
  $t \leftarrow t + 1$;
  $\hat{e} \leftarrow \arg\min\{C(T \oplus_e t) : e \in E(T)\}$;
  $T \leftarrow T \oplus_{\hat{e}} t$;
return T;

Algorithm 7: Agglomerative approach - Greedy search

Input: Set of taxa $\Gamma = \{t_1, \ldots, t_q\}$, distance matrix D for Γ,
Output: A phylogeny of Γ
$V \leftarrow \Gamma \cup \{v_1\}$;
$E \leftarrow \{\{t_i, v_1\} : t_i \in \Gamma\}$;
while $|E| < (2n-3)$ **do**
  $u \leftarrow$ A vertex of degree at least four in (V, E) that minimizes a selection criterion $C(P)$ for a bipartition $P = (N_1, N_2)$ of its neighbours;
  $V \leftarrow V \setminus \{u\} \cup \{u_1, u_2\}$;
  $E \leftarrow E \setminus \{(u, v) : (u, v) \in E\} \cup \{(u_1, v) : v \in N_1\} \cup \{(u_2, v) : v \in N_2\} \cup \{\{u_1, u_2\}\}$;
return (V, E);

edges (see Algorithm 7). We have shown how both approaches operate in Figures 5.1 and 5.3 and now added a greedy selection criterion as a specification of each step for both strategies. Thus, Algorithms 6 and 7 are correct greedy algorithms for the BMEP. To specify these algorithms we choose greedy selection criteria with a low computational cost that yield good upper bounds on the optimal solution in our following experiments. For Algorithm 6, we consider

$$C(T) = L(T) + \sum_{t \in \Gamma \setminus \text{taxa}(T)} \min_{\substack{i,j \in \text{taxa}(T) \\ i<j<t}} \frac{1}{2} \left(d_{it} + d_{jt} - d_{ij} \right),$$

the lower bound on the optimal solution that we used in the example shown in Figure 5.2, which was introduced as a selection criterion for the greedy method by Pardi [2009]. For Algorithm 7, we pick u according to the bipartition $P = (N_1, N_2)$ with $|N_1| > |N_2| = 2$ which minimizes the change in Semple and Steel [2004]'s extension of length function (2.1) to unrooted non-binary trees. This method is known in the literature as the *Neighbour-Joining algorithm* (NJ) (see Section 2.5.2)

In the next section we will use all the techniques presented so far to improve upon Catanzaro et al. [2012]'s exact solution algorithm for the BMEP.

5.2.3 Experiments

We assemble all algorithms discussed in this chapter to solve the BMEP in parallel and establish new benchmarks for the solution time of the BMEP on the same instances used by Catanzaro et al. [2012] in their experiments. Recall that these instances are built from real aligned DNA data sets Catanzaro et al. [2006b]. To this end, we call the parallelized SAS consisting of Algorithm 4 and 5 where

(A) the order of taxa Γ is identical with the order of taxa in the distance matrix D,

(B) the edges F are ordered by the time at which each edge was created during the execution of current SAS,

(C) the jobs are added and removed from the queue Q in a first-in-first-out order,

(D) parameters p and c are set to 2 and 1, respectively, and subsets F_1 and F_2 contain the first $\lfloor |F|/2 \rfloor$ edges of F and last $\lceil |F|/2 \rceil$ edges of F, respectively,

the *Parallel Stepwise Addition Strategy With Default Configuration* (PSAS-C1111). Furthermore, we preprocess Algorithms 6 and 7 as specified in the last section to calculate two upper bounds on the optimal solution to the BMEP and store the better one before starting the PSAS. Moreover, we apply all possible NNI operations to the phylogeny underlying the stored upper bound and every newly stored phylogeny T^* during the execution of PSAS to potentially improve them. We refer to this extension of the PSAS-C1111 as EPSAS-C1111. Next, we consider alternative configurations by considering variations of specifications (A) to (C). This means,

(A) the order of taxa Γ is identical with the order of taxa

(A.1) in the distance matrix D,

(A.2) in the optimal solution to the Traveling Salesman Problem for distance matrix D,

(A.3) obtained by shortcutting the tree returned by Algorithm 6,

(A.4) obtained by shortcutting the tree returned by Algorithm 7;

(B) the edges F are ordered

(B.1) by the time at which each edge was created during the execution of the current SAS,

(B.2) non-decreasing in the pre-calculated values lowerbound(T) one would obtain in the next recursion step by branching on the corresponding edge;

(C) the jobs are added and removed from the queue Q

(C.1) in a first-in-first-out order,

(C.2) in a first-in-last-out order,

(C.3) in non-increasing order of the number of edges contained in the underlying sub-phylogeny,

(C.4) in non-decreasing order of the number of edges contained in the underlying sub-phylogeny.

When referring to one fixed combination of specifications to modify the implementation of algorithm EPSAS-C1111 we change the suffix C1111 accordingly, e.g., C4231 means we chose (A.4), (B.2) and (C3) instead of (A.1), (B.1) and (C.1), respectively.

To test the performance of algorithms EPSAS-C1111 to EPSAS-C4241, we used the same data sets as Catanzaro et al. [2012] and set one hour as the maximum time limit. All numerical experiments were performed on the Cesam server provided by the supercomputing facilities of the Université catholique de Louvain which is equipped with a 2x8 Core E5-2667v3 Processor at 3.2 GHz and 256GB RAM. The source code is written in C++ and compiled with the GCC compiler version 4.8.5.

The results of the first series of tests are shown in Figures 5.6 to 5.10. First, we observe that the choice for the order of jobs in the queue (C) has an impact of multiple magnitudes smaller than the choice for taxa order (A) and edge order (B) and that no queue order yields a statistical significant advantage compared to all others.

Next, we can see that edge order (B.1) has a positive impact on the overall runtime of all tested configurations compared to (B.2). Specifically, the worst case and 75% quartile runtimes improve drastically for increasing n. This leads to a twofold conclusion. On the one hand, the additional preprocessing overhead for the calculation of lower bounds required in each iteration to determine edge order (B.2) affects the runtime for small n negatively. On the other hand, the calculation of lower bounds throughout the tested algorithms dominates all other contributions to the total computation time. This means, preprocessing lower bounds for future iterations does not scale well with a higher number of cores and increasing n because at the time a pre-calculated lower bound is employed by one core the corresponding node of the search-tree might already be pruned by a different core.

Concerning the taxa order, the test results are less clear. The best worst-case performance is always reached by taxa order (A.4), but the lowest best-case and 25% quartile runtime is achieved by taxa orders different from (A.4) for most n. However, since taxa order (A.4) performs statistically significant better than all other configurations for $n = 20$ we keep this order as the best choice for further experiments.

Thus, overall we conclude that configurations C411 to C414 perform the best. Now, we consider further configurations extending specification (D) with

$$\text{(D.1) } p = 2, c = 1, \ \text{(D.2) } p = 3, c = 1, \ldots, \ \text{(D.8) } p = 9, c = 1.$$

Then, Figures 5.11 to 5.13 show the computational results for configurations C4111 to C4148. Here we can observe that no choice of parameters for specification (D) yields a statistical significant advantage compared to all other evaluated choices. However, for sufficiently large n,

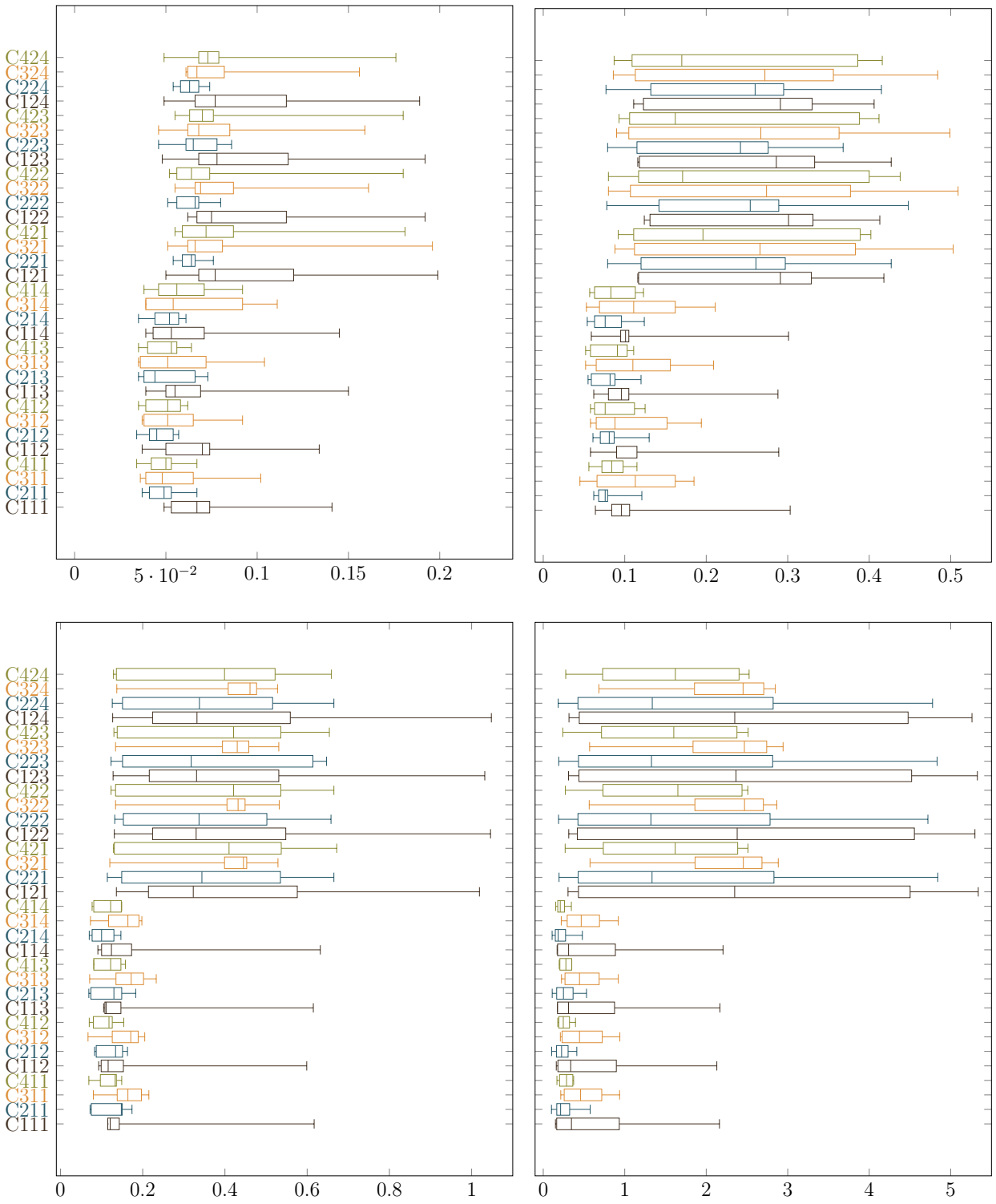


Fig. 5.6. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C1111 to EPSAS-C4241 for $n = 10$ (upper left), $n = 11$ (upper right), $n = 12$ (lower left) and $n = 13$ (lower right).

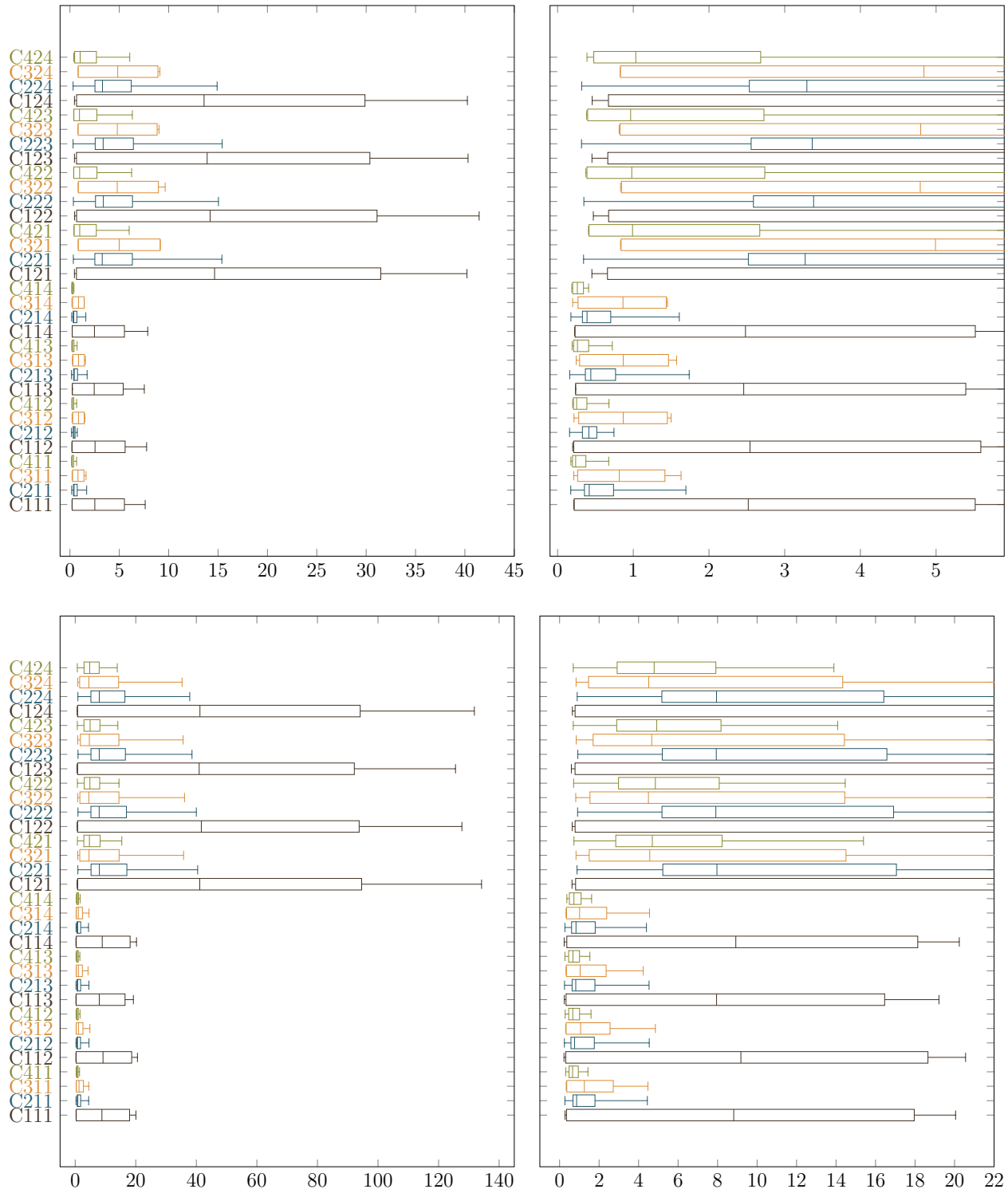


Fig. 5.7. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C1111 to EPSAS-C4241 for $n = 14$ (top) and $n = 15$ (bottom).

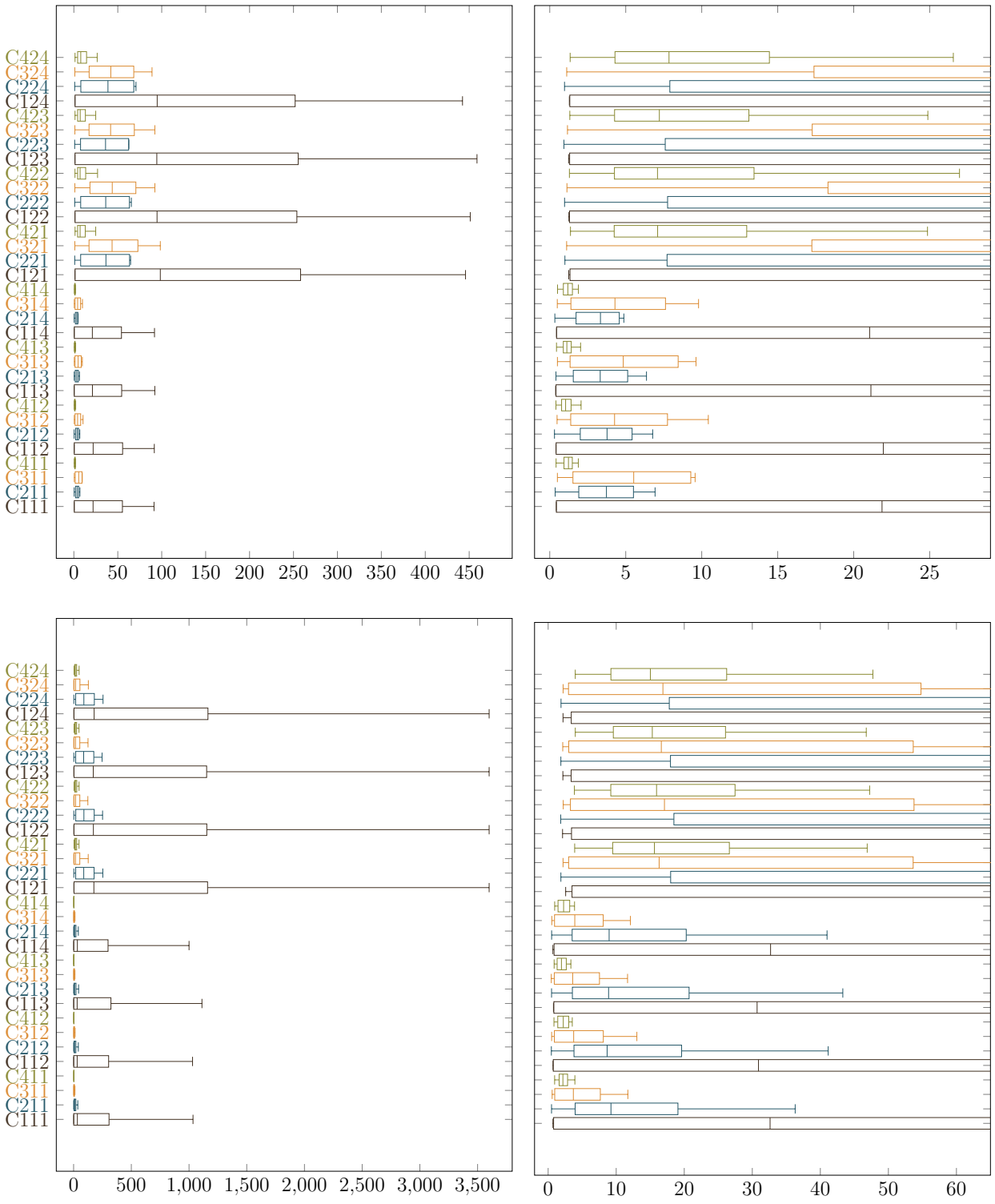


Fig. 5.8. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C1111 to EPSAS-C4241 for $n = 16$ (top) and $n = 17$ (bottom).

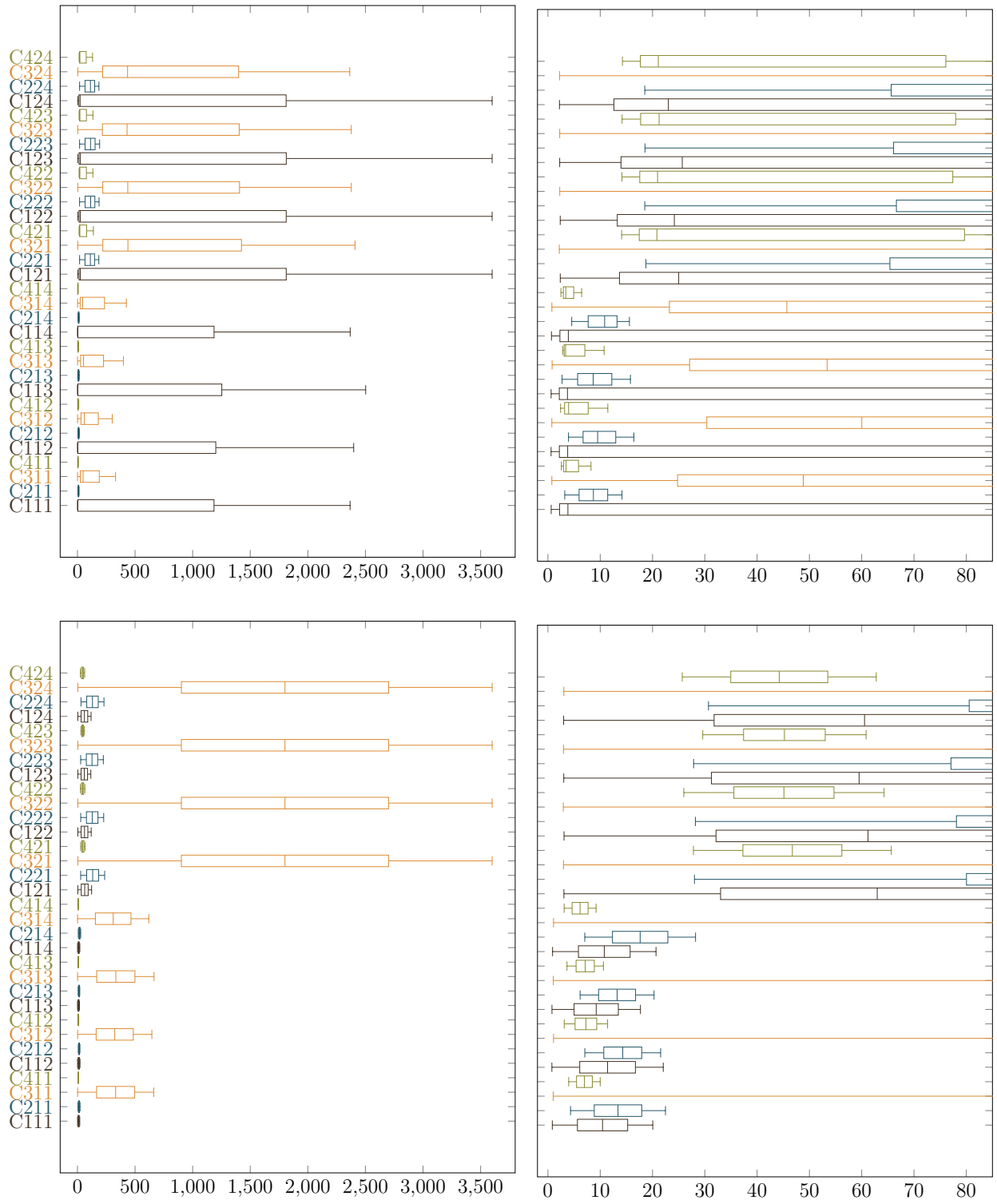


Fig. 5.9. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C1111 to EPSAS-C4241 for $n = 18$ (top) and $n = 19$ (bottom).

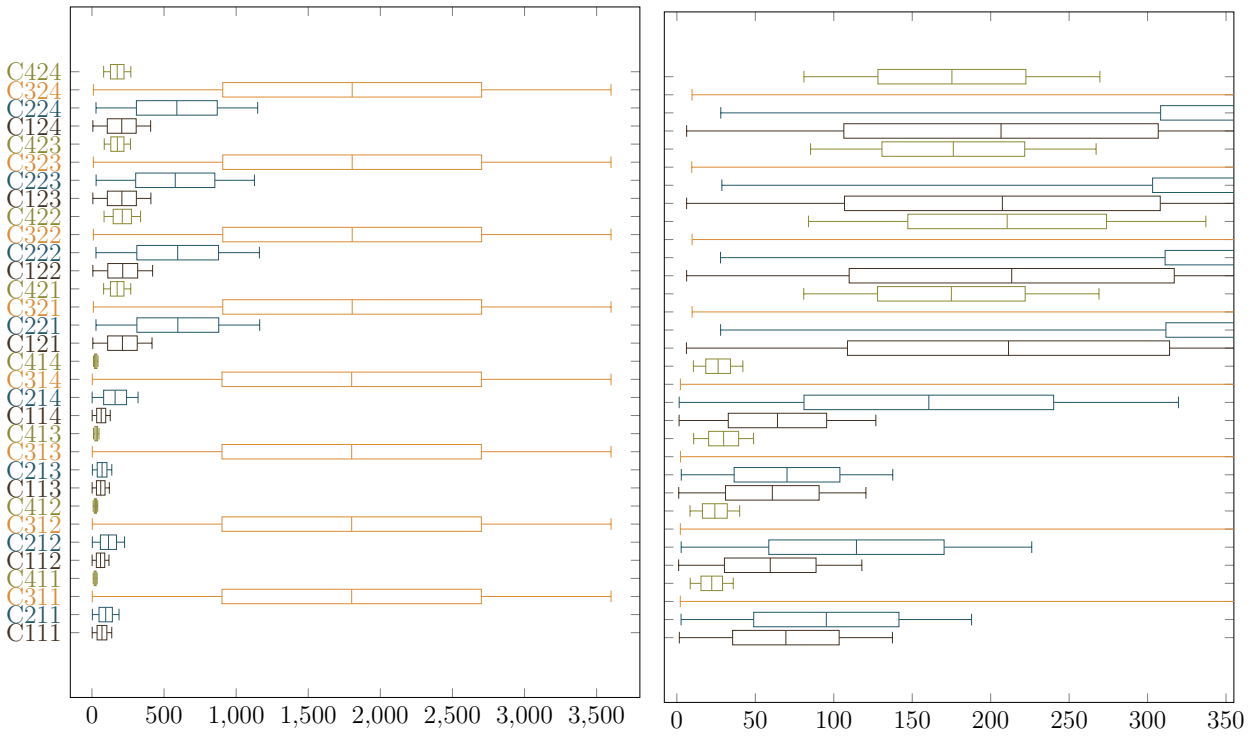


Fig. 5.10. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C1111 to EPSAS-C4241 for $n = 20$.

increasing the number of partitions sets p seems to slow down the execution of algorithm EPSAS. This observation matches our previous assessment that smaller partition sets lead to an overabundance of jobs in the queue. This increase in computation time for managing the distribution of jobs can not be compensated by the decrease in idle time for each individual core. Hence, we only consider $p = 2$ and $p = 3$ as part of the best performing configurations in the worst case and extend specification (D) further as follows:

$$\text{(D.9) } p = 2, c = 2, \text{ (D.10) } p = 2, c = 3, \text{ (D.11) } p = 2, c = 4,$$
$$\text{(D.12) } p = 3, c = 2, \text{(D.13) } p = 3, c = 3, \text{ (D.14) } p = 3, c = 4.$$

Then, Figures 5.14 to 5.16 show the computational results for configurations C4119 to C41414 and C4112, C4113 to C4142, C4143, respectively. Again, we observe that no specification of (D) is significantly better compared to all tested configurations. Moreover, executions of EPSAS with smaller choices for the constant c remain among the best performances for increasing n. This strengthens our suspicion that the best strategy to partition the set of jobs of a sub-phylogeny can not be defined a-priori and requires the employment of more adaptive methods.

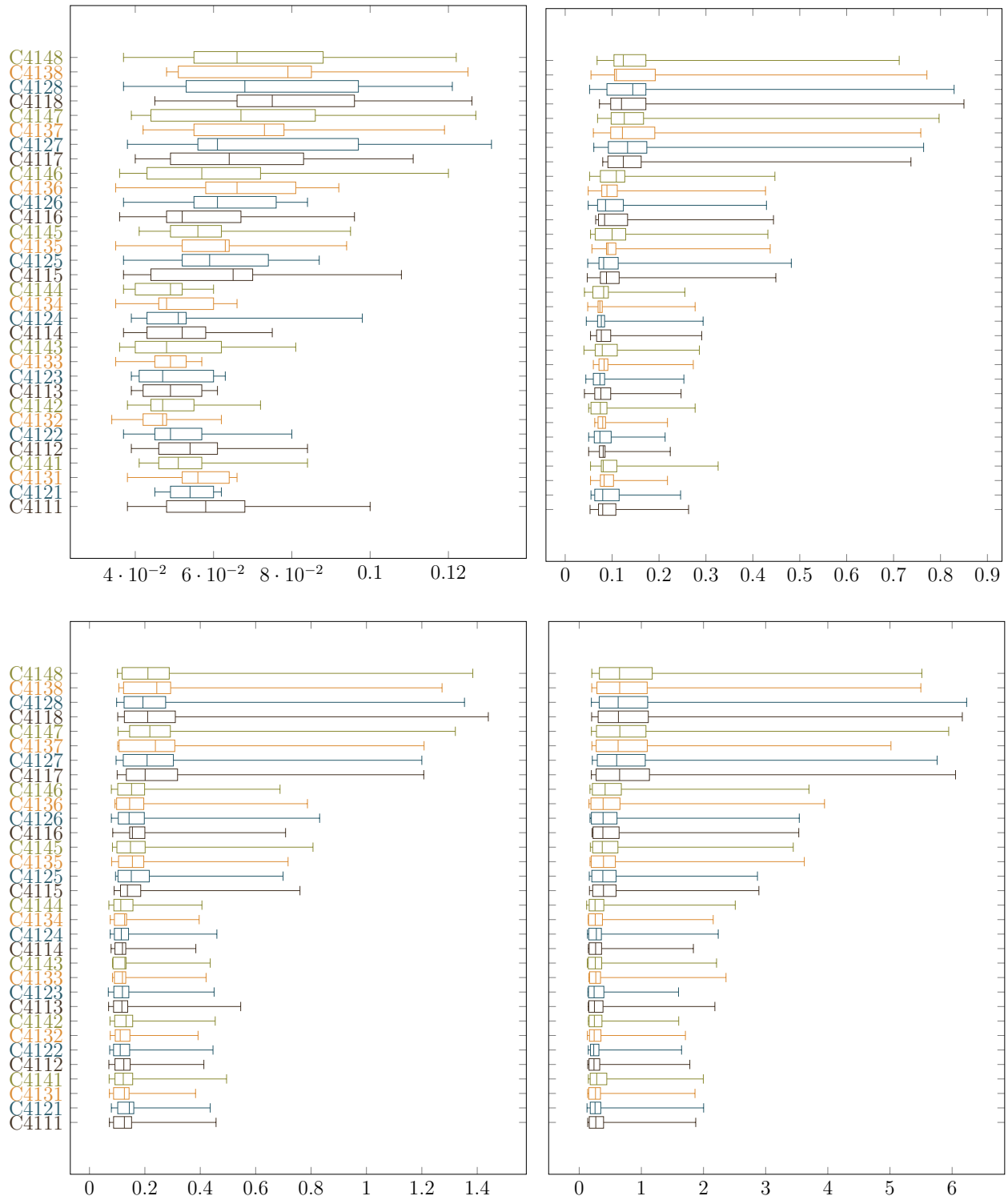


Fig. 5.11. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4111 to EPSAS-C4148 for $n = 10$ (upper left), $n = 11$ (upper right), $n = 12$ (lower left) and $n = 13$ (lower right).

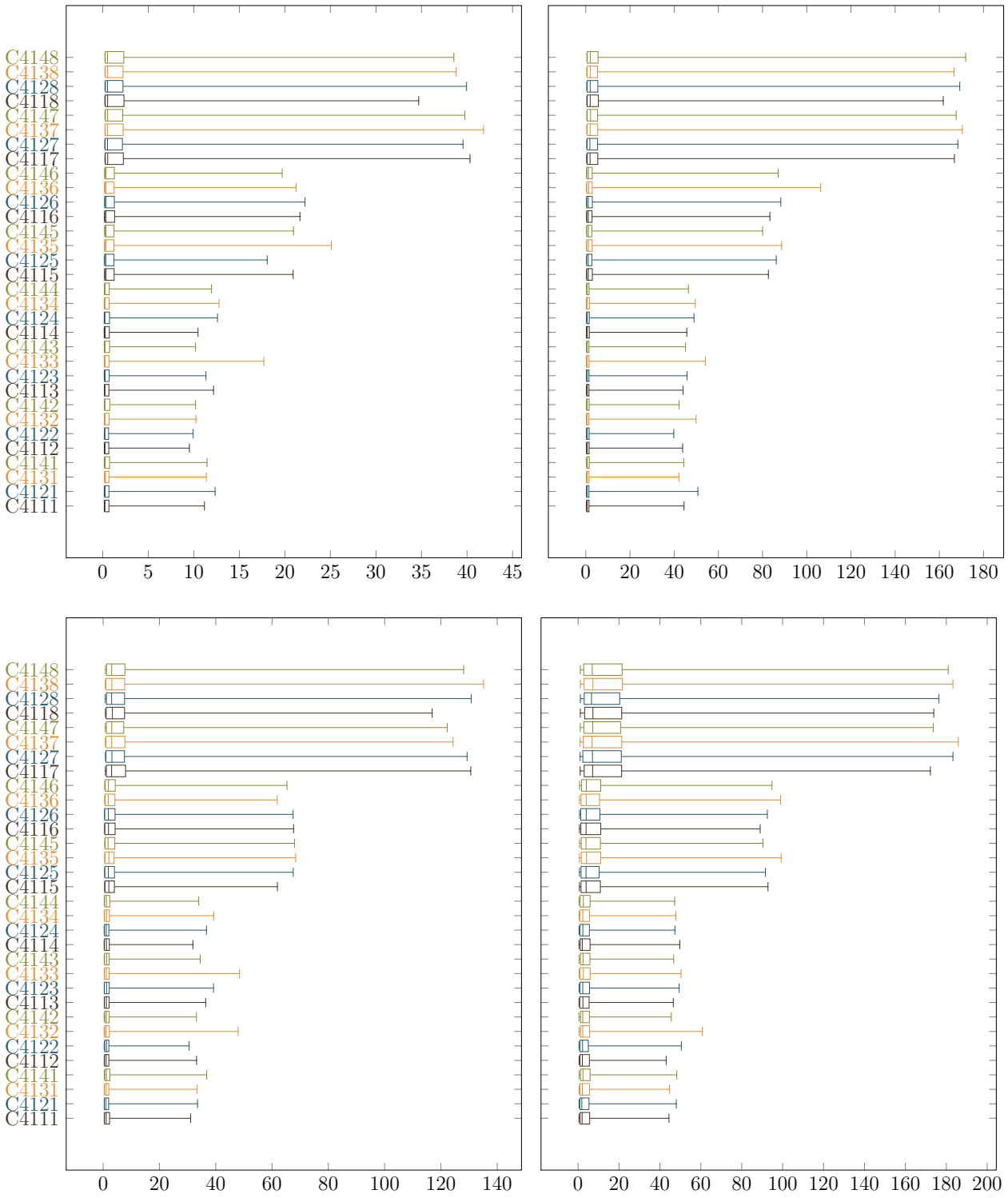


Fig. 5.12. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4111 to EPSAS-C4148 for $n = 14$ (upper left), $n = 15$ (upper right), $n = 16$ (lower left) and $n = 17$ (lower right).

Fig. 5.13. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4111 to EPSAS-C4148 for $n = 18$ (upper left), $n = 19$ (upper right) and $n = 20$ (bottom).

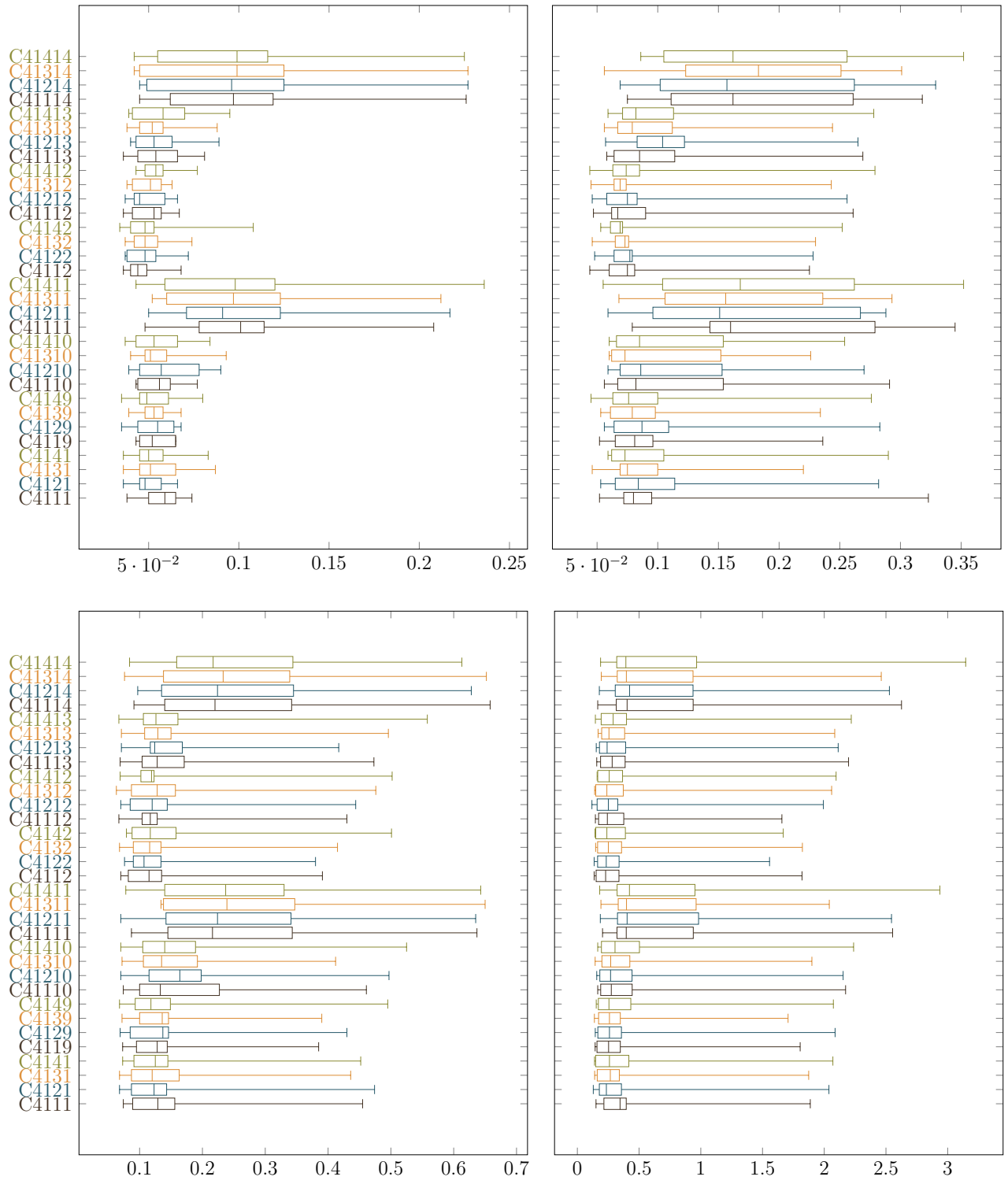


Fig. 5.14. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4119 to EPSAS-C41416 for $n = 10$ (upper left), $n = 11$ (upper right), $n = 12$ (lower left) and $n = 13$ (lower right).

Fig. 5.15. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4119 to EPSAS-C41416 for $n = 14$ (upper left), $n = 15$ (upper right), $n = 16$ (lower left) and $n = 17$ (lower right).

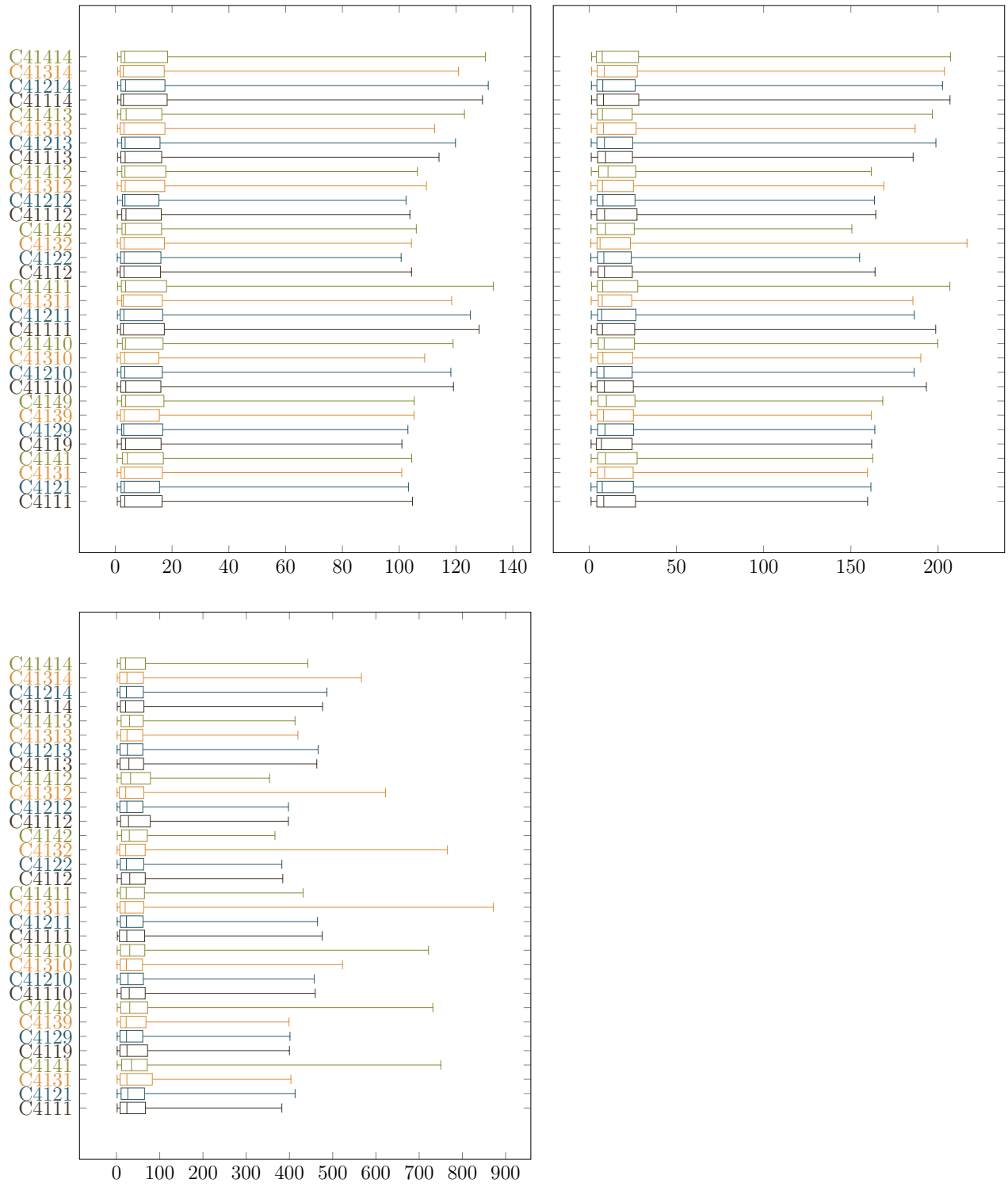


Fig. 5.16. A boxplot of the runtime in seconds for the 32 algorithms EPSAS-C4119 to EPSAS-C41416 for $n = 18$ (upper left), $n = 19$ (upper right) and $n = 20$ (bottom).

To conclude our experiments for the instances evaluated so far, we establish new benchmarks for the solution time of the BMEP by using algorithm EPSAS-C41412, which offers the best worst case performance among all analysed configurations of EPSAS for $n = 20$. Recall that this means we implement algorithm EPSAS such that

(A) the order of taxa Γ is obtained by shortcutting the tree returned by Algorithm 7,

(B) the edges F are ordered by the time at which each edge was created during the execution of current SAS,

(C) the jobs are added and removed from the queue Q in non-decreasing order of the number of edges contained in the underlying sub-phylogeny,

(D) parameters p and c are set to 3 and 2, respectively, and subsets F_1 and F_2 contain the first $\lfloor |F|/2 \rfloor$ edges of F and last $\lceil |F|/2 \rceil$ edges of F, respectively,

Tables 5.3 and 5.4 summarize the numerical results with respect to the solution time and the number of branchings needed in the EPSAS-C41412 to solve a specific instance. Table 5.5 completes Tables 5.3 and 5.4 by listing all the instances that can be solved by the multi core solver but not the single core version. In general we see that the single core EPSAS-C41412 as well as the multi core version solve the BMEP faster than Catanzaro et al. [2012]'s state-of-the-art solution algorithm. Moreover, the tables show that using a single core is only beneficial when dealing with very small sets of taxa. This is explained by the fact that the computational overhead of managing multiple cores for EPSAS-C41412 on a few taxa is more expensive than the potential for improvements in the solution time through parallelization. Furthermore, we observe that the multi core EPSAS-C41412 is twice as fast as the single core version 99% of the time. The relative improvements when comparing both versions of the EPSAS-C41412 also show that the parallel version is at least one order of magnitude faster than the single core version 28% of the time. This is particularly interesting for instances like *M82* on 20 taxa for which the BMEP was not solvable before in under an hour. Thus, EPSAS-C41412 increases the number of taxa for which the BMEP can be solved in a reasonable amount of time. This raises the question how much farther faster algorithms than the EPSAS-C41412 could go before meeting the numerical limits of the BMEP. We discuss this issue and a possible resolution in the next section.

5.3 On the numerical stability of the BMEP

The objective function of the BMEP is subjected to numerical instabilities whenever the generic addend $d_{ij}2^{-\tau_{ij}}$ approaches (or get smaller than) the machine epsilon ϵ, i.e., if there exist at least a pair of distinct taxa $i, j \in \Gamma$, such that

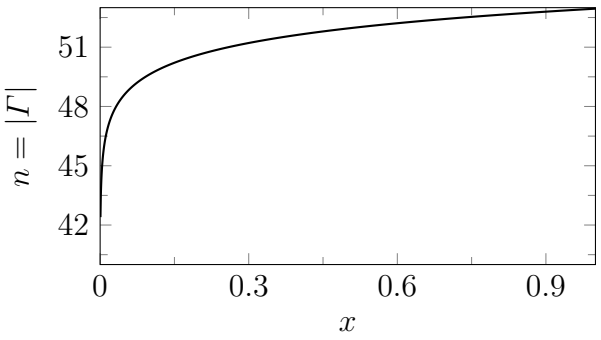


Fig. 5.17. Behavior of the function $\log_2\left(\frac{x}{\epsilon}\right) + 1$ for $0 < x \leq 1$.

$$d_{ij} \cdot 2^{-\tau_{ij}} \leq \epsilon. \tag{5.25}$$

This situation limits the analysis of practical instances involving hundreds or thousands taxa (see, e.g., the phylogenetic analysis of the datasets related to SARS-Cov2 in Chapter 2) and may occur because of specific values of either the path-length τ_{ij} or the distance d_{ij} (or both). In order to characterize the impact of path-lengths on the numerical stability of the BMEP, we start by recalling that in any feasible solution to the problem the value of each τ_{ij} may vary within the discrete interval $[2, \ldots, n-1]$ (see Section 2.4.3). Then, it is licit to wonder which is the smallest number n of taxa for which numerical instabilities occur when solving an instance of the BMEP. A trivial algebraic manipulation of (5.25) provides the following answer: numerical instabilities occur whenever

$$n \geq \left\lfloor \log_2\left(\frac{d_{ij}}{\epsilon}\right) + 1 \right\rfloor. \tag{5.26}$$

Now, in 64-bit floating point arithmetic, $\epsilon \approx 2.3 \cdot 10^{-16}$; then, if the smallest non-diagonal entry of $\mathbf{D}$ is at most 1 (e.g. for our previously cited instance "Primates12" on 12 taxa) the BMEP becomes numerically instable for

$$n \geq \left\lfloor \log_2\left(\frac{1}{2.3 \cdot 10^{-16}}\right) + 1 \right\rfloor = 53. \tag{5.27}$$

As shown in Figure 5.17, this bound can become even smaller in function of the smallest non-diagonal entry of $\mathbf{D}$.

To overcome numerical instabilities it would be desirable to pull away as far as possible the bound on n for which numerical stability issues arise. One way to achieve this goal consists of finding a scalar $\alpha > 1$ such that

$$n \geq \left\lfloor \alpha \cdot \log_2\left(\frac{d_{ij}}{\epsilon}\right) + 1 \right\rfloor \tag{5.28}$$

may be as large as possible. Imposing (5.28) means considering the following new length function for the BMEP:

$$L_\alpha(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{d_{ij}}{2^{\tau_{ij}/\alpha}}. \tag{5.29}$$

We observe that any rescaling of the length function (2.1) can be written in the form of (5.29) and we refer to the problem of minimizing (5.29) as the *Rescaled Balanced Minimum Evolution Problem* (RBMEP). An interesting property of (5.29) is that in some circumstances it allows to characterize the optimal solution to the rescaled problem, as illustrated by the following proposition:

Proposition 5.2. *For a sufficiently large scalar $\alpha > 1$, there exists an input distance matrix $\mathbf{D}$ for which the caterpillar phylogeny is an optimal solution to the RBMEP.*

Proof. Assume $n = 11$ and first observe that

$$\lim_{\alpha \to \infty} L_\alpha(T) = \frac{n(n-1)}{2} - \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij} d_{ij}.$$

Then, for sufficiently large α, the phylogeny T that maximizes

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij} d_{ij}$$

is optimal for the RBMEP. Choose a distance matrix $\mathbf{D}$ with entries defined by a positive constant. Then, the phylogeny T with maximal

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij}$$

is optimal to the RBMEP. Since we assumed $n = 11$, the maximum

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij} = 310$$

is achieved by the caterpillar phylogeny.

Proposition 5.2 shows us that there exist scalars $\alpha > 1$ for which the optimal solution to the RBMEP can be identified independently of a given input distance matrix $\mathbf{D}$. This is not the case for the BMEP. Hence, a licit question is whether the new length function (5.29) allows to preserve the statistical consistency of the optimal solution to the RBMEP for sufficiently small $\alpha > 1$. Unfortunately, the next propositions provide a negative answer.

Proposition 5.3. *For any scalar $\alpha > 1$, $L_\alpha(T)$ overestimates the length of the true phylogeny of the set Γ of taxa encoded by the input distance matrix $\mathbf{D}$.*

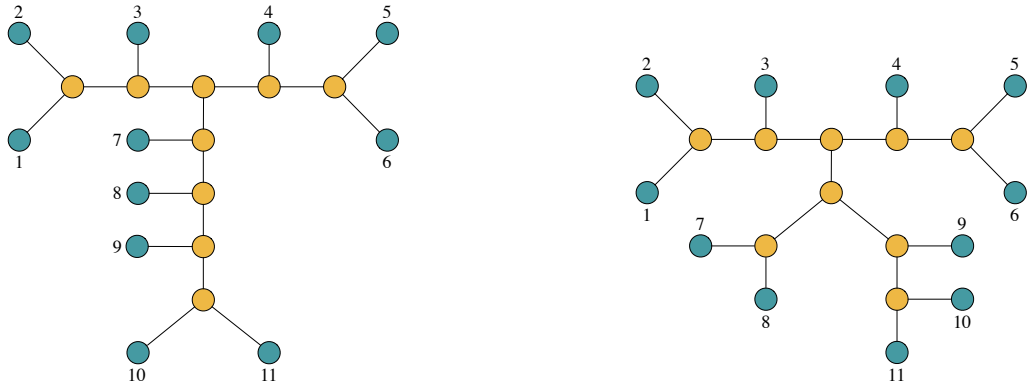


Fig. 5.18. Two different phylogenies T (left) and T' (right) of 11 taxa.

Proof. As shown in Desper and Gascuel [2004], $L(T)$ is a consistent estimation of the length of the true phylogeny associated to $\mathbf{D}$. Then, the statement follows by observing that $2^{-\tau_{ij}/\alpha} > 2^{-\tau_{ij}}$ holds true for any scalar $\alpha > 1$, hence $L_\alpha(T) > L(T)$, for any phylogeny T of Γ.

Proposition 5.4. *There exists an input distance matrix* $\mathbf{D}$ *for which the RBMEP is not statistically consistent, for any* $\alpha \geq 1.2447$.

Proof. Assume $n = 11$ and consider the phylogenies T and T', shown in Figure 5.18, and the associated path-length matrices τ and τ', respectively, shown in Figure 5.19. Set $D = \tau'$. Then, it is possible to see numerically that $L_\alpha(T) < L_\alpha(T')$ for any $\alpha \geq 1.2447$. Now, D corresponds precisely to T', hence T' is the true phylogeny for this specific instance of the RBMEP. Thus, for any $\alpha \geq 1.2447$, the problem of minimizing $L_\alpha(T)$ over the set of the phylogenies for Γ does not yield the desired true phylogeny.

$$\tau = \begin{bmatrix}
0 & 2 & 3 & 5 & 6 & 6 & 5 & 6 & 7 & 8 & 8 \\
2 & 0 & 3 & 5 & 6 & 6 & 5 & 6 & 7 & 8 & 8 \\
3 & 3 & 0 & 4 & 5 & 5 & 4 & 5 & 6 & 7 & 7 \\
5 & 5 & 4 & 0 & 3 & 3 & 4 & 5 & 6 & 7 & 7 \\
6 & 6 & 5 & 3 & 0 & 2 & 5 & 6 & 7 & 8 & 8 \\
6 & 6 & 5 & 4 & 2 & 0 & 5 & 6 & 7 & 8 & 8 \\
5 & 5 & 4 & 4 & 5 & 5 & 0 & 3 & 4 & 5 & 5 \\
6 & 6 & 5 & 5 & 6 & 6 & 3 & 0 & 3 & 4 & 4 \\
7 & 7 & 6 & 6 & 7 & 7 & 4 & 3 & 0 & 3 & 3 \\
8 & 8 & 7 & 7 & 8 & 8 & 5 & 4 & 3 & 0 & 2 \\
8 & 8 & 7 & 7 & 8 & 8 & 5 & 4 & 3 & 2 & 0
\end{bmatrix} \qquad \tau' = \begin{bmatrix}
0 & 2 & 3 & 5 & 6 & 6 & 6 & 6 & 6 & 7 & 7 \\
2 & 0 & 3 & 5 & 6 & 6 & 6 & 6 & 6 & 7 & 7 \\
3 & 3 & 0 & 4 & 5 & 5 & 5 & 5 & 5 & 6 & 6 \\
5 & 5 & 4 & 0 & 3 & 3 & 5 & 5 & 5 & 6 & 6 \\
6 & 6 & 5 & 3 & 0 & 2 & 6 & 6 & 6 & 7 & 7 \\
6 & 6 & 5 & 3 & 2 & 0 & 6 & 6 & 6 & 7 & 7 \\
6 & 6 & 5 & 5 & 6 & 6 & 0 & 2 & 4 & 5 & 5 \\
6 & 6 & 5 & 5 & 6 & 6 & 2 & 0 & 4 & 5 & 5 \\
6 & 6 & 5 & 5 & 6 & 6 & 4 & 4 & 0 & 3 & 3 \\
7 & 7 & 6 & 6 & 7 & 7 & 5 & 5 & 3 & 0 & 2 \\
7 & 7 & 6 & 6 & 7 & 7 & 5 & 5 & 3 & 2 & 0
\end{bmatrix}$$

Fig. 5.19. Two path-length matrices τ and τ' encoding the two distinct phylogenies T and T' of Figure 5.18, respectively.

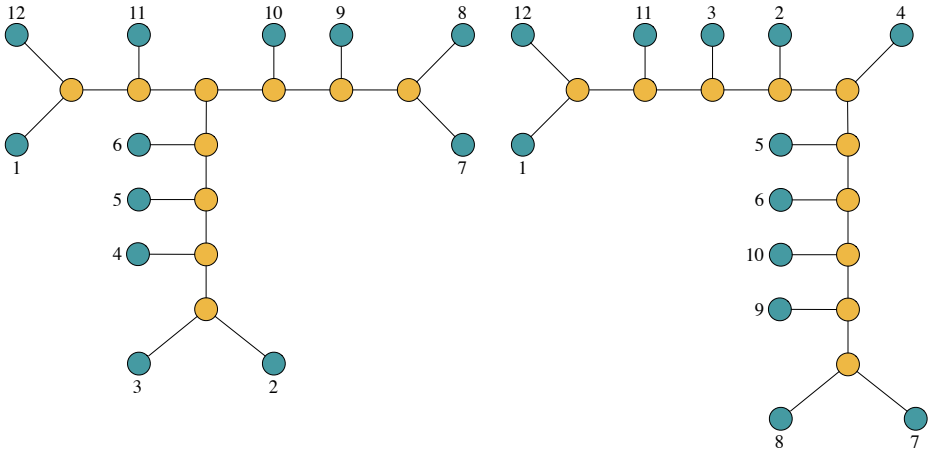


Fig. 5.20. The optimal solutions T (left) and T' (right) to the BMEP and the RBMEP for $\alpha = 2$, respectively, for the instance "Primates12" of 12 taxa.

Proposition 5.4 still leaves some hope to find a scalar $1 < \alpha < 1.2447$ for which the RBMEP could be statistically consistent. However, checking whether this hope is well-grounded is of little interest in practice mainly because, in the best case, this α (if it existed) would pull away the arising numerical instability up to

$$n \geq \left\lfloor 1.2446 \cdot \log_2 \left(\frac{1}{2.3 \cdot 10^{-16}} \right) + 1 \right\rfloor = 65, \tag{5.30}$$

a value that is not suited for large scale phylogenetic analyses involving hundreds or thousands taxa.

As a striking example of the implications of the above propositions, consider again the input distance matrix "Primates12". The optimal solutions to the BMEP and the RBMEP for $\alpha = 2$ are shown as T and T' in Figure 5.20, respectively, and are unfortunately different.

This means, there exists, depending on the input distance matrix $\mathbf{D}$, an upper bound on the number of taxa any exact solution algorithm for the BMEP can consistently solve. We have seen in Section 3.5 that the BMEP uses maximal information about the topology of a phylogeny among all phylogenetic estimation problems under minimum evolution. This poses the question how the information aspect of the BMEP and its numerical stability are connected. Indeed, does there exist a phylogenetic estimation problem which has a better trade-off between statistical inconsistencies and numerical instabilities than the BMEP? This question opens up another research direction to improve our understanding of statistical consistency in computational phylogenetics.

				Single Core		Multi Core		
Data set	Taxa	Optimum	Gap (%)	Time (s)	Branchings	Time (s)	Branchings	Improvement (%)
Primates12	10	124.968	2.57	**0.030**	216	0.800	216	-0.77
	11	332.834	2.59	0.086	344	**0.075**	344	0.01
	12	802.589	2.47	0.160	579	**0.074**	579	0.08
M17	10	105.134	1.05	0.202	1454	**0.071**	1,709	0.13
	11	261.233	1.18	0.493	2,796	**0.085**	2,779	0.41
	12	541.632	1.21	0.640	4,749	**0.137**	5,222	0.50
	13	1,181.598	1.65	3.121	18,553	**0.317**	20,685	2.80
	14	2,408.066	1.65	0.504	2,917	**0.150**	4,813	0.35
	15	4,998.295	1.65	7.246	32,492	**0.845**	47,878	6.40
	16	10,225.560	1.70	11.177	43,126	**1.233**	65,529	9.94
	17	20,788.112	1.64	12.930	39,161	**1.611**	65,403	11.32
M18	10	190.331	2.11	0.147	744	**0.049**	934	0.10
	11	396.942	2.57	0.232	1,598	**0.073**	1,681	0.16
	12	805.937	2.71	0.596	3,875	**0.126**	3,845	0.47
	13	1,758.111	3.27	2.486	17,495	**0.315**	18,224	2.17
	14	3,599.678	3.62	7.397	44,057	**0.663**	48,547	6.73
	15	7,746.218	3.47	18.412	94,430	**1.462**	99,285	16.95
	16	15,973.016	3.51	24.440	122,068	**1.984**	138,598	22.45
	17	32,345.662	3.53	43.348	215,505	**3.704**	250,581	39.64
	18	66,029.112	3.55	128.097	584,269	**10.343**	691,333	117.75
SeedPlant25	10	91.458	5.51	0.065	581	**0.046**	926	0.02
	11	206.250	5.10	0.097	717	**0.065**	1,482	0.03
	12	427.816	4.87	0.109	595	**0.088**	595	0.02
	13	943.774	5.30	0.289	1,788	**0.146**	2,074	0.14
	14	1,929.219	5.06	0.483	2,378	**0.207**	3,506	0.28
	15	4,085.763	4.66	0.771	3,173	**0.352**	13,154	0.42
	16	8,353.457	5.76	1.846	8,963	**0.387**	12,092	1.46
	17	17,314.429	6.21	6.211	39,328	**0.816**	55,241	5.39
	18	36,156.527	6.92	33.931	187,267	**3.722**	295,186	30.21
	19	75,261.323	6.97	107.280	516,428	**9.758**	705,960	97.53
	20	164,507.262	9.11	391.165	1,738,599	**43.789**	3,094,726	347.38
M43	10	105.020	1.29	0.105	490	**0.048**	477	0.06
	11	209.282	2.39	0.515	4,940	**0.087**	4,943	0.43
	12	434.326	1.79	0.693	4,584	**0.128**	5,053	0.56
	13	895.424	1.74	1.021	8,062	**0.164**	8,168	0.86
	14	1,808.970	1.75	1.900	12,414	**0.268**	12,477	1.63
	15	3,965.234	1.53	4.354	20,495	**0.507**	24,909	3.85
	16	8,219.835	1.51	9.291	43,883	**0.900**	48,793	8.39
	17	16,798.759	1.82	33.213	129,312	**2.802**	154,811	30.41
	18	35,383.158	1.44	23.212	82,653	**1.940**	86,798	21.27
	19	71,769.903	1.43	40.449	141,845	**3.495**	162,376	36.95
	20	157,472.424	1.48	88.241	312,258	**7.586**	355,902	80.65

Table 5.3. Numerical results obtained for the algorithm EPSAS-C41412 on a single core and on 32 cores. The last column shows the improvement in the solution time of the multi core approach relative to the time required to terminate the EPSAS-C41412 on one core. Bold indicates the approach that performed the best. The improvement measures the difference in solution time between the single core and multi core version.

				Single Core		Multi Core		
Data set	Taxa	Optimum	Gap (%)	Time (s)	Branchings	Time (s)	Branchings	Improvement (%)
RbcL55	10	152.482	1.79	0.182	1,201	**0.070**	1,412	0.11
	11	328.645	1.61	0.286	1,960	**0.077**	1,393	0.21
	12	685.972	2.20	1.021	6,249	**0.177**	7,206	0.84
	13	1,502.870	2.33	3.408	16,105	**0.363**	20,661	3.05
	14	3,094.888	2.27	6.913	35,580	**0.730**	40,253	6.18
	15	6,448.259	2.46	17.498	66,903	**1.386**	68,164	16.11
	16	13,455.292	2.50	26.492	96,772	**2.101**	105,599	24.39
	17	27,804.246	2.94	146.577	622,052	**9.796**	649,156	136.7 8
	18	56,175.236	2.90	288.358	1,208,182	**21.176**	1,339,498	267.18
	19	114,623.865	2.97	390.419	1,482,378	**27.242**	1,557,301	363.18
	20	236,424.336	4.63	908.280	3,515,861	**74.330**	4,482,361	833.95
M62	10	163.876	1.07	0.074	235	**0.041**	297	0.03
	11	350.425	1.50	0.137	514	**0.076**	605	0.06
	12	753.821	2.06	0.197	813	**0.083**	1,131	0.11
	13	1,580.764	2.21	0.575	3,195	**0.153**	3,131	0.42
	14	3,345.564	2.35	0.423	1,975	**0.164**	3,931	0.26
	15	7,161.078	2.64	2.357	8,785	**0.352**	11,519	2.00
	16	14,980.693	2.54	4.308	16,508	**0.576**	24,084	3.73
	17	31,293.082	2.50	10.388	37,089	**1.185**	53,167	9.20
	18	66,187.974	2.14	13.251	48,506	**2.170**	114,739	11.08
	19	145,642.424	2.19	47.787	159,650	**6.259**	309,798	41.53
	20	298,561.239	2.81	65.398	220,435	**11.339**	583,732	54.06
Rana64	10	41.223	5.74	0.301	2,134	**0.084**	2,134	0.22
	11	87.162	4.94	0.463	1,841	**0.092**	1,841	0.37
	12	183.042	4.64	0.457	1,905	**0.100**	1,905	0.36
	13	382.700	5.39	0.633	2,698	**0.153**	2,698	0.48
	14	730.466	5.60	0.824	3,098	**0.209**	3,126	0.62
	15	1,603.120	6.47	1.163	3,968	**0.268**	4,134	0.89
	16	3,290.745	7.78	2.055	8,094	**0.376**	8,535	1.68
	17	6,745.447	8.01	2.457	8,871	**0.435**	9,480	2.02
	18	15,435.268	5.07	4.428	15,584	**0.667**	19,080	3.76
	19	36,052.455	4.48	7.865	27,222	**1.047**	33,430	6.82
	20	81,105.131	4.74	12.124	36,885	**1.412**	43,256	10.71
M82	10	53.170	4.39	0.373	3,939	**0.101**	4,567	0.27
	11	106.443	4.52	2.599	37,818	**0.272**	44,071	2.33
	12	225.836	4.43	5.356	54,323	**0.385**	46,455	4.97
	13	543.127	3.92	29.519	264,235	**1.627**	211,810	27.89
	14	1,238.322	4.51	231.497	2,076,753	**10.104**	1,384,383	221.39
	15	2,515.808	6.26	891.329	7,101,372	**40.006**	4,888,322	851.32
	16	5,098.459	5.13	563.475	3,276,068	**29.507**	2,710,938	533.97
	17	10,483.717	4.76	727.497	3,716,014	**41.165**	3,331,369	686.33
	18	21,508.330	5.53	1,662.013	7,903,728	**92.955**	6,915,563	1,569.06
	19	43,997.450	6.36	2,648.185	10,924,884	**152.720**	9,662,546	2,495.46
	20	89,431.696	8.21	>3,600	n.a.	**355.224**	20,547,942	3,244.78

Table 5.4. The continuation of Table 5.3.

Data set	Taxa	Optimum	Gap (%)	Time (s)	Branchings
SeedPlant25	20	149,517.567	7.08	37.86	2,638,772
	21	308,829.021	7.30	110.97	6,038,213
	22	619,759.057	7.66	184.25	9,742,797
	23	1,259,681.860	8.24	625.89	33,099,791
	24	2,565,765.151	21.46	>3,600	n.a.
	25	5,403,865.964	25.09	>3,600	n.a.
M43	20	155,142.306	1.48	9.23	454,778
	21	320,726.566	1.45	11.49	531,549
	22	627,364.757	4.86	>3,600	n.a.
	23	1,259,716.112	4.87	>3,600	n.a.
	24	2,661,312.124	8.77	>3,600	n.a.
	25	5,432,750.885	8.63	>3,600	n.a.
RbcL55	20	225,473.659	2.96	65.77	3,846,795
	21	480,122.632	2.79	130.04	7,032,882
	22	995,596.435	2.73	234.98	11,627,808
	23	2,049,037.184	10.33	>3,600.	n.a.
	24	4,145,906.240	10.69	>3,600	n.a.
	25	8,595,977.714	9.74	>3,600	n.a.
M62	20	290,179.592	2.18	11.43	589,755
	21	576,663.719	4.87	13.07	563,435
	22	1,162,187.379	5.37	>3,600	n.a.
	23	2,366,103.293	8.83	>3,600	n.a.
	24	4,903,286.768	8.80	>3,600	n.a.
	25	9,927,436.569	8.86	>3,600	n.a.
Rana64	20	77,259.420	4.74	1.40	42,519
	21	179,623.137	3.48	3.47	127,983
	22	372,354.375	3.53	4.24	146,933
	23	763,885.643	5.22	11.27	285,718
	24	1,558,525.951	5.18	21.58	516,498
	25	3,413,097.109	8.75	>3,600	n.a.
M82	20	82,092.395	5.77	359.83	20,752,303
	21	169,136.637	5.43	852.38	44,877,124
	22	368,956.166	5.90	3,043.69	135,930,327
	23	752,462.574	6.38	>3,600	n.a.
	24	1,549,887.516	6.91	>3,600	n.a.
	25	3,206,058.298	11.92	>3,600	n.a.

Table 5.5. The continuation of Tables 5.3 and 5.4 for the multi core version.

6

Conclusion and Perspective

In this thesis we have provided a study of the Balanced Minimum Evolution Problem from multiple different perspectives. We have (i) put the BMEP into its historical context and evaluated the insights that arise from all previous investigations of the problem; (ii) given alongside the connection of the BMEP to the Traveling Salesman Problem and the Huffman Coding Problem an interpretation of the BMEP as a cross-entropy minimization problem; (iii) studied the computational complexity of the problem and proposed ways of exploiting this knowledge; (iv) significantly improved upon the state-of-the-art exact solution algorithm for the BMEP through theoretical considerations and empirical investigations; (v) addressed issues with the statistical consistency and numerical stability of the problem which have not be acknowledged in earlier works on the topic.

Beyond the BMEP we studied other phylogenetic estimation problems under the minimum evolution criterion and considered their statistical consistency and numerical stability aspects from an information theory perspective. Furthermore, we discussed the advantages and disadvantages of optimizing over rooted and unrooted binary trees, respectively, and we highlighted questions on the expansion of the information theory perspective and the structure of unrooted binary trees that require additional research efforts. Major obstacles in the development of answers to these questions are on the one hand the lack of sufficient conditions for the set of path-length matrices Θ and our limited understanding of the facets of the BME polytope. On the other hand difficulties in finding well structured sets of probability distributions (3.29) that provide a concretisation of Problem 6 and avoid statistical inconsistencies, e.g., a structure provided by a majorization ordering (see Section 4.2.3).

Further research directions not yet mentioned in the concluding remarks of previous chapters are as follows: (i) the investigation of the approximation quality of agglomerative approaches because of the empirical evidence that the NJ algorithm consistently produces favourable taxa orders; (ii) studying the performance of the new state-of-the-art exact solution algorithm for structured sets of distance matrices; (iii) the influence of the cross-entropy reformulation of the BMEP on the computational complexity in practice, e.g., an average-case complexity analysis; (iv) variations of Formulation 3 which use special ordered sets different from the sets $P_{ij}(T_S)$ derived from the stepwise addition strategy.

References

C. Abbosh, N. Birkbak, G. Wilson, and et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545:446–451, April 2017.

V. A. Albert. *Parsimony, phylogeny, and genomics*. Oxford Bioscience, UK, 2005.

S. Amiroch, M. S. Pradana, M. I. Irawan, and I. Mukhlash. Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method. *Journal of Physics: Conference Series*, 890(1):012080, September 2017.

R. Aringhieri, D. Catanzaro, and M. Di Summa. Optimal solutions for the balanced minimum evolution problem. *Computers and Operations Research*, 38:1845–1854, 2011.

K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2-3):251–278, 1999.

A. F. Auch, Henz S. R., B. R. Holland, and Göker M. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics*, 7(350):1–5, July 2006.

D. A. Bader, B. M. E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In *SPIE ITCom: Commercial application for high-performance computing*, pages 159–168. SPIE, Bellingham, WA, July 2001.

N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2015.

W. A. Beyer, M. Stein, T. Smith, and S. Ulam. A molecular sequence metric and evolutionary trees. *Mathematical Biosciences*, 19:9–25, 1974.

L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, November 2001.

B. E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the USA*, 83:5155–5159, 1986.

Georg Böcherer and R. A. Amjad. Informational divergence and entropy rate on rooted trees with probabilities. In *IEEE International Symposium on Information Theory*, pages 176–180. IEEE Computer Society, Honolulu, HI, USA, 2014.

M. Bordewich, O. Gascuel, K. Huber, and V. Moulton. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:110–117, 2009a.

M. Bordewich, O. Gascuel, K. T. Huber, and V. Moulton. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):110–117, 2009b.

C. Bowern and Q. Atkinson. Computational phylogenetics and the internal structure of pama-nyungan. *Language*, 88(4):817–845, December 2012.

R. A. Brualdi. The DAD theorem for arbitrary row sums. In *Proceedings of the American Mathematical Society*, volume 45, pages 189–194, Providence, Rhode Island, USA, August 1974.

P. Buneman. The recovery of trees from measure of dissimilarities. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Archaeological and Historical Science*, pages 387–395. Edinburgh University Press, Edinburgh, UK, 1971.

P. Buneman. A note on the metric properties of trees. *Journal of combinatorial theory*, 17 (B):48–50, February 1974.

R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science*, 286(5446):1921–1925, December 1999.

S. Caminiti, I. Finocchi, and R. Petreschi. On coding labeled trees. *Theoretical Computer Science*, 382:97–108, 2007.

E. Castro-Nallar, M. Pérez-Losada, G. F. Burton, and K. A. Crandall. The evolution of HIV: Inferences using phylogenetics. *Molecular Phylogenetics and Evolution*, 62(2):777–792, February 2012.

D. Catanzaro. The minimum evolution problem: Overview and classification. *Networks*, 53 (2):112–125, 2009.

D. Catanzaro. Estimating phylogenies from molecular data. In R. Bruni, editor, *Mathematical approaches to polymer sequence analysis and related problems*, pages 149–176. Springer, NY, 2011.

D. Catanzaro and R. Pesenti. Enumerating vertices of the balanced minimum evolution polytope. *Computers and Operations Research*, 109:209–217, 2019.

D. Catanzaro, L. Gatto, and M. Milinkovitch. Assessing the applicability of the GTR nucleotide substitution model through simulations. *Evolutionary Bioinformatics*, 2:145–155, 2006a.

D. Catanzaro, R. Pesenti, and M. Milinkovitch. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model. *Bioinformatics*, 22(6):708–715, 2006b.

D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-Gonzáles. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. *Networks*, 53(2): 126–140, 2009.

D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-Gonzáles. The balanced minimum evolution problem. *INFORMS Journal on Computing*, 24(2):276–294, 2012.

D. Catanzaro, M. Labbé, and R. Pesenti. The balanced minimum evolution problem under uncertain data. *Discrete Applied Mathematics*, 161(13–14):1789–1804, 2013a.

D. Catanzaro, R. Ravi, and R. Schwartz. A mixed integer linear programming model to reconstruct phylogenies from single nucleotide polymorphism fragments under the maximum parsimony criterion. *BMC Algorithms for Molecular Biology*, 8(1):3, 2013b.

D. Catanzaro, R. Aringhieri, M. di Summa, and R. Pesenti. A branch-price-and-cut algorithm for the minimum evolution problem. *European Journal of Operational Research*, 244(3): 753–765, 2015.

D. Catanzaro, S. E. Schackney, A. A. Schäffer, and R. Schwartz. Classifying the progression of Ductal Carcinoma from single-cell sampled data via integer linear programming: A case study. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(4): 643–655, 2016.

D. Catanzaro, M. Frohn, and R. Pesenti. An information theory perspective on the balanced minimum evolution problem. *Operations Research Letters*, 48(3):362–367, May 2020a.

D. Catanzaro, R. Pesenti, and L. A. Wolsey. On the Balanced Minimum Evolution polytope. *Discrete Optimization*, 36:1–33, 2020b.

A. L. Cauchy. *L'Imprimerie Royale, Debure frères, Libraires du Roi et de la Bibliothèque du Roi*. Cours d'Analyse de l'École Royale Polytechnique, 1821.

L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19:233–257, May 1967.

E. Çela. *The Quadratic Assignment Problem*. Kluwer Academic Publishers, Boston, MA, 1997.

B. S. W. Chang and M. J. Donoghue. Recreating ancestral proteins. *Trends in Ecology and Evolution*, 15(3):109–114, March 2000.

X. Cheng and D. Z. Du. *Steiner Trees in Industry*. Kluwer Academic Publishers, Boston, MA, 2001.

S. A. Chowdhury, S. E. Shackney, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*, 29(13):i189–i198, 2013.

D. Cieslik. *Steiner minimal trees*. Springer, Boston, MA, USA, 1998.

A. Criscuolo and C. J. Michel. Phylogenetic inference with weighted codon evolutionary distances. *Journal of Molecular Evolution*, 68:377–392, March 2009.

C. Darwin. *On the Origin of Species*. Harvard University Press, Cambridge, MA, USA, 1964.

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

F. Denis and O. Gascuel. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127:66–77, 2003.

R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *Journal of Computational Biology*, 9(5):687–705, 2002.

R. Desper and O. Gascuel. Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3):587–598, December 2004.

R. Desper and O. Gascuel. The minimum-evolution distance-based approach to phylogeny inference. *Mathematics of Evolution and Phylogeny*, pages 1–32, January 2005.

R. Desper and O. Gascuel. *Encyclopedia of Algorithms*, chapter Distance-based Phylogeny Reconstruction (Optimal Radius). Springer, 2008.

S. L. Devadoss, C. Durell, and S. Forcey. Split network polytopes and network spaces. In *The 31st international conference on Formal Power Series and Algebraic Combinatorics*, volume 82B, page 68. Séminaire Lotharingien de Combinatoire, July 2019.

D. Z. Du and X. Hu. *Steiner tree problems in computer communication networks*. World Scientific Publishing Company, Singapore, 2008.

D. Z. Du, J. M. Smith, and J. H. Rubinstein. *Advances in Steiner trees*. Kluwer Academic Publishers, Boston, MA, USA, 2000.

W. E. Duellman, A. B. Marion, and S. B. Hedges. Phylogenetics, classification, and biogeography of the treefrogs (Amphibia: Anura: Arboranae). *Zootaxa*, 4104(1):1–109, April 2016.

K. Eickmeyer, P. Huggins, L. Pachter, and R. Yoshida. On the optimality of the neighbor-joining algorithm. *Algorithms for Molecular Biology*, 3(1):5, February 2008.

P. L. Erdös, M. A. Steel, L. A. Székély, and T. Warnow. A few logs suffice to build (almost) all trees: Part I. *Random Structures and Algorithms*, 14(2):153–184, 1999.

J. S. Farris. Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92, 1970.

J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

J. Felsenstein. An alternating least-squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46:101–111, 1997.

J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

S. Fiorini and G. Joret. Approximating the balanced minimum evolution problem. *Operations Research Letters*, 40(1):31–35, January 2012.

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

S. Forcey, L. Keefe, and W. Sands. Split-facets for balanced minimal evolution polytopes and the permutoassociahedron. *Bulletin of Mathematical Biology, in press*, 79:975–994, March 2017.

Stefan Forcey, Logan Keefe, and William Sands. Facets of the balanced minimal evolution polytope. *Journal of Mathematical Biology*, 73(2):447–468, December 2015.

P. Forster, L. Forster, Colin Renfrew, and M. Forster. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the USA*, 117 (11):9241–9243, 2020.

M. Frohn. On the approximability of the fixed-tree balanced minimum evolution problem. *To appear in Optimization Letters*, 2020.

M. R. Garey and D. S. Johnson. *Computers and Intractability: A guide to the theory of NP-Completeness*. Freeman, New York, NY, 2003.

O. Gascuel. On the optimization principle in phylogenetic analysis and the minimum evolution criterion. *Journal of Classification*, 19(1):67–69, 2000.

O. Gascuel. *Mathematics of evolution and phylogeny*. Oxford University Press, New York, NY, 2005.

O. Gascuel and D. Levy. A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *Journal of Classification*, 13:129–155, 1996.

O. Gascuel and A. McKenzie. Performance analysis of hierarchical clustering algorithms. *Journal of Classification*, 21:3–18, 2004.

O. Gascuel and M. Steel. A 'stochastic safety radius' for distance-based tree reconstruction. *Algorithmica*, 74:1386–1403, 2016.

O. Gascuel and M. A. Steel. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23 (11):1997–2000, 2006.

O. Gascuel, D. Bryant, and F. Denis. Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50:621–627, 2001.

E. Gawrilow and M. Joswig. Polymake: A framework for analyzing convex polytopes. In G. Kalai and G. M. Ziegler, editors, *Polytopes - combinatorics and computation*, pages 43–73. Birkhäuser, 2000.

Xing-Yi Ge, Jia-Lu Li, Xing-Lou Yang, Aleksei A. Chmura, Guangjian Zhu, Jonathan H. Epstein, Jonna K. Mazet, Ben Hu, Wei Zhang, Cheng Peng, Yu-Ji Zhang, Chu-Ming Luo, Bing Tan, Ning Wang, Yan Zhu, Gary Crameri, Shu-Yi Zhang, Lin-Fa Wang, Peter Daszak, and Zheng-Li Shi. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, 503:535–538, October 2013.

S. Gerschgorin. Über die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, 6:749–754, 1931.

G. Golub and C. van Loan. *Matrix computations*. The Johns Hopkins University Press, London, 1996.

D. Gusfield. The steiner tree problem in phylogeny. Technical Report 334, Yale University, New Haven, CT, 1984.

G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, UK, 1978.

P. H. Harvey, A. J. L. Brown, J. M. Smith, and S. Nee. *New uses for new phylogenies*. Oxford University Press, Oxford, UK, September 1996.

M. Hasegawa, H. Kishino, and T. Yano. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, November 1981.

D. C. Haws, T. L. Hodge, and R. Yoshida. Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. *Bulletin of Mathematical Biology*, 73: 2627–2648, March 2011.

M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 1982.

T. Hirschler and W. Woess. Comparing entropy rates on finite and infinite rooted trees. *IEEE Transactions on Information Theory*, 64(8):5570–5580, August 2018.

L. J. Hubert and P. Arabie. Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology*, 48:281–317, 1995.

J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314, 2001.

J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic networks*. Cambridge University Press, Cambridge, UK, 2011.

F. K. Hwang, D. S. Richards, and P. Winter. *The Steiner tree problem*. North-Holland, Amsterdam, The Netherlands, 1992.

M. Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. arXiv: 1609.06349, 2016.

G. Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5:180189, October 2018.

D. S. Johnson, J. K. Lenstra, and A. H. G. Rinnooy Kan. The complexity of the network design problem. *Networks*, 8:279–285, 1978.

C. Jordan. Sur les assemblages des lignes. *Journal für die reine und angewandte Mathematik*, 70:185–190, 1869.

T. H. Jukes and C.R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York, NY, 1969.

M. Jung. *Évolution du VIH: méthodes, modèles et algorithmes*. PhD thesis, Université Montpellier II - Sciences et Techniques du Languedoc, France, 2012.

Suhas Kadam, Tri D. Vuong, Dan Qiu, Clinton G. Meinhardt, Li Song, Rupesh Deshmukh, Gunvant Patil, Jinrong Wan, Babu Valliyodan, Andrew M. Scaboo, J. Grover Shannon, and Henry T. Nguyen. Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. *Plant Science*, 242:342–350, January 2016.

M. M. Kapranov. The permutoassociahedron, Mac Lane's coherence theorem and asymptotic zones for the KZ equation. *Journal of Pure and Applied Algebra*, 85(2):119–142, March 1993.

L. Khachiyan. Diagonal matrix scaling is $\mathcal{NP}$-hard. *Linear algebra and its applications*, 234: 173–179, 1996.

K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics*, 23:235–252, 1971.

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

W. S. Klung, M. R. Cummings, C. A. Spencer, M. A. Palladino, and D. Killian. *Concepts of Genetics*. Pearson, NY, 2019.

D. L. Kreher and D. R. Stinson. *Combinatorial algorithms: Generation, enumeration, and search*. CRC Press, Boca Raton, FL, March 1999.

W. John Kress, David L. Erickson, Nathan G. Swenson, Jill Thompson, Maria Uriarte, and Jess K. Zimmermann. Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a puerto rican forest dynamics plot. *PLoS One*, 5(11):e15409, November 2010.

D. Lähnemann, J. Köster, E. Szczurek, D. J McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. Stephan-Otto Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. de Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. Jessurun Lobo, E. M. Keizer, I. Khatri, S. M Kielbasa, J. O. Korbel, A. M. Kozlov, T. Kuo, B. P. F. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. de Ridder, A. E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

C. C. Lai, T. P. Shih, W. C. Ko, H. J. Tang, and P. R. Hsue. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The

epidemic and the challenges. *International Journal of Antimicrobial Agents*, 55(3):105924, 2020.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, February 1984.

E. Lawler, J. Lenstra, A. Rinnooy Kan, and D. Shmoys. *The Traveling Salesman Problem: A guided tour of combinatorial optimization*. John Wiley and Sons Ltd., New York, NY, 1985.

V. Lefort, R. Desper, and O. Gascuel. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10): 2798–2800, October 2015.

M. A. Leibold, E. P. Economo, and P. Peres-Neto. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. *Ecology Letters*, 13(10):1290–1299, August 2010.

E. M. Lemmon and A. R. Lemmon. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44:99–121, November 2013.

F. Lemoine, L. Blassel, J. Voznica, and O. Gascuel. COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM. *Bioinformatics*, 2020.

X. Li and Y. Mao. *Generalized connectivity of graphs*. Springer, Boston, MA, USA, 2016.

H. R. Lourenço, O. Martin, and T. Stützle. Iterated local search. In F. Glover and G. A. Kochenberger, editors, *Handbook of Metaheuristics*, volume 57 of *International Series in Operations Research & Management Science*, pages 320–353. Springer, Boston, MA, USA, 2003.

C. L. Lu, C. Y. Tang, and R. C. T. Lee. The full Steiner tree problem. *Theoretical Computer Science*, 306:55–67, 2003.

V. Makarenkov and B. Leclerc. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification*, 16:3–26, 1999.

Vladimir Makarenkov and Bruno Leclerc. Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In B. Mirkin, F.R. McMorris, F. Roberts, and A. Rzhetsky, editors, *Mathematical hierarchies and Biology*, volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 183–208. American Mathematical Society, Providence, RI, 1997.

M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper. The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404, May 2003.

R. Kipp Martin. *Large Scale Linear and Integer Optimization: A Unified Approach*. Springer-Verlag, New York, NY, 1999.

C. Mavian, S. K. Pond, S. Marini, B. Rife Magalis, A. M. Vandamme, S. Dellicour, S. V. Scarpino, C. Houldcroft, J. Villabona-Arenas, T. K. Paisie, N. S. Trovão, C. Boucher, Y. Zhang, R. H. Scheuermann, O. Gascuel, T. T.-Y. Lam, M. A. Suchard, A. Abecasis, E. Wilkinson, T. de Oliveira, A. I. Bento, H. A. Schmidt, D. Martin, J. Hadfield, N. Faria, D. N. Grubaugh, R. A. Neher, G. Baele, P. Lemey, T. Stadler, J. Albert, K. A. Crandall, T. Leitner, A. Stamatakis, M. Prosperi, and M. Salemi. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proceedings of the National Academy of Sciences of the USA*, 117(23):12522–12523, 2020.

J. E. McCormack, S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66(2):526–538, February 2013.

J. A. McGuire, C. C. Witt, J. V. Remsen Jr., A. Corl, D. L. Rabosky, D. L. Altshuler, and R. Dudley. Molecular phylogenetics and the diversification of hummingbirds. *Current Biology*, 24(8):910–916, April 2014.

L. Mirsky. *An Introduction to Linear Algebra*. Dover Publications, New York, 1955.

N. Misra, G. E. Blelloch, R. Ravi, and R. Schwartz. Generalized buneman pruning for inferring the most parsimonious multi-state phylogeny. *Journal of Computational Biology*, 18(3):445–57, 2011.

D. S. Mitrinović, J. E. Pečarić, and A. M. Fink. *A. L. Cauchy*. Kluwer Academic Publishers Group, Dordrecht, 1993.

M. A. Myers, G. Satas, and B. J. Raphael. Calder: Inferring phylogenetic trees from longitudinal tumor samples. *Cell Systems*, 8(5):514–522, 2019.

M. Nei. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, UK, 2000.

G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, New York, NY, 1999.

H.-C. Ng, S. Liu, and W. Luk. Reconfigurable acceleration of genetic sequence alignment: A survey of two decades of efforts. In M. Santambrogio, D. Göhringer, D. Stroobandt, N. Mentens, and J. Nurmi, editors, *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–8, 2017.

C. Notredame. Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics*, 3(1):131–144, 2004.

C. Y. Ou, C. A. Ciesielski, 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. Maclnnes, J. W. Curran, and H. W. Jaffe. Molecular epidemiology of HIV transmission in a dental practice. *Science*, 256(5060):1165–1171, May 1992.

L. Pachter and B. Sturmfels. The mathematics of phylogenomics. *SIAM Review*, 49(1):3–31, 2007.

R. D. M. Page and E. C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, UK, 1998.

F. Pardi. *Algorithms on Phylogenetic Trees*. PhD thesis, University of Cambridge, UK, 2009.

F. Pardi and O. Gascuel. Combinatorics of distance-based tree inference. *PNAS*, 109(41): 16443–16448, October 2012.

F. Pardi and O. Gascuel. *Encyclopedia of Evolutionary Biology*, chapter Distance-based methods in phylogenetics. Elsevier, 2016.

F. Pardi, S. Guillemot, and O. Gascuel. Robustness of phylogenetic inference based on minimum evolution. *Bulletin of Mathematical Biology*, 72:1820–1839, 2010.

D. S. Parker and P. Ram. The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees. *SIAM Journal on Computing*, 28(5): 1875–1905, 1996.

B. N. Parlett and T. L. Landis. Methods for scaling to doubly stochastic form. *Linear Algebra and its Applications*, 48:53–79, December 1982.

Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51(1):41–47, July 2000.

G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz. Reconstructing tumor phylogenies from heterogeneous single-cell data. *Journal of Bioinformatics and Computational Biology*, 5(2a):407–427, 2006.

V. R. Perovic. Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses. *Journal of Mathematical Chemistry*, 51:2238–2255, June 2013.

Art F. Y. Poon, Jeffrey B. Joy, Conan K. Woods, Susan Shurgold, Guillaume Colley, Chanson J. Brumme, Robert S. Hogg, Julio S. G. Montaner, and P. Richard Harrigan. The impact of clinical, demographic and risk factors on rates of HIV transmission: A population-based phylogenetic analysis in British Columbia, Canada. *The Journal of Infectious Diseases*, 211(6):926–935, March 2015.

P. C. Pop. *Generalized network design problems: Modeling and optimization*. De Gruyter, Berlin, Germany, 2012.

K. Popper. *The logic of scientific discovery*. Routledge, London, UK, 2002.

H. J. Prömel and A. Steger. *The Steiner tree problem: A tour through graphs, algorithms, and complexity*. Vieweg+Teubner Verlag, Berlin, 2002.

V. Reiner and G. M. Ziegler. Coxeter-associahedra. *Zuse Institute Berlin (ZIB), Berlin, Takustrasse 7, 14195, Germany*, 41(2):364–393, December 1994.

D. Reinhard. *Graph Theory*. Springer-Verlag, New York, NY, 2005.

M. Riester, C. Stephan-Otto Attolini, R. J. Downey, S. Singer, and F. Michor. A differentiation-based phylogeny of cancer subtypes. *PLoS Computational Biology*, 6: e100077, 2010a.

M. Riester, C. Stephan-Otto Attolini, R. J. Downey, S. Singer, and F. Michor. A differentiation-based phylogeny of cancer subtypes. *PLoS Computational Biology*, 6(5): e1000777, 2010b.

F. Rodriguez, J. L. Oliver, A. Marin, and J. R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142(4):485–501, February 1990.

M. S. Rosenberg. *Sequence Alignment: Methods, Models, Concepts, and Strategies*. University of California Press, LA, 2011.

H. A. Ross and A. G. Rodrigo. Immune-mediated positive selection drives Human Immunodeficency Virus type 1 molecular variation and predicts disease duration. *Journal of Virology*, 76(22):11715–11720, November 2002.

A. Rzhetsky and M. Nei. A simple method for estimating and testing minimum evolution trees. *Computer Applications in the Biosciences*, 10:409–412, 1992.

A. Rzhetsky and M. Nei. Theoretical foundations of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5):1073–1095, September 1993.

A. Rzhetsky and M. Nei. METREE: A program package for inferring and testing minimum evolution trees. *Computer Applications in the Biosciences*, 10:409–412, 1994.

N. Saitou and T. Imanishi. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbour-joining methods of phylogenetic tree construction in obtaining the correct tree. *Molecular Biology and Evolution*, 6:514–525, 1989.

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann, San Francisco, CA, 5th edition, 2017.

S. M. Scheiner and D. P. Mindell. *The theory of evolution: Principles, concepts and assumptions*. University of Chicago Press, Chicago, IL, USA, 2020.

S. Schulmeister. Inconsistency of maximum parsimony revisited. *Systematic Biology*, 53(4): 521–528, 2004.

R. Schwartz. Computational models for cancer phylogenetics. In T. Warnow, editor, *Bioinformatics and Phylogenetics. Computational Biology*, volume 29, pages 243–275. Springer, Cham, 2019.

C. Semple and M. A. Steel. *Phylogenetics*. Oxford University Press, New York, NY, 2003.

C. Semple and M. A. Steel. Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32(4):669–680, May 2004.

M. Simonsen, T. Mailund, and C. N. S. Pedersen. Rapid neighbour joining. *Lecture Notes in Bioinformatics*, 5251:113–122, October 2008.

R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35:876–879, 1964.

R. R. Sokal and C. D. Michener. A quantitative approach to a problem in classification. *Evolution*, 1957.

S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi, and R. Schwartz. Algorithms for efficient near-perfect phylogenetic tree reconstruction in theory and practice. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):561–571, 2007.

S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi, and R. Schwartz. Mixed integer linear programming for maximum parsimony phylogeny inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3):323–331, 2008.

G. W. Stuart, K. Moffett, and S. Baker. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18(1):100–108, January 2002a.

G. W. Stuart, K. Moffett, and J. J. Leader. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19(4):554–562, April 2002b.

J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

A. Subramanian, S. Shackney, and R. Schwartz. Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(6):1422–1431, 2013.

Simon Tavare. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1987.

A. Valiente-Banuet and M. Verdú. Plant facilitation and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44:347–366, November 2013.

S. Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 15(3):376–389, May 2014.

S. Vinga and J. Almeida. Alignment-free sequence comparison - A review. *Bioinformatics*, 19(4):513–523, March 2003.

L. S. Vinh and A. von Haeseler. Shortest triplet clustering: Reconstructing large phylogenies using representative sets. *BMC Bioinformatics*, 6(92):1–14, 2005.

M. V. Volkenstein and M. A. Livshits. *Speciation and bifurcations*, 23(1):1–5, 1989.

P. J. Waddell and M. A. Steel. General time-reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. *Molecular Phylogenetics and Evolution*, 8:398–414, 1997.

Alex D. Washburne, James T. Morton, Jon Sanders, Daniel McDonald, Qiyun Zhu, Angela M. Oliverio, and Rob Knight. Methods for phylogenetic analysis of microbiome data. *Nature Microbiology*, 3:652–661, May 2018.

M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64:199–213, 1977.

B. Y. Wu and K. M. Chao. *Spanning trees and optimization problems*. Chapman and Hall/CRC, Boca Raton, FL, 2004.

Z. Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39:105–111, 1994.

Z. Yang. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, UK, 2014.

P. Zhou, X. L. Yang, Z. L. Shi, and *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579:270–273, 2020.

A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186, October 2017.