A Tutorial on the Balanced Minimum Evolution Problem

Daniele Catanzaro, Martin Frohn, Olivier Gascuel, Raffaele Pesenti

A Tutorial on the Balanced Minimum Evolution Problem

Daniele Catanzaro[a,*,1], Martin Frohn[a,2], Olivier Gascuel[c,3] and Raffaele Pesenti[d,4]

[a]*Center for Operations Research and Econometrics, Université Catholique de Louvain, Voie du Roman Pays 34, 1348, Louvain-la-Neuve, Belgium.*
[c]*Institut de Systématique, Evolution, Biodiversité (ISYEB - UMR 7205 CNRS & Muséum National d'Histoire Naturelle), Paris, France.*
[d]*Department of Management, University Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venice, Italy.*

ARTICLE INFO

Keywords:
Combinatorial optimization; balanced minimum evolution problem; network design; information entropy; mathematics of evolution; phylogenetics;

ABSTRACT

The *Balanced Minimum Evolution Problem* (BMEP) is an $\mathcal{APX}$-hard network design problem that consists of finding a minimum length unrooted binary tree (also 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. In this article, we introduce the reader to the problem and present the current state-of-the-art; we include the most important achievements reached so far and the challenges that still remain to be addressed.

1. Introduction

Consider a set $\Gamma = \{1, 2, \dots, n\}$ of $n \geq 3$ distinct aligned molecular sequences (such as DNA, RNA, codon sequences or whole genomes) hereafter referred to as *taxa* (see Figure 1). 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 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.

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 – encodes a measure of the dissimilarity between the input taxa i and j in Γ (see, e.g., Figure 3). Then, the *Balanced Minimum Evolution Problem* (BMEP) consists of finding a *phylogeny* T of Γ that minimizes the following *length function*

$$L(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{d_{ij}}{2^{\tau_{ij}}}, \tag{1}$$

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 [3, 21, 25, 28, 118] (see Figure 2). The aim of this paper is to celebrate the first 20 years of research activities on the BMEP. We introduce the reader to the biological and statistical foundations of the BMEP, to its combinatorics, to its optimization aspects, and to its connections with information theory. 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. In order to widen the readability of this tutorial, we assume no a-priori technical knowledge on the topic and we will not hesitate to recall fundamental concepts from the mathematics of evolution and phylogenetics whenever it will be necessary, by starting from the fundamental concept and uses of a phylogeny presented in the next section.

*Corresponding author
ORCID(s): 0000-0001-9427-1562 (D. Catanzaro); 0000-0002-5002-4049 (M. Frohn); 0000-0001-5890-4238 (R. Pesenti)
[1]Email: daniele.catanzaro@uclouvain.be (Daniele Catanzaro)
[2]Email: martin.frohn@uclouvain.be (Martin Frohn)
[3]Email: olivier.gascuel@mnhn.fr (Olivier Gascuel)
[4]Email: pesenti@unive.it (Raffaele Pesenti)

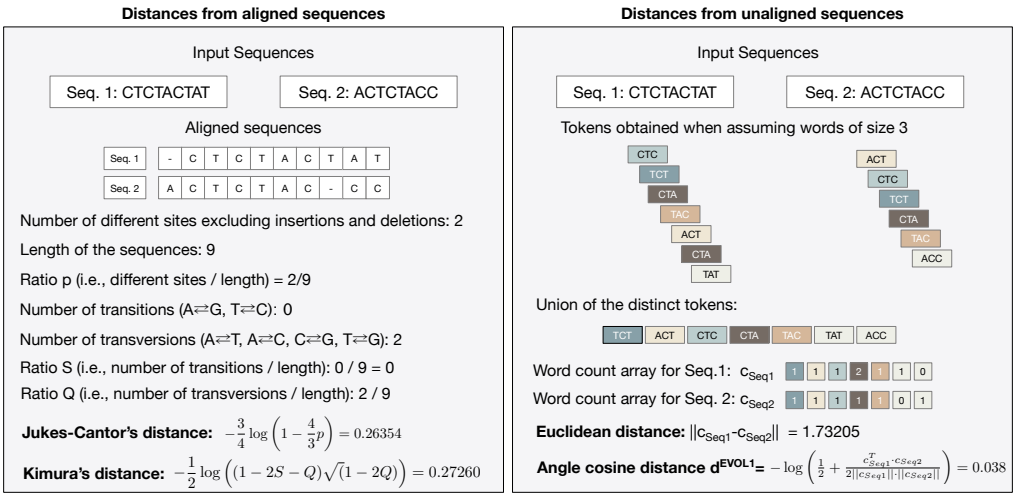


Figure 1: Some examples of how to derive the evolutionary distance for a pair of dummy molecular sequences, namely "CTCTACTAT" and "ACTCTACC". In particular, the subfigure on the left shows two possible distances derived from the dissimilarity measures proposed by Jukes and Cantor [82] and Kimura [87], respectively. Both measures are based on the time-reversible Markov model of DNA sequence evolution (see [22, 27, 51, 70, 93, 131, 153, 159, 163]) and require that the two sequences undergo an *alignment process* [132] before being analyzed, i.e., a process that generates two new equal length sequences from the input ones by maximizing their regions of similarity (i.e., their character matching). 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 [51] for details). The subfigure on the right shows two possible distances derived from the alignment-free dissimilarity measures proposed by Blaisdell [9] and Stuart et al. [149, 150], respectively. Because solving the multiple alignment problem is notoriously $\mathcal{NP}$-hard [132], this family of distances may prove particularly useful to process very large molecular datasets. We refer the reader to [132] and [98, 113, 114, 145] for an introduction and recent developments on the alignment problem as well as to [51, 117] for a review of the numerous alignment-based distances currently discussed in the literature. We refer instead the reader to [155, 156, 166] for an introduction and recent surveys on alignment-free distances.

1.1. Phylogenies

A phylogeny describes the hierarchical evolutionary relationships of a given set of taxa, based on their observed inherited traits (e.g., their molecular sequences) [19, 51, 60, 117, 143]. The *topology* of a phylogeny (i.e., the structure of U in the case of the BMEP) and its corresponding biological meaning may depend on the specific application or use. For example, in the context of tumor evolution, the topology of a phylogeny can be represented as an arborescence having as a root the taxon relative to a healty cell and as descendent vertices the taxa representing tumor cells (see, e.g., [6, 30, 122]). In systematics, instead, the topology of a phylogeny is usually encoded as an UBT in which the terminal vertices (or *leaves*) represent the observed taxa; *internal vertices* represent speciation events that occurred throughout evolution of taxa; *edges* represent estimated evolutionary relationships; and *edge weights* represent a measure of the similarity between pairs of taxa [18].

By providing core insights on the evolutionary dynamics of many fine-scale molecular data, phylogenies 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 [45, 79, 83, 97, 108, 116, 154]. For example, phylogenies have been used to predict evolution of human influenza A [15, 123]; to understand the relationships between the virulence and evolution of HIV [17, 115, 124, 133]; to identify emerging viruses such as SARS [2, 67, 104]; to recreate and investigate ancestral proteins [33, 160]; to design neuropeptides causing smooth muscle contraction [5]; to relate geographic patterns to ecological and macro-evolutionary processes [69, 90, 96, 107]; or to uncover similarities in evolution of a number of human languages [12, 79]. Phylogenies have also been used to study the evolutionary processes underlying the genetic factors involved in common human diseases [24, 29, 109, 122, 147, 148] as well as those at the core of the progression of carcinomas over time [30, 35, 91, 110, 129, 130, 142, 152]. In particular, in the cancer context, phylogenies allowed the remarkable classification of tumor cells of given pathologies in subfamilies characterized by specific evolutionary traits [142]. A shared hope is that this classification could enable a better understanding of cellular atypia over time and, in the long run, suggest new therapeutic targets [6, 91, 110, 142].

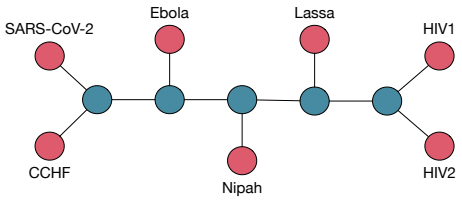


Figure 2: An example of a phylogeny of a set of seven taxa (red vertices), including the whole genomes of the Crimean-Congo Hemorrhagic Fever (CCHF) orthonairovirus, Ebolavirus, the Lassa mammarenavirus, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the Human Immunodeficiency Viruses (HIVs) 1 and 2, and the Nipah virus. The internal vertices of the phylogeny are marked in blue. Edge weights have been removed for the sake of readability. The above complete genomes 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). All of these genomes can be downloaded in a unique fasta-format file at `https://perso.uclouvain.be/daniele.catanzaro/SupportingMaterial/viruses.fasta`. The evolutionary distances for the considered genomes have been computed via Alfree (`http://150.254.123.165/alfree/`) by using d^{EVOL1} metrics (i.e., Stuart *et al.*' alignment-free *angle cosine metric*, see [149, 150] and Figure 1). The resulting distance matrix is shown in Figure 3. The above phylogeny 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

$$\tau = \begin{array}{c} \begin{matrix} \text{CCHF} & \text{Ebola} & \text{Lassa} & \text{SARS-CoV-2} & \text{HIV1} & \text{HIV2} & \text{Nipah} \end{matrix} \\ \begin{bmatrix} 0 & 3 & 5 & 2 & 6 & 6 & 4 \\ 3 & 0 & 4 & 3 & 5 & 5 & 3 \\ 5 & 4 & 0 & 5 & 3 & 3 & 3 \\ 2 & 3 & 5 & 0 & 6 & 6 & 4 \\ 6 & 5 & 3 & 6 & 0 & 2 & 4 \\ 6 & 5 & 3 & 6 & 2 & 0 & 4 \\ 4 & 3 & 3 & 4 & 4 & 4 & 0 \end{bmatrix} \end{array} \begin{array}{l} \text{CCHF} \\ \text{Ebola} \\ \text{Lassa} \\ \text{SARS-CoV-2.} \\ \text{HIV1} \\ \text{HIV2} \\ \text{Nipah} \end{array}$$

More recently, phylogenies 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 phylogeny of SARS-CoV-2 (see, e.g., Figure 4) shows how distinctive migratory histories and founder events may modulate the clinical presentation in affected patients [92, 106, 165]. This key information, in turn, assists the search for potential vaccines.

In the context of systematics (central for this article), the internal vertices of a phylogeny of n taxa usually have degree three [19, 51, 60, 117, 143]. This constraint is biologically founded for all the evolutionary processes that are

$$\begin{array}{c} \begin{matrix} \text{CCHF} & \text{Ebola} & \text{Lassa} & \text{SARS-CoV-2} & \text{HIV1} & \text{HIV2} & \text{Nipah} \end{matrix} \\ \begin{bmatrix} 0.0000000 & 0.0048683 & 0.0094948 & 0.0139234 & 0.0109681 & 0.0114508 & 0.0081327 \\ 0.0048683 & 0.0000000 & 0.0081206 & 0.0118547 & 0.0164171 & 0.0184611 & 0.0038627 \\ 0.0094948 & 0.0081206 & 0.0000000 & 0.0154856 & 0.0326514 & 0.0306672 & 0.0148241 \\ 0.0139234 & 0.0118547 & 0.0154856 & 0.0000000 & 0.0338913 & 0.0406092 & 0.0143275 \\ 0.0109681 & 0.0164171 & 0.0326514 & 0.0338913 & 0.0000000 & 0.0025788 & 0.0141298 \\ 0.0114508 & 0.0184611 & 0.0306672 & 0.0406092 & 0.0025788 & 0.0000000 & 0.0203479 \\ 0.0081327 & 0.0038627 & 0.0148241 & 0.0143275 & 0.0141298 & 0.0203479 & 0.0000000 \end{bmatrix} \end{array} \begin{array}{l} \text{CCHF} \\ \text{Ebola} \\ \text{Lassa} \\ \text{SARS-CoV-2} \\ \text{HIV1} \\ \text{HIV2} \\ \text{Nipah} \end{array}$$

Figure 3: The distance matrix computed by Alfree (`http://150.254.123.165/alfree/`) when analyzing the genomes listed in Figure 2.

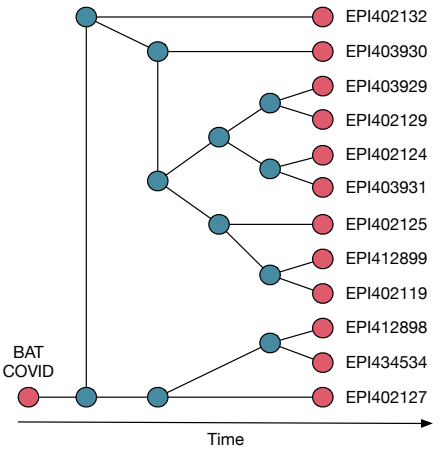


Figure 4: An example of the phylogeny of SARS-CoV-2 Chinese genomes sampled from affected patients during December 2019. Data, phylogeny, and taxa labels refer to GISAID nomenclature (https://www.gisaid.org/epiflu-applications/next-hcov-19-app/). The BAT COVID label refers to the bat (R. affinis) coronavirus isolate BatCoVRaTG13 from Yunnan Province, used as an outgroup in [56].

bifurcating (i.e., those processes for which the occurrence of mutations over time generates two distinct species at each speciation event [51, 117, 158]), but as shown in [117], it can also model evolutionary processes that include *polytomies*, i.e., multi-furcations during evolution of taxa. In fact, any m-ary tree can be transformed into an unrooted binary tree by just adding "dummy edges" (i.e., edges with zero length), as shown, e.g., in Figure 5. The degree constraint on the internal vertices of a phylogeny implicitly fixes the overall number of edges and internal vertices of a phylogeny to be $2n-3$ and $n-2$, respectively (see [19] for details). Hence, the topology of a phylogeny looks unrooted and binary (see, e.g., Figure 2). Moreover, the degree constraint also implies that the overall number of possible phylogenies for a set of n taxa is equal to $(2n-5)!! = 1\times 3\times 5\times 7\times \cdots \times (2n-5)$ [18, 51]. This exponential number of plausible alternatives entails the search for specific estimation criteria to select the proper phylogeny of a given set of taxa [19, 20, 23].

1.2. 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 [51, 60, 117]. Examples include the *parsimony criterion*, the criteria based on *distance methods*, the *likelihood criterion*, and the criteria based on *Bayesian inference* [51, 60, 74, 75, 117]. Discussing all of these estimation criteria goes beyond the scope of the present article. The reader interested in such a topic may refer however both to comprehensive sources such as [1, 51, 60, 111, 117, 143, 164] and surveys, such as those discussed in [18, 19]. 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 [19, 20, 59]:

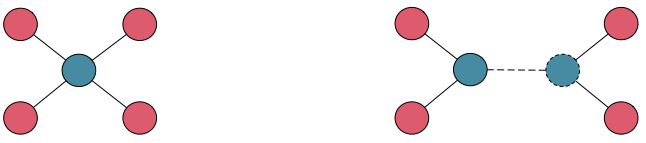


Figure 5: The 4-ary tree (on the left) can be transformed into an unrooted binary tree by adding a dummy vertex and a dummy edge (dashed, on the right).

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

$$\begin{aligned} &\textit{optimize} && \Lambda_\Gamma(T) && (2)\\ &\textit{s.t.} && \Omega_\Gamma(T) \geq 0 && (3)\\ & && T \in \mathcal{T} && (4) \end{aligned}$$

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; and $\Omega : \Gamma \times \mathcal{T} \to \mathbb{R}^m$, for some $m \geq 1$, is an auxiliary function that imposes 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 [51]. 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 and Ω_Γ. A phylogeny $T^* \in \mathcal{T}$ that optimizes Λ_Γ and satisfies the additional constraints described by 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) [51, 60]. This criterion was introduced by Desper and Gascuel [39] based on Pauplin's *Direct Calculation Formulae of branch-lengths* (DCF) discussed in [121] (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 [20]; (ii) a particular version of the *network design problem* [80, 125] 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 [34, 36, 43, 44, 68, 77, 99, 101, 127, 162]. BME is proven to be one of the most accurate estimation criteria in phylogenetics [95]. It inherits from distance methods the benefit of being able to handle virtually any kind of data for which a dissimilarity measure is available [142]; 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 article 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 we review the theoretical foundations of the BMEP, by presenting a brief historical perspective on the mathematical development that led to Pauplin's DCF [121]. In Section 3 we review the main combinatorial and optimization aspects of the BMEP, by reporting on the recent advances on its polyhedral combinatorics [25, 28, 46, 54, 55, 71]. In Section 4 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 [3, 21, 25, 41, 52, 57, 60, 118]. In Section 5 we recall the connections between the BMEP and information entropy [21, 139]. Finally, in Section 6 we provide a perspective on future developments related to the BMEP, by highlighting the currently open theoretical and computational challenges.

2. A brief history of the BMEP

We chronicle in this section the main stages of the mathematical development that led to *Pauplin's DCF* [121], by starting from the parsimony criterion and by proceeding then with the minimum evolution criterion and the balanced minimum evolution criterion. Figure 6 graphically summarizes the main stages of this development. Before proceeding, we introduce some notation and definitions that will prove useful throughout the article. 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}.

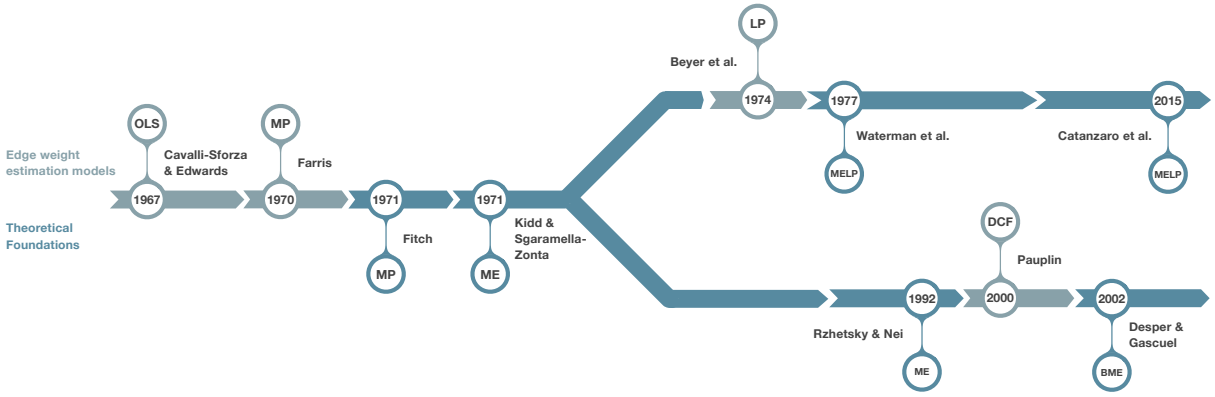


Figure 6: Graphical representation of the main stages of the mathematical development that led to Pauplin's Direct Calculation Formulae of branch-lengths (DCF). The light gray stages represent articles introducing specific edge weight estimation models. We can distinguish between the Ordinary Least-Squares (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 [18, 19, 20, 24]).

2.1. 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 [126]. 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 [126]. Inspired by the rationale of this law, Farris [48] and Fitch [53] laid in early 1970s the theoretical ground for the *parsimony criterion* of phylogenetic estimation [1], 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 7 and [7, 37, 48, 53, 140, 161]). 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) [88]. 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 evolve from speciation events to speciation events by means of *local minimum changes* [1, 7, 117]. Locally rather than globally minimum mostly because the selective pressure that acts on a taxon may not be constant throughout its evolution [1, 7]. 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 [48] and Fitch [53]'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 (see Figure 8) [1, 7, 51, 117, 161]. 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 $T \in \mathcal{T}$ [51]. 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:

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 8 shows a phylogeny T and a set of molecular sequences associated to the internal vertices of T that satisfies the above constraints and whose length is provably minimum for the six taxa considered in Figure 7. 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.

2.2. The minimum evolution criterion

In 1978, Felsenstein [49] 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 [51] for supplementary details and [141] 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, Denis and Gascuel [38], Gascuel [59, 60], Gascuel et al. [61], Rzhetsky and Nei [134, 135, 136], and Gascuel and McKenzie [63] – 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 [86] 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 [51, 60, 137]. In particular, Kidd and Sgaramella-Zonta introduced the concept of *true 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), and linked the true evolutionary distances with the estimated evolutionary distances d_{ij} from taxa (i.e., the entries of **D** defined in Section 1) through the following conjecture (proved true subsequently by Rzhetsky and Nei [135]):

Conjecture 1. *(Kidd and Sgaramella-Zonta [86])*
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*

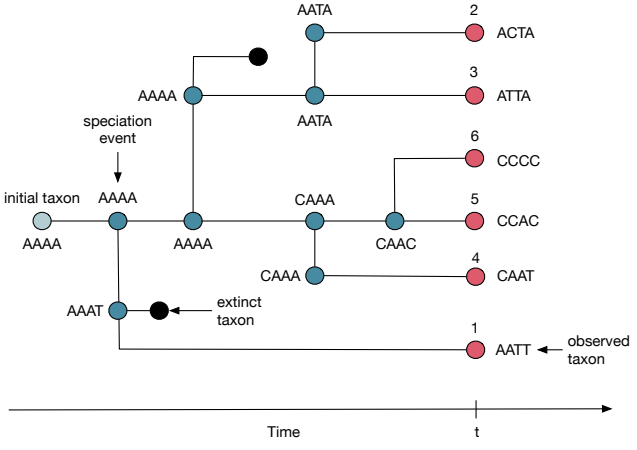


Figure 7: An example of a hypothetical speciation process that starting from an ancestral taxon AAAA (light blue vertex) generates six distinct observed taxa at time t (red vertices), namely AATT, ACTA, ATTA, CAAT, CCAC and CCCC. Random mutations as well as mutation caused by the selective pressure may change the original sequence, by giving rise to the bifurcations represented by the dark blue vertices. The sum of the Hamming distances between adjacent vertices is 11.

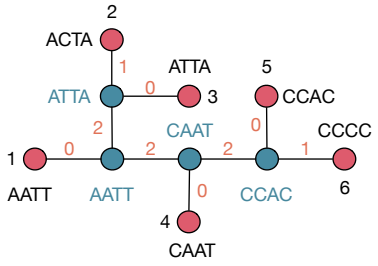


Figure 8: An example of the most parsimonious phylogeny approximating the speciation process shown in Figure 7. The sequences associated to the observed taxa (red vertices) are given and their hamming distances are reported on Figure 9. The sequences associated to the internal vertices of the phylogeny (marked in blue) are unknown and are estimated by solving an optimization problem that consists of finding a set of $n-2$ strings associated to the internal vertices of the given phylogeny so as to minimize the sum of the Hamming distances between pairs of adjacent vertices. These optimal internal sequences for the given phylogeny are marked in blue. The relative Hamming distances are marked in orange and represent the edge weights of the phylogeny, i.e., estimates of the changes between the predicted speciation events. The length of this phylogeny is 8 as opposed to 11 for the true speciation process in Figure 7. 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.

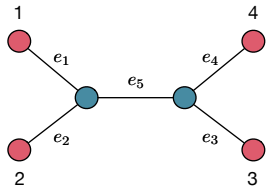
condition: for each pair of taxa i and j in Γ, the sum of the (positive) edge weights belonging to the path p_{ij} in T from i to j is greater than or equal to their estimated evolutionary distance d_{ij}.

Farris [48] and Fitch [53]'s parsimony criterion as well as Kidd and Sgaramella-Zonta [86]'s minimum evolution share the idea of searching for the phylogeny having minimum 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 [86] and Rzhetsky and Nei [135]' 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 [51, 60, 164]. 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 [164]. We refer the reader interested in deepening the fundamental theory on models and algorithms for computing these distances to [51, 60, 164]. 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 [135]. This model encodes a phylogeny $T \in \mathcal{T}$ by means of an *Edge–Path incidence matrix of a Tree* (EPT) (see [112] for a general introduction and [24] 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^e_{ij} 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 10(b) shows the EPT matrix corresponding to the phylogeny shown in Figure 10(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

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

Figure 9: Hamming distances associated to the six taxa in Figure 7.



(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)

Figure 10: (a) An example of a phylogeny of four taxa and the associated edge-path incidence matrix of a tree (b).

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 models asserts that

$$\sum_{e \in E(p_{ij})} y_{ij}^{e} w_e = d_{ij} \qquad \forall\, i, j \in \Gamma : i \neq j. \tag{5}$$

In general, (5) 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

$$\mathbf{w} = (\mathbf{Y}^t \mathbf{Y})^{-1} \mathbf{Y}^t \mathbf{D}^{\triangle} = \mathbf{Y}^{\dagger} \mathbf{D}^{\triangle} \tag{6}$$

i.e., when

$$w_e = \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij}, \quad \forall\, e \in E(T), \tag{7}$$

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

$$w_e = \sum_{\substack{i,j \in \Gamma \\ i \neq j}} \sum_{e \in E(p_{ij})} y_{ij}^{e\dagger} d_{ij}, \quad \forall\, e \in E(T). \tag{8}$$

It is easy to see that Problem 3 is a particular case of Problem 1 where $\Lambda_\Gamma = L_{ME}$ and the constraints imposed by Ω_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 $e \in E(T)$.

Rzhetsky and Nei [135] showed that if the distance matrix $\mathbf{D}$ is an unbiased estimate of the true distance matrix from taxa, then the optimal solution to (8) is statistical consistent. This milestone result proved Kidd and Sgaramella-Zonta [86]'s conjecture and constitutes the main pillar of the theoretical foundations of ME.

2.3. The balanced minimum evolution criterion

Rzhetsky and Nei [135] observed an undesirable aspect related to the optimal solution to (8): whenever $\mathbf{D}$ is a biased estimate of the true distance matrix from taxa, the optimal solution to (8) may be characterized by negative edge weights, which are void of biological meaning [60]. 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., [50, 62, 73, 103]) tried to find ways around the occurrence of negative edge weights, including exploring their possible biological interpretations [51] as well as the design of methods and algorithms to include in (8) 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

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

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

However, as observed in [60], these research efforts did not lead to satisfying conclusions. The publication of Pauplin's article in 2000 marked a turning point on this issue [121]. 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 11: 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

$$w_e = \frac{1}{2}\left[\lambda\left(\delta_{\Gamma_1|\Gamma_4} + \delta_{\Gamma_2|\Gamma_3}\right) + (1-\lambda)\left(\delta_{\Gamma_1|\Gamma_3} + \delta_{\Gamma_2|\Gamma_4}\right) - \left(\delta_{\Gamma_1|\Gamma_2} + \delta_{\Gamma_3|\Gamma_4}\right)\right] \tag{9}$$

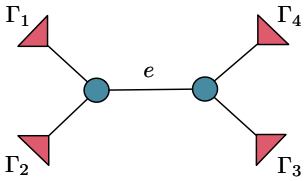


Figure 11: An internal edge e of a phylogeny T and the associated and four subtrees inducing a partition of Γ into the subsets Γ_1, Γ_2, Γ_3, and Γ_4, respectively.

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

$$\lambda = \frac{|\Gamma_1||\Gamma_3| + |\Gamma_2||\Gamma_4|}{|\Gamma_1 \cup \Gamma_2| \cdot |\Gamma_3 \cup \Gamma_4|} \tag{10}$$

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

$$\delta_{\Gamma_i|\Gamma_j} = \frac{1}{|\Gamma_i||\Gamma_j|} \sum_{\substack{r \in \Gamma_i \\ s \in \Gamma_j}} d_{rs}. \tag{11}$$

Similarly, in the case of an external edge, such as e in Figure 12, the corresponding weight can be computed as

$$w_e = \frac{1}{2}\left(\delta_{\{i\}|\Gamma_1} + \delta_{\{i\}|\Gamma_2} - \delta_{\Gamma_1|\Gamma_2}\right). \tag{12}$$

Equations (9) and (12) 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 (9). Now, Pauplin [121] observed that, the average distance (11) can be equivalently defined in the following recursive way:

$$\delta_{\Gamma_i|\Gamma_j} = \begin{cases} d_{ij} & \text{if } \Gamma_i = \{i\} \text{ and } \Gamma_j = \{j\}; \\ \frac{|\Gamma_p|}{|\Gamma_j|}\delta_{\Gamma_i|\Gamma_p} + \frac{|\Gamma_q|}{|\Gamma_j|}\delta_{\Gamma_i|\Gamma_q} & \text{otherwise,} \end{cases} \tag{13}$$

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 (13) by the following recursive formula

$$\delta^T_{\Gamma_i|\Gamma_j} = \begin{cases} d_{ij} & \text{if } \Gamma_i = \{i\} \text{ and } \Gamma_j = \{j\}; \\ \frac{1}{2}\delta^T_{\Gamma_i|\Gamma_p} + \frac{1}{2}\delta^T_{\Gamma_i|\Gamma_q} & \text{otherwise.} \end{cases} \tag{14}$$

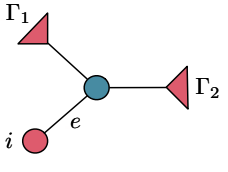


Figure 12: An external edge e of a phylogeny T and the associated and three subtrees inducing a partition of Γ into the subsets $\{i\}$, Γ_1, and Γ_2, respectively.

Pauplin's showed that under this definition of average distance, (9) becomes

$$w_e = \frac{1}{4}\left(\delta^T_{\Gamma_1|\Gamma_4} + \delta^T_{\Gamma_2|\Gamma_3} + \delta^T_{\Gamma_1|\Gamma_3} + \delta^T_{\Gamma_2|\Gamma_4}\right) - \frac{1}{2}\left(\delta^T_{\Gamma_1|\Gamma_2} + \delta^T_{\Gamma_3|\Gamma_4}\right), \tag{15}$$

(12) becomes

$$w_e = \frac{1}{2}\left(\delta^T_{\{i\}|\Gamma_1} + \delta_{\{i\}|\Gamma_2} - \delta^T_{\Gamma_1|\Gamma_2}\right), \tag{16}$$

and the length of the given phylogeny T becomes precisely (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 [60], 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*

$$L(T) = \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} \frac{d_{ij}}{2^{\tau_{ij}}}. \tag{17}$$

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 $T \in \mathcal{T}$. 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 showed in Desper and Gascuel [40], 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 [95, 157]. 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

$$\mathbf{w} = (\mathbf{Y}^t\mathbf{V}^{-1}\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{V}^{-1}\mathbf{D}^{\triangle}, \tag{18}$$

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 (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 [60] 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.

3. 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) [94] and the *Huffman Coding Problem* (HCP) [20, 23, 25, 28, 120]. In this section we review these aspects, by starting from Semple and Steel's interpretation of the BMEP length function (1) and by subsequently discussing the complexity aspects of the problem as well as the recent advances on its polyhedral combinatorics. Figure 13 graphically summarizes the main stages of this development.

3.1. 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 [144] to investigate the combinatorial nature of the BMEP objective

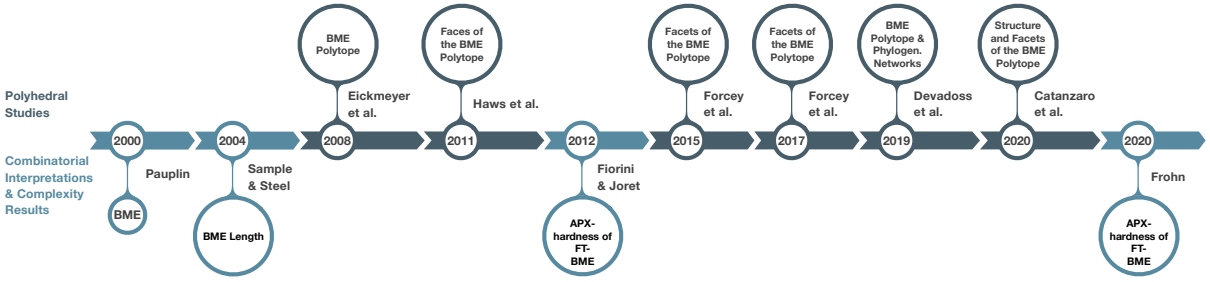


Figure 13: Graphical representation of the development of the BMEP's combinatorics, since Pauplin [121]'s article. The dark gray stages represent articles studying the polyhedral combinatorics of the BME polytope, including Eickmeyer et al. [46]'s conjectures, Haws et al. [71]'s faces, Forcey et al. [54]'s complete description of the BME polytope for $n = 5$, Forcey et al. [55]'s further facets of the BME polytope and combinatorial connections with the permutoassociahedron, Devadoss et al. [42]'s connections of the BME polytope with split-networks, and Catanzaro et al. [28]'s recent advances. The blue stages represent articles presenting the combinatorial interpretation of the BME length function [144] as well as complexity results [52, 57].

function. The authors started from Makarenkov and Leclerc [102]'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 [102]. As an example, consider the phylogeny T of six taxa in Γ shown in Figure 14(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 14(b), the circular order becomes $\pi_b = (1, 4, 5, 6, 3, 2)$. In both figures these circular orders are highlighted by the orange 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 orange dashed arrows depicted in Figure 14(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 black dashed arrows in Figure 14(a)), a possible circular order can be 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 $p_{\pi(i)\pi(i+1)}$, $i \in \{1, \dots, n\}$; (ii) that every edge that is incident to a taxon occurs in exactly two of the paths $p_{\pi(1)\pi(2)}, \dots, p_{\pi(n)\pi(1)}$;

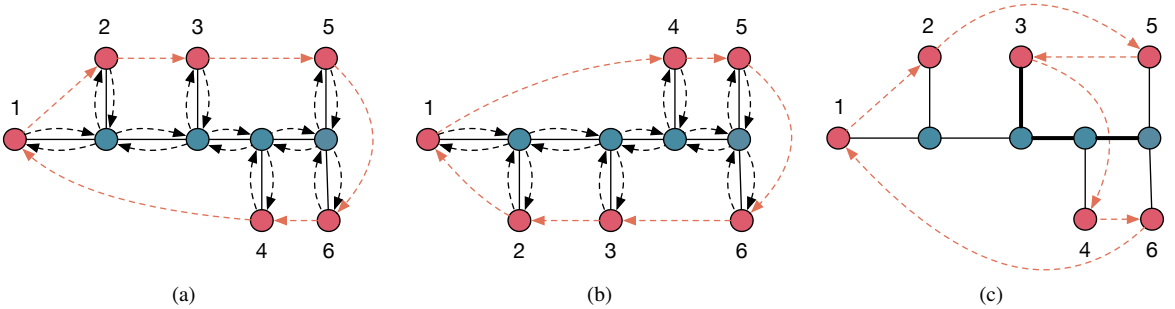


Figure 14: (a) By starting from taxon 1 and by following the orange dashed arrows, we obtain the permutation $\pi_a = (1, 2, 3, 5, 6, 4)$, which constitutes an example of a circular order on the given phylogeny T of six taxa. This permutation can be seen as derived from an Eulerian circuit on T (shown in black 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.

and (iii) that every edge that is incident only to internal vertices occurs in a positive and even number γ of paths $p_{\pi(1)\pi(2)}, \cdots, p_{\pi(n)\pi(1)}$. 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 14(a) whereas it is not satisfied by $\pi_c = (1, 2, 5, 3, 4, 6)$ in Figure 14(c), as the bold internal edges occur more than twice in $p_{\pi(1)\pi(2)}, \cdots, p_{\pi(n)\pi(1)}$. 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

$$d_{ij} = \sum_{e \in E(p_{ij})} w_e \quad \forall\, i, j \in \Gamma \tag{19}$$

then

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

In other words, if (19) 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:

$$L(T) = \frac{1}{|\Pi(T)|} \sum_{\pi \in \Pi(T)} \left(\frac{1}{2} \sum_{k=1}^{n} d_{\pi(k)\pi(k+1)} \right) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{\mu_{ij}}{2} d_{ij} \tag{20}$$

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 (20) becomes precisely (17).

3.2. Complexity results

Deciding the complexity of the BMEP remained an open problem for more than a decade since its introduction. Fiorini and Joret [52] 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) [58], 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* [81] 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.

Interestingly, Frohn [57] recently showed 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. Frohn [57]'s result on this particular restriction of the BMEP, known in the literature as the *Fixed-Tree BMEP* (FT-BMEP) [3], seems to indicate the presence of strong ties between the combinatorics of the BMEP and the *Quadratic Assignment Problem* (QAP) [32]. Exploring systematically these ties could provide new approaches to solution of the BMEP.

3.3. 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.

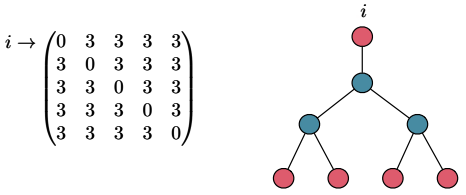


Figure 15: The i-th row of the τ matrix on the left encodes the binary tree rooted in taxon i on the right. Note how the entries of this specific row satisfy (24).

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 $\{2, \dots, n-1\}$, satisfying the following necessary conditions [28]:

$$\tau_{ii} = 0 \qquad \forall\, i \in \Gamma \qquad \text{(null diagonal path-lengths)} \tag{21}$$

$$\tau_{ij} = \tau_{ji} \qquad \forall\, i, j \in \Gamma : i < j \qquad \text{(symmetry equalities on path-lengths)} \tag{22}$$

$$\tau_{ij} + \tau_{jk} - \tau_{ik} \geq 2 \qquad \forall\, i, j, k \in \Gamma : i \neq j \neq k \qquad \text{(triangular inequalities on path-lengths).} \tag{23}$$

Moreover, because any path-length matrix $\tau \in \Theta$ must encode an unrooted binary tree, the rows of τ must satisfy *Kraft's equalities* [25, 28, 120]:

$$\sum_{j \in \Gamma \setminus \{i\}} 2^{-\tau_{ij}} = \frac{1}{2} \quad \forall\, i \in \Gamma. \tag{24}$$

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 15). In other words, each of the Kraft's equalities allows to capture 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 (24) 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 15 satisfy (24) 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 [21] and Section 5) and it is usually referred to as the equation of the *phylogenetic manifold* [25]:

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \tau_{ij} 2^{-\tau_{ij}} = (2n - 3). \tag{25}$$

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., $2n - 3$) [21].

Equations (21)-(25) 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 (25). 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 (24). 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 (23). 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 (23) can be further generalized by means of Buneman's *additive property* [14, 47, 161], 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:

$$\begin{cases} \tau_{ij} + \tau_{pq} + 2 \leq \tau_{iq} + \tau_{jp} = \tau_{ip} + \tau_{jq} \\ \tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq} \\ \tau_{ip} + \tau_{jq} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{iq} + \tau_{jp}. \end{cases} \tag{26}$$

It is easy to see that (23) is a restriction of (26) when i, j, p and q belong to Γ and e.g., $p = q$. Now, conditions (21)-(22) and (24)-(26) are necessary to characterize Θ, but proving (or disproving) the following conjecture still remains an open problem:

Conjecture 2. *Are conditions* (21)-(22) *and* (24)-(26) *sufficient to characterize* Θ.

Addressing this conjecture implies proving (or disproving) whether Buneman's additive property restricted just to quartets of taxa (rather than to the whole vertex set of a phylogeny) may imply the acyclicity of the combinatorial structure encoded by τ. If these conditions should prove sufficient, it would be possible to develop compact mixed integer programming formulations for the BMEP.

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 (24) and (25) can be linearized. Forcey et al. [54, 55], and Catanzaro et al. [28] 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

$$x_{ij} = 2^{n-1-\tau_{ij}} \quad \forall\, i, j \in \Gamma. \tag{27}$$

Specifically, by (27), 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 (27) induces the following $n(n+3)/2$ linear independent conditions on the entries of each matrix $\mathbf{X}$, which are analogous to conditions (21), (22), and (24) for the path-length matrices $\tau \in \Theta$:

$$x_{ii} = 2^{n-1} \qquad \forall\, i \in \Gamma \tag{28}$$

$$x_{ij} = x_{ji} \qquad \forall\, i, j \in \Gamma : i \neq j \tag{29}$$

$$\sum_{j \in \Gamma_i} x_{ij} = 2^{n-2} \qquad \forall\, i \in \Gamma. \tag{30}$$

Note, in particular, how Kraft's equalities becomes linear in (30). Catanzaro et al. [28] defined $\mathcal{X}$ as the set of the matrices $\mathbf{X}$ encoding the phylogenies of Γ via (27) 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

$$\binom{n}{2} - n. \tag{31}$$

From a historical perspective, the first use of Space($\mathcal{X}$) was introduced in the literature by Eickmeyer et al. [46] 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 [65, 138, 151] and Section 4.2). The authors considered the vectors

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

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 (24) for each $i \in \Gamma$, then the dimension of the BME polytope should have been equal to (31). Three years later, a coauthor of Eickmeyer et al. [46], namely Ruriko Yoshida, tried a second attack to the BME polytope in Haws et al. [71], by providing the description of some faces of the BME polytope. The proofs of Eickmeyer et al. [46]'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. [54, 55] and, more recently, of Catanzaro et al. [28]. In particular, Forcey et al. [54] 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 1. Remarkably, Forcey et al. [54] showed that some of these inequalities remain facet-defining even for generic n.

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. [54] they studied the connections between the BME polytope and the *Birkhoff polytope*, by showing their combinatorial equivalence for $n = 5$. In Forcey et al. [55], the authors focused on the connections between the BME polytope and the *permutoassociahedron* $\mathcal{KP}_n$ [8, 55, 84, 128] 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 1). This result, combined with the previous findings discussed in Forcey et al. [54], 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. [28] extended Forcey et al. [54, 55]'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. [54], the assistance of Polymake [66] was fundamental 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($\mathcal{X}$).

Facet Number	Facet-defining inequalities		Reference
1	$x_{ij} \geq 1$	For all distinct $i, j \in \Gamma$	Forcey et al. [54, 55]
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 $i_1, i_2, i_3 \in \Gamma$	Catanzaro et al. [28]
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	$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} \geq (3+\rho)2^{k-1}$	$k = \lfloor \frac{n}{3} \rfloor, \rho = n - 3k$	Catanzaro et al. [28]

Table 1
List of the known facet-defining inequalities of the BME polytope when assuming $n = 6$.

Recently, Devadoss et al. [42] extended the polyhedral combinatorics of the BME polytope also to the context of phylogenetic networks. In particular, starting from the results discussed in [55], the authors studied the combinatorics of the *level-1 split networks* [76], by showing that their convex hull shares families of facets with the BME polytope.

4. 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 enumeration 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 16 provides a graphical representation of this taxonomy and highlights the current state-of-the-art in the respective areas.

4.1. Exact approaches

The earliest exact approach to solution of the BMEP was proposed by Pardi [118] in 2009. Based on Semple and Steel [144]'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 [118] 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 [25].

An alternative exact solution approach for the BMEP was proposed by Aringhieri et al. [3] in 2011. The authors exploited the concept of *tree-isomorphism* [89] 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

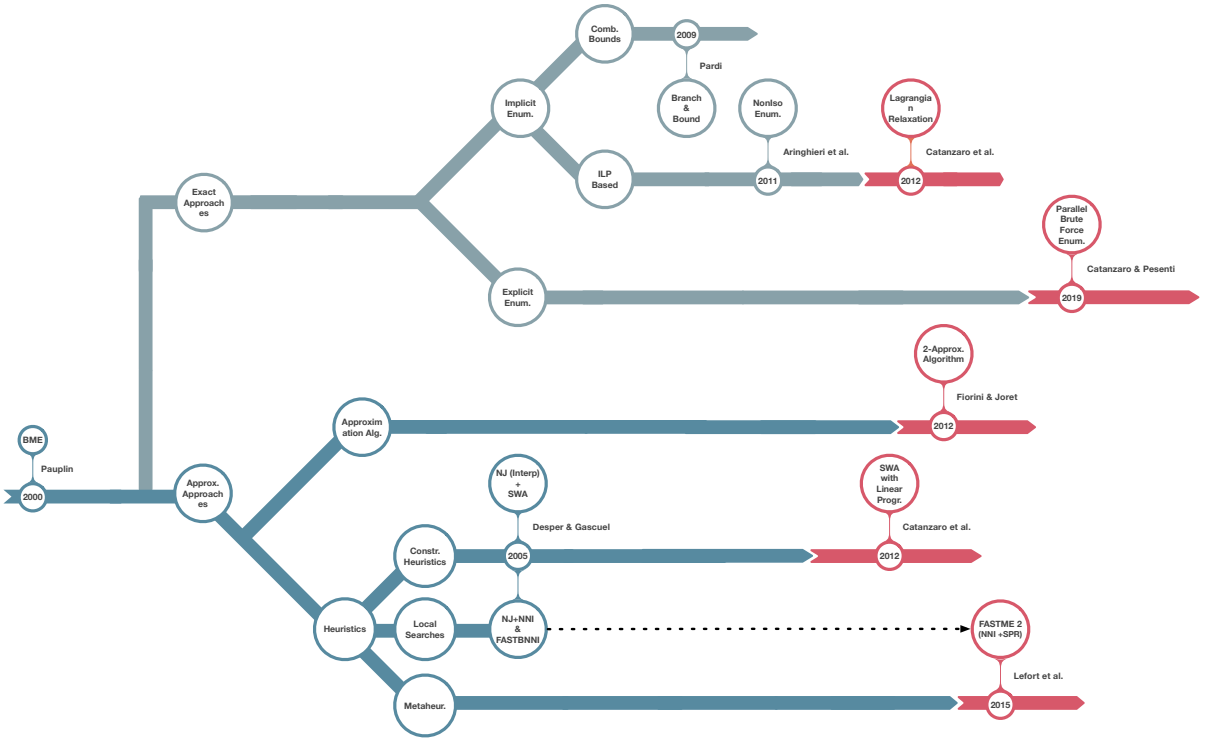


Figure 16: A graphical taxonomy of the solution approaches for the BMEP currently described in the literature. We can distinguish between exact (gray line) and approximate (blue line) solution approaches. The exact approaches can be classified into implicit and explicit enumerations; the implicit enumerations 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.

in $\mathcal{T}$ which is captured by the concept of *unlabeled phylogeny*. For example, the phylogeny in Figure 2 belongs to the equivalence class of the unlabeled phylogeny shown in Figure 17.

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. [3] 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 (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); 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. [3] to solve instances of the BMEP larger than Pardi [118] within the same time limit.

Catanzaro et al. [25] introduced in 2012 the first implicit enumeration algorithm for the BMEP based on math-

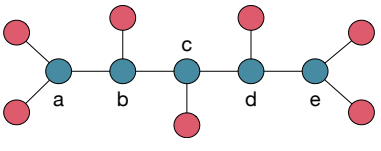


Figure 17: An unlabeled phylogeny with seven leaves.

Number of Taxa	Phylogenies Labeled	Unlabeled	Number of Taxa	Phylogenies 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
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.

ematical programming. The authors inspired to the strong combinatorial connections existing between the BMEP and the *Huffman Coding Problem* (HPC) [120] to identify some fundamental equalities describing UBTs (namely (24) and (25)). Then, they combined these equalities with Buneman's conditions (26) 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 $\{1, 2, 3, \dots, (n-1)\}$. 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. [25]'s ILP formulation of the BMEP reads as follows:

Formulation.

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

$$\text{s.t.} \sum_{k \in L} x_{ij}^k = 1 \qquad \forall\, i \neq j \in \Gamma \cup V \tag{32}$$

$$x_{ji}^k = x_{ij}^k \qquad \forall\, i < j \in \Gamma \cup V, k \in L \tag{33}$$

$$\sum_{\substack{j \in \Gamma \\ j \neq i}} \sum_{k \in L \setminus \{1\}} 2^{-k} x_{ij}^k = \frac{1}{2} \qquad \forall\, i \in \Gamma \tag{34}$$

$$\sum_{\substack{k \in L \\ k \neq 1}} k 2^{-k} \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} x_{ij}^k = (2n-3) \tag{35}$$

$$\sum_{k \in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k \in L} k(x_{iq}^k + x_{jt}^k) + (2n-2) y_{ij}^{qt} \qquad \forall\, i \neq j \neq q \neq t \in \Gamma \cup V \tag{36}$$

$$\sum_{k \in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k \in L} k(x_{it}^k + x_{jq}^k) + (2n-2)(1 - y_{ij}^{qt}) \qquad \forall\, i \neq j \neq q \neq t \in \Gamma \cup V \tag{37}$$

$$x_{ij}^1 = 0 \qquad \forall\, i \neq j \in \Gamma \tag{38}$$

$$\sum_{j \in V} x_{ij}^1 = 1 \qquad \forall\, i \in \Gamma \tag{39}$$

$$\sum_{\substack{j\in\Gamma\cup V\\ j\neq i}} x^1_{ij} = 3 \qquad \forall\, i \in V \tag{40}$$

$$x^1_{ij} + x^1_{il} + x^1_{lj} \leq 2 \qquad \forall\, i \neq j \neq l \in V \tag{41}$$

$$x^k_{ij} + 1 \geq x^{(k-1)}_{il} + x^1_{lj} \qquad \forall\, i \neq j \in \Gamma, l \in V, k \in L \setminus \{1, n-1\} \tag{42}$$

$$x^k_{ij} + x^{(k-2)}_{ij} + 1 \geq x^{(k-1)}_{il} + x^1_{lj} \qquad \forall\, i \neq j \neq l \in \Gamma \cup V, k \in L \setminus \{1, 2, n-1\} \tag{43}$$

$$x^k_{ij} \in \{0, 1\} \qquad \forall\, i, j \in \Gamma \cup V, k \in L \tag{44}$$

$$y^{qt}_{ij} \in \{0, 1\} \qquad \forall\, i \neq j \neq q \neq t \in \Gamma \cup V. \tag{45}$$

The convexity constraints (32) 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 (33) impose equation (22). Constraints (34) encode Kraft's equalities (24). Constraint (35) impose equation (25). Constraints (36) and (37) impose the Buneman's conditions (26). The subsequent set of constraints (38)-(43) impose that the set of edges encoded by variables x^1_{ij} form an UBT. Finally, the last two constraints impose the integrality of variables x and y. Catanzaro et al. [25]'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 [105]) 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 [118] 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 (32)-(35). 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's and Aringhieri et al. [3]'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. [25]'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. [25]'s algorithm is highly desirable. The use of specific tree-coding schemes mixed to ILP and parallel programming seems a promising way to ensure this result. Specifically, Catanzaro and Pesenti [26] 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 [16] 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 [26]'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.

4.2. 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.

4.2.1. Approximation algorithms

The first – and currently unique – approximation algorithm for the BMEP was proposed by Fiorini and Joret [52] 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 **D** is a metric matrix. This statement is based on some combinatorial relationships existing between the TSP, the *Minimum Spanning Tree Problem* (MST) [112], 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 **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 3.1) [52]. Moreover, the authors also recalled that the optimal solution to the TSP instance encoded by **D** is an upper bound on the value of the corresponding MST instance [58]. 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

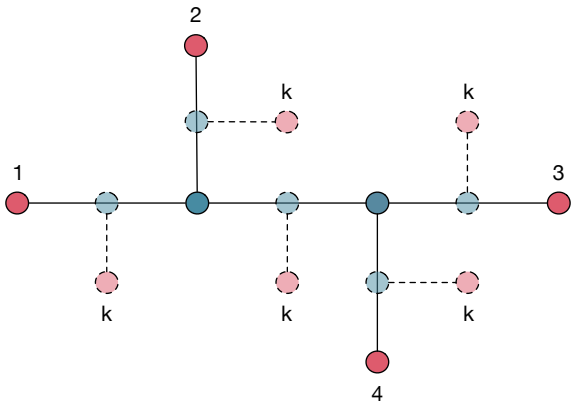


Figure 18: All of the possible insertions of the taxon k into the edges of a partial phylogeny of four taxa. The result of this insertion is a phylogeny of five taxa.

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.

4.2.2. 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 [51] and those based on the *agglomerative* paradigm [60]. 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 18). 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 [39] and Pardi [118] proposed the use of proxies derived from Semple and Steel [144]'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. [25], instead, proposed the use of a linear programming relaxation of Formulation 4.1 constituted by constraints (32)-(35). The insertion criterion proposed by Catanzaro et al. [25] is computationally more expensive than those proposed by Desper and Gascuel [39] and Pardi [118]; the BME length of the resulting solution, however, may prove shorter than those of Desper and Gascuel [39] and Pardi [118].

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 19). The agglomerative approach is well-known in the literature on phylogenetics. It was first introduced by Saitou and Nei [138] in 1987, under the name of *Neighbor Joining* (NJ) method [65, 138, 151], and subsequently declined in a myriad of variants, many of which may not be directly related to the BMEP [51, 60]. The validity of the NJ method (see Figure 20) 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 [60]). The reader interested in a historical perspective on this development may refer to Gascuel and Steel [65]'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 [51] and Figure 20). 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 [39]. 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 [39]. 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 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 [95], 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* [100] 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.

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

The statistical consistency results presented in Desper and Gascuel [40] 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 [4]'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

$$||D^T - D^{\Delta}||_{\infty} < \alpha w_{\min} \tag{46}$$

holds true, then A still reconstructs T when processing the input matrix D^{Δ}. As shown by Atteson [4], 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. [119] 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 [4]. 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 [11]. Based on empirical evidence, Pardi et al. [119] 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 [11]. 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 [64] 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 lowerbounds for the resulting safety radius.

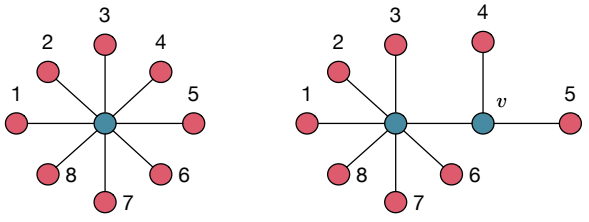


Figure 19: The phylogeny on the right is obtained from the star-tree on the left by "agglomerating" taxa 4 and 5, i.e., by connecting the two taxa to a common ancestor and by joining this ancestor to the center of the remaining star.

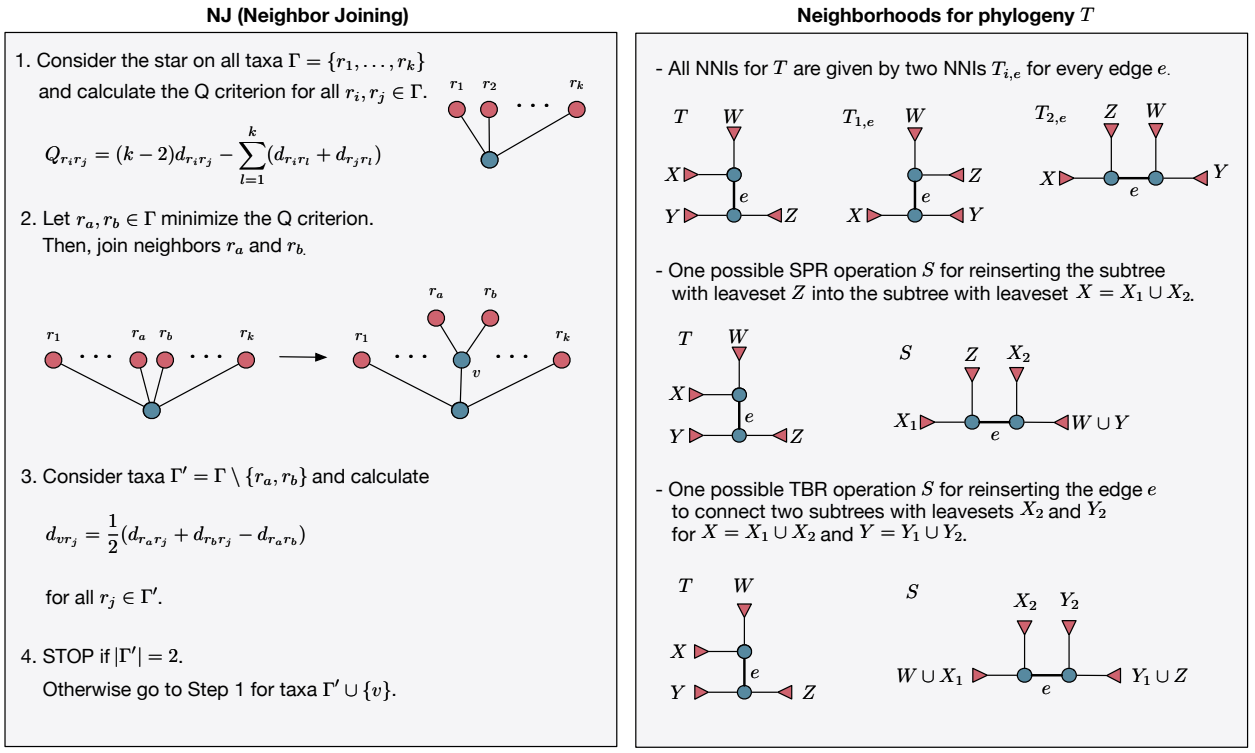


Figure 20: An example of two possible approximate approaches to solution of the BMEP. The figure on the left outlines the so called *Neighbor Joining* (NJ) method [65, 138, 151] i.e., a greedy constructive heuristic that starting from an infeasible solution to the problem (i.e., a star-tree) iteratively 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 [51, 60].

5. An information theory perspective on the BMEP

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 [28]. 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, as recently discussed in Catanzaro et al. [21], there exist further shared aspects between the two problems that enable a new original interpretation of the BMEP in terms of information theory. A key aspect of this interpretation 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. An analogous interpretation is possible also for the input matrix $\mathbf{D}$, provided an appropriate *rescaling* of $\mathbf{D}$ into a *doubly-stochastic matrix*, i.e., a square matrix having nonnegative entries and such that each of its rows and columns sums to one. As shown in Catanzaro et al. [21], the rescaling of $\mathbf{D}$ (i) is always possible, thanks to the results described in [13, 78, 85, 146]; (ii) is computable in polynomial-time; and (iii) does not alter the optimal solution to the BMEP. Thus, in the light of these rescaling operations, the length function of the BMEP can be expressed in terms of *cross-entropy* (or equivalently *Kullback-Leibler (KL) Divergence*) [139] associated to the probability distributions related to the path-lengths encoded by τ and the evolutionary distances encoded by (the rescaled version of) $\mathbf{D}$, respectively. In particular, let us denote by γ_i and δ_i the probability distributions encoded by the rescaled form of the i-th row vectors

of τ and $\mathbf{D}$, respectively. Then, the BME length function can be rewritten as [21]:

$$\begin{aligned} L(T) &= \frac{1}{2}\sum_{i\in\Gamma}\Big(\mathcal{H}(\gamma_i,\delta_i)-\mathcal{H}(\gamma_i)\Big)+Q \\ &= \frac{1}{2}\sum_{i\in\Gamma} D_{KL}(\gamma_i\|\delta_i)+Q \end{aligned} \tag{47}$$

where

$$\mathcal{H}(\gamma_i) = -\sum_{j\in\Gamma\setminus\{i\}}\gamma_{ij}\log_2(\gamma_{ij})$$

is the *information entropy* associated to γ_i [139]; $\mathcal{H}(\gamma_i,\delta_i)$ is the *cross-entropy* associated to γ_i and δ_i [10, 72], i.e.,

$$\mathcal{H}(\gamma_i,\delta_i) = -\sum_{j\in\Gamma_i}\gamma_{ij}\log_2(\delta_{ij}); \tag{48}$$

$D_{KL}(\gamma_i\|\delta_i)$ is the *Kullback-Leibler (KL) Divergence* [139] between a pair of probability distributions γ_i and δ_i, i.e.,

$$D_{KL}(p\|q) = \mathcal{H}(\gamma_i,\delta_i)-\mathcal{H}(\gamma); \tag{49}$$

and Q is a positive constant. Each term $\mathcal{H}(\gamma_i)$ encodes the information entropy of the evolutionary process "seen" from the perspective of taxon $i\in\Gamma$, i.e., the evolutionary process in which taxon i is the root from which all of the other taxa have been generated over time. The term $\mathcal{H}(\gamma_i,\delta_i)$, instead, encodes the error of approximating the evolutionary process "seen" from the perspective of taxon i with the process represented by the estimated evolutionary distances. The optimal solution to the BMEP, then, can be interpreted as a the *consensus tree* [51] 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 $\mathcal{H}(\gamma_i,\delta_i)-\mathcal{H}(\gamma_i)$.

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 [51] and Bayesian inference [74, 75], it is proven to be statistically consistent and by far the most accurate estimation model from among distance methods [60, 65]. 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, the information theory, and the optimization aspects of the BMEP.

In this article 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 [66] – 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 [54, 55]. We also know that there exist combinatorial connections between the BMEP and the Quadratic Assignment Problem [32]. 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. Furthermore, 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., [64]). Finally, we know that the BMEP is related to the Huffman Coding Problem [120] and this fact enables a particular interpretation of the BMEP in terms of

information theory. However, we do not know how to exploit this insight in order to extend this estimation model to phylogenetic networks. This fact, e.g., could accurately capture how viruses mutate in time, space, and in presence of multiple spreading events. 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.

Acknowledgment

The authors sincerely thank the anonymous referees for their valuable comments. The first author also acknowledge support from the Université Catholique de Louvain via the Fonds Spéciaux de Recherche (FSR) 2017-2021 and the Fondation Louvain via the research grant COALESCENS. The computational resources used during the development of this article have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI), funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11.

References

[1] Albert, V.A., 2005. Parsimony, phylogeny, and genomics. Oxford Bioscience, UK.
[2] Amiroch, S., Pradana, M.S., Irawan, M.I., Mukhlash, I., 2017. Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method. Journal of Physics: Conference Series 890, 012080.
[3] Aringhieri, R., Catanzaro, D., Di Summa, M., 2011. Optimal solutions for the balanced minimum evolution problem. Computers and Operations Research 38, 1845–1854.
[4] Atteson, K., 1999. The performance of the neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25, 251–278.
[5] Bader, D.A., Moret, B.M.E., Vawter, L., 2001. Industrial applications of high-performance computing for phylogeny reconstruction, in: SPIE ITCom: Commercial application for high-performance computing, SPIE, Bellingham, WA. pp. 159–168.
[6] Beerenwinkel, N., Schwarz, R.F., Gerstung, M., Markowetz, F., 2015. Cancer evolution: Mathematical models and computational inference. Systematic Biology 64, e1–e25.
[7] Beyer, W.A., Stein, M., Smith, T., Ulam, S., 1974. A molecular sequence metric and evolutionary trees. Mathematical Biosciences 19, 9–25.
[8] Billera, L.J., Holmes, S.P., Vogtmann, K., 2001. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics 27, 733–767.
[9] Blaisdell, B.E., 1986. 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.
[10] Böcherer, G., Amjad, R.A., 2014. Informational divergence and entropy rate on rooted trees with probabilities, in: IEEE International Symposium on Information Theory. IEEE Computer Society, Honolulu, HI, USA, pp. 176–180.
[11] Bordewich, M., Gascuel, O., Huber, K., Moulton, V., 2009. 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.
[12] Bowern, C., Atkinson, Q., 2012. Computational phylogenetics and the internal structure of pama-nyungan. Language 88, 817–845.
[13] Brualdi, R.A., 1974. The DAD theorem for arbitrary row sums, in: Proceedings of the American Mathematical Society, Providence, Rhode Island, USA. pp. 189–194.
[14] Buneman, P., 1971. The recovery of trees from measure of dissimilarities, in: Hodson, F.R., Kendall, D.G., Tautu, P. (Eds.), Archaeological and Historical Science. Edinburgh University Press, Edinburgh, UK, pp. 387–395.
[15] Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J., Fitch, W.M., 1999. Predicting the evolution of human influenza A. Science 286, 1921–1925.
[16] Caminiti, S., Finocchi, I., Petreschi, R., 2007. On coding labeled trees. Theoretical Computer Science 382, 97–108.
[17] Castro-Nallar, E., Pérez-Losada, M., Burton, G.F., Crandall, K.A., 2012. The evolution of HIV: Inferences using phylogenetics. Molecular Phylogenetics and Evolution 62, 777–792.
[18] Catanzaro, D., 2009. The minimum evolution problem: Overview and classification. Networks 53, 112–125.
[19] Catanzaro, D., 2011. Estimating phylogenies from molecular data, in: Bruni, R. (Ed.), Mathematical approaches to polymer sequence analysis and related problems. Springer, NY, pp. 149–176.
[20] Catanzaro, D., Aringhieri, R., di Summa, M., Pesenti, R., 2015. A branch-price-and-cut algorithm for the minimum evolution problem. European Journal of Operational Research 244, 753–765.
[21] Catanzaro, D., Frohn, M., Pesenti, R., 2020a. An information theory perspective on the balanced minimum evolution problem. Operations Research Letters 48, 362–367.
[22] Catanzaro, D., Gatto, L., Milinkovitch, M., 2006a. Assessing the applicability of the GTR nucleotide substitution model through simulations. Evolutionary Bioinformatics 2, 145–155.
[23] Catanzaro, D., Labbé, M., Pesenti, R., 2013a. The balanced minimum evolution problem under uncertain data. Discrete Applied Mathematics 161, 1789–1804.
[24] Catanzaro, D., Labbé, M., Pesenti, R., Salazar-Gonzáles, J.J., 2009. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. Networks 53, 126–140.
[25] Catanzaro, D., Labbé, M., Pesenti, R., Salazar-Gonzáles, J.J., 2012. The balanced minimum evolution problem. INFORMS Journal on Computing 24, 276–294.

[26] Catanzaro, D., Pesenti, R., 2019. Enumerating vertices of the balanced minimum evolution polytope. Computers and Operations Research 109, 209–217.
[27] Catanzaro, D., Pesenti, R., Milinkovitch, M., 2006b. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model. Bioinformatics 22, 708–715.
[28] Catanzaro, D., Pesenti, R., Wolsey, L.A., 2020b. On the Balanced Minimum Evolution polytope. Discrete Optimization 36, 1–33.
[29] Catanzaro, D., Ravi, R., Schwartz, R., 2013b. 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, 3.
[30] Catanzaro, D., Schackney, S.E., Schäffer, A.A., Schwartz, R., 2016. 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, 643–655.
[31] Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: Models and estimation procedures. American Journal of Human Genetics 19, 233–257.
[32] Çela, E., 1997. The Quadratic Assignment Problem. Kluwer Academic Publishers, Boston, MA.
[33] Chang, B.S.W., Donoghue, M.J., 2000. Recreating ancestral proteins. Trends in Ecology and Evolution 15, 109–114.
[34] Cheng, X., Du, D.Z., 2001. Steiner Trees in Industry. Kluwer Academic Publishers, Boston, MA.
[35] Chowdhury, S.A., Shackney, S.E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A.A., Schwartz, R., 2013. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. Bioinformatics 29, i189–i198.
[36] Cieslik, D., 1998. Steiner minimal trees. Springer, Boston, MA, USA.
[37] Darwin, C., 1964. On the Origin of Species. Harvard University Press, Cambridge, MA, USA.
[38] Denis, F., Gascuel, O., 2003. On the consistency of the minimum evolution principle of phylogenetic inference. Discrete Applied Mathematics 127, 66–77.
[39] Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. Journal of Computational Biology 9, 687–705.
[40] Desper, R., Gascuel, O., 2004. 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, 587–598.
[41] Desper, R., Gascuel, O., 2005. The minimum-evolution distance-based approach to phylogeny inference. Mathematics of Evolution and Phylogeny , 1–32.
[42] Devadoss, S.L., Durell, C., Forcey, S., 2019. Split network polytopes and network spaces, in: The 31st international conference on Formal Power Series and Algebraic Combinatorics, Séminaire Lotharingien de Combinatoire. p. 68.
[43] Du, D.Z., Hu, X., 2008. Steiner tree problems in computer communication networks. World Scientific Publishing Company, Singapore.
[44] Du, D.Z., Smith, J.M., Rubinstein, J.H., 2000. Advances in Steiner trees. Kluwer Academic Publishers, Boston, MA, USA.
[45] Duellman, W.E., Marion, A.B., Hedges, S.B., 2016. Phylogenetics, classification, and biogeography of the treefrogs (Amphibia: Anura: Arboranae). Zootaxa 4104, 1–109.
[46] Eickmeyer, K., Huggins, P., Pachter, L., Yoshida, R., 2008. On the optimality of the neighbor-joining algorithm. Algorithms for Molecular Biology 3, 5.
[47] Erdös, P.L., Steel, M.A., Székély, L.A., Warnow, T., 1999. A few logs suffice to build (almost) all trees: Part I. Random Structures and Algorithms 14, 153–184.
[48] Farris, J.S., 1970. Methods for computing wagner trees. Systematic Biology 19, 83–92.
[49] Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27, 401–410.
[50] Felsenstein, J., 1997. An alternating least-squares approach to inferring phylogenies from pairwise distances. Systematic Biology 46, 101–111.
[51] Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.
[52] Fiorini, S., Joret, G., 2012. Approximating the balanced minimum evolution problem. Operations Research Letters 40, 31–35.
[53] Fitch, W.M., 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20, 406–416.
[54] Forcey, S., Keefe, L., Sands, W., 2015. Facets of the balanced minimal evolution polytope. Journal of Mathematical Biology 73, 447–468.
[55] Forcey, S., Keefe, L., Sands, W., 2017. Split-facets for balanced minimal evolution polytopes and the permutoassociahedron. Bulletin of Mathematical Biology, in press 79, 975–994.
[56] Forster, P., Forster, L., Renfrew, C., Forster, M., 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences of the USA 117, 9241–9243.
[57] Frohn, M., 2020. On the approximability of the fixed-tree balanced minimum evolution problem. To appear in Optimization Letters .
[58] Garey, M.R., Johnson, D.S., 2003. Computers and Intractability: A guide to the theory of NP-Completeness. Freeman, New York, NY.
[59] Gascuel, O., 2000. On the optimization principle in phylogenetic analysis and the minimum evolution criterion. Journal of Classification 19, 67–69.
[60] Gascuel, O., 2005. Mathematics of evolution and phylogeny. Oxford University Press, New York, NY.
[61] Gascuel, O., Bryant, D., Denis, F., 2001. Strengths and limitations of the minimum evolution principle. Systematic Biology 50, 621–627.
[62] Gascuel, O., Levy, D., 1996. A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. Journal of Classification 13, 129–155.
[63] Gascuel, O., McKenzie, A., 2004. Performance analysis of hierarchical clustering algorithms. Journal of Classification 21, 3–18.
[64] Gascuel, O., Steel, M., 2016. A 'stochastic safety radius' for distance-based tree reconstruction. Algorithmica 74, 1386–1403.
[65] Gascuel, O., Steel, M.A., 2006. Neighbor-joining revealed. Molecular Biology and Evolution 23, 1997–2000.
[66] Gawrilow, E., Joswig, M., 2000. Polymake: A framework for analyzing convex polytopes, in: Kalai, G., Ziegler, G.M. (Eds.), Polytopes - combinatorics and computation. Birkhäuser, pp. 43–73.
[67] Ge, X.Y., Li, J.L., Yang, X.L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., Peng, C., Zhang, Y.J., Luo, C.M., Tan,

B., Wang, N., Zhu, Y., Crameri, G., Zhang, S.Y., Wang, L.F., Daszak, P., Shi, Z.L., 2013. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 503, 535–538.
[68] Gusfield, D., 1984. The Steiner Tree Problem in Phylogeny. Technical Report 334. Yale University, New Haven, CT.
[69] Harvey, P.H., Brown, A.J.L., Smith, J.M., Nee, S., 1996. New uses for new phylogenies. Oxford University Press, Oxford, UK.
[70] Hasegawa, M., Kishino, H., Yano, T., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17, 368–376.
[71] Haws, D.C., Hodge, T.L., Yoshida, R., 2011. Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. Bulletin of Mathematical Biology 73, 2627–2648.
[72] Hirschler, T., Woess, W., 2018. Comparing entropy rates on finite and infinite rooted trees. IEEE Transactions on Information Theory 64, 5570–5580.
[73] Hubert, L.J., Arabie, P., 1995. Iterative projection strategies for the least-squares fitting of tree structures to proximity data. British Journal of Mathematical and Statistical Psychology 48, 281–317.
[74] Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of bayesian inference of phylogeny. Systematic Biology 51, 673–688.
[75] Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.
[76] Huson, D.H., Rupp, R., Scornavacca, C., 2011. Phylogenetic networks. Cambridge University Press, Cambridge, UK.
[77] Hwang, F.K., Richards, D.S., Winter, P., 1992. The Steiner tree problem. North-Holland, Amsterdam, The Netherlands.
[78] Idel, M., 2016. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. arXiv: 1609.06349.
[79] Jäger, G., 2018. Global-scale phylogenetic linguistic inference from lexical resources. Scientific Data 5, 180189.
[80] Johnson, D.S., Lenstra, J.K., Kan, A.H.G.R., 1978. The complexity of the network design problem. Networks 8, 279–285.
[81] Jordan, C., 1869. Sur les assemblages des lignes. Journal für die reine und angewandte Mathematik 70, 185–190.
[82] Jukes, T.H., Cantor, C., 1969. Evolution of protein molecules, in: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, NY, pp. 21–123.
[83] Kadam, S., Vuong, T.D., Qiu, D., Meinhardt, C.G., Song, L., Deshmukh, R., Patil, G., Wan, J., Valliyodan, B., Scaboo, A.M., Shannon, J.G., Nguyen, H.T., 2016. Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. Plant Science 242, 342–350.
[84] Kapranov, M.M., 1993. The permutoassociahedron, Mac Lane's coherence theorem and asymptotic zones for the KZ equation. Journal of Pure and Applied Algebra 85, 119–142.
[85] Khachiyan, L., 1996. Diagonal matrix scaling is $\mathcal{NP}$-hard. Linear algebra and its applications 234, 173–179.
[86] Kidd, K.K., Sgaramella-Zonta, L.A., 1971. Phylogenetic analysis: Concepts and methods. American Journal of Human Genetics 23, 235–252.
[87] Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16, 111–120.
[88] Klung, W.S., Cummings, M.R., Spencer, C.A., Palladino, M.A., Killian, D., 2019. Concepts of Genetics. Pearson, NY.
[89] Kreher, D.L., Stinson, D.R., 1999. Combinatorial algorithms: Generation, enumeration, and search. CRC Press, Boca Raton, FL.
[90] Kress, W.J., Erickson, D.L., Swenson, N.G., Thompson, J., Uriarte, M., Zimmermann, J.K., 2010. 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, e15409.
[91] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.O., Aparicio, S., Baaijens, J., Balvert, M., de Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., de Ridder, J., Saliba, A.E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A., 2020. Eleven grand challenges in single-cell data science. Genome Biology 21, 1–35.
[92] Lai, C.C., Shih, T.P., Ko, W.C., Tang, H.J., Hsue, P.R., 2020. 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, 105924.
[93] Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20, 86–93.
[94] Lawler, E., Lenstra, J., Kan, A.R., Shmoys, D., 1985. The Traveling Salesman Problem: A guided tour of combinatorial optimization. John Wiley and Sons Ltd., New York, NY.
[95] Lefort, V., Desper, R., Gascuel, O., 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Molecular Biology and Evolution 32, 2798–2800.
[96] Leibold, M.A., Economo, E.P., Peres-Neto, P., 2010. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. Ecology Letters 13, 1290–1299.
[97] Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44, 99–121.
[98] Lemoine, F., Blassel, L., Voznica, J., Gascuel, O., 2020. COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM. Bioinformatics .
[99] Li, X., Mao, Y., 2016. Generalized connectivity of graphs. Springer, Boston, MA, USA.
[100] Lourenço, H.R., Martin, O., Stützle, T., 2003. Iterated local search, in: Glover, F., Kochenberger, G.A. (Eds.), Handbook of Metaheuristics. Springer, Boston, MA, USA. volume 57 of *International Series in Operations Research & Management Science*, pp. 320–353.
[101] Lu, C.L., Tang, C.Y., Lee, R.C.T., 2003. The full Steiner tree problem. Theoretical Computer Science 306, 55–67.
[102] Makarenkov, V., Leclerc, B., 1997. Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees,

in: Mirkin, B., McMorris, F., Roberts, F., Rzhetsky, A. (Eds.), Mathematical hierarchies and Biology. American Mathematical Society, Providence, RI. volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 183–208.

[103] Makarenkov, V., Leclerc, B., 1999. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. Journal of Classification 16, 3–26.

[104] Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C., Roper, R.L., 2003. The genome sequence of the SARS-associated coronavirus. Science 300, 1399–1404.

[105] Martin, R.K., 1999. Large Scale Linear and Integer Optimization: A Unified Approach. Springer-Verlag, New York, NY.

[106] Mavian, C., Pond, S.K., Marini, S., Magalis, B.R., Vandamme, A.M., Dellicour, S., Scarpino, S.V., Houldcroft, C., Villabona-Arenas, J., Paisie, T.K., Trovão, N.S., Boucher, C., Zhang, Y., Scheuermann, R.H., Gascuel, O., Lam, T.T.Y., Suchard, M.A., Abecasis, A., Wilkinson, E., de Oliveira, T., Bento, A.I., Schmidt, H.A., Martin, D., Hadfield, J., Faria, N., Grubaugh, D.N., Neher, R.A., Baele, G., Lemey, P., Stadler, T., Albert, J., Crandall, K.A., Leitner, T., Stamatakis, A., Prosperi, M., Salemi, M., 2020. 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, 12522–12523.

[107] McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66, 526–538.

[108] McGuire, J.A., Witt, C.C., Remsen Jr., J.V., Corl, A., Rabosky, D.L., Altshuler, D.L., Dudley, R., 2014. Molecular phylogenetics and the diversification of hummingbirds. Current Biology 24, 910–916.

[109] Misra, N., Blelloch, G.E., Ravi, R., Schwartz, R., 2011. Generalized buneman pruning for inferring the most parsimonious multi-state phylogeny. Journal of Computational Biology 18, 445–57.

[110] Myers, M.A., Satas, G., Raphael, B.J., 2019. Calder: Inferring phylogenetic trees from longitudinal tumor samples. Cell Systems 8, 514–522.

[111] Nei, M., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, UK.

[112] Nemhauser, G.L., Wolsey, L.A., 1999. Integer and combinatorial optimization. Wiley-Interscience, New York, NY.

[113] Ng, H.C., Liu, S., Luk, W., 2017. Reconfigurable acceleration of genetic sequence alignment: A survey of two decades of efforts, in: Santambrogio, M., Göhringer, D., Stroobandt, D., Mentens, N., Nurmi, J. (Eds.), 2017 27th International Conference on Field Programmable Logic and Applications (FPL), pp. 1–8.

[114] Notredame, C., 2004. Recent progress in multiple sequence alignment: A survey. Pharmacogenomics 3, 131–144.

[115] Ou, C.Y., Ciesielski, C.A., Myers, G., Bandea, C.I., Luo, C.C., Korber, B.T.M., Mullins, J.I., Schochetman, G., Berkelman, R.L., Economou, A.N., Witte, J.J., Furman, L.J., Satten, G.A., Maclnnes, K.A., Curran, J.W., Jaffe, H.W., 1992. Molecular epidemiology of HIV transmission in a dental practice. Science 256, 1165–1171.

[116] Pachter, L., Sturmfels, B., 2007. The mathematics of phylogenomics. SIAM Review 49, 3–31.

[117] Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford, UK.

[118] Pardi, F., 2009. Algorithms on Phylogenetic Trees. Ph.D. thesis. University of Cambridge, UK.

[119] Pardi, F., Guillemot, S., Gascuel, O., 2010. Robustness of phylogenetic inference based on minimum evolution. Bulletin of Mathematical Biology 72, 1820–1839.

[120] Parker, D.S., Ram, P., 1996. The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees. SIAM Journal on Computing 28, 1875–1905.

[121] Pauplin, Y., 2000. Direct calculation of a tree length using a distance matrix. Journal of Molecular Evolution 51, 41–47.

[122] Pennington, G., Smith, C.A., Shackney, S., Schwartz, R., 2006. Reconstructing tumor phylogenies from heterogeneous single-cell data. Journal of Bioinformatics and Computational Biology 5, 407–427.

[123] Perovic, V.R., 2013. Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses. Journal of Mathematical Chemistry 51, 2238–2255.

[124] Poon, A.F.Y., Joy, J.B., Woods, C.K., Shurgold, S., Colley, G., Brumme, C.J., Hogg, R.S., Montaner, J.S.G., Harrigan, P.R., 2015. 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, 926–935.

[125] Pop, P.C., 2012. Generalized network design problems: Modeling and optimization. De Gruyter, Berlin, Germany.

[126] Popper, K., 2002. The logic of scientific discovery. Routledge, London, UK.

[127] Prömel, H.J., Steger, A., 2002. The Steiner tree problem: A tour through graphs, algorithms, and complexity. Vieweg+Teubner Verlag, Berlin.

[128] Reiner, V., Ziegler, G.M., 1994. Coxeter-associahedra. Zuse Institute Berlin (ZIB), Berlin, Takustrasse 7, 14195, Germany 41, 364–393.

[129] Riester, M., Attolini, C.S.O., Downey, R.J., Singer, S., Michor, F., 2010a. A differentiation-based phylogeny of cancer subtypes. PLoS Computational Biology 6, e1000777.

[130] Riester, M., Attolini, C.S.O., Downey, R.J., Singer, S., Michor, F., 2010b. A differentiation-based phylogeny of cancer subtypes. PLoS Computational Biology 6, e100077.

[131] Rodriguez, F., Oliver, J.L., Marin, A., Medina, J.R., 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology 142, 485–501.

[132] Rosenberg, M.S., 2011. Sequence Alignment: Methods, Models, Concepts, and Strategies. University of California Press, LA.

[133] Ross, H.A., Rodrigo, A.G., 2002. Immune-mediated positive selection drives Human Immunodeficency Virus type 1 molecular variation and predicts disease duration. Journal of Virology 76, 11715–11720.

[134] Rzhetsky, A., Nei, M., 1992. A simple method for estimating and testing minimum evolution trees. Computer Applications in the Biosciences 10, 409–412.
[135] Rzhetsky, A., Nei, M., 1993. Theoretical foundations of the minimum evolution method of phylogenetic inference. Molecular Biology and Evolution 10, 1073–1095.
[136] Rzhetsky, A., Nei, M., 1994. METREE: A program package for inferring and testing minimum evolution trees. Computer Applications in the Biosciences 10, 409–412.
[137] Saitou, N., Imanishi, T., 1989. 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.
[138] Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425.
[139] Sayood, K., 2017. Introduction to Data Compression. 5th ed., Morgan Kaufmann, San Francisco, CA.
[140] Scheiner, S.M., Mindell, D.P., 2020. The theory of evolution: Principles, concepts and assumptions. University of Chicago Press, Chicago, IL, USA.
[141] Schulmeister, S., 2004. Inconsistency of maximum parsimony revisited. Systematic Biology 53, 521–528.
[142] Schwartz, R., 2019. Computational models for cancer phylogenetics, in: Warnow, T. (Ed.), Bioinformatics and Phylogenetics. Computational Biology. Springer, Cham. volume 29, pp. 243–275.
[143] Semple, C., Steel, M.A., 2003. Phylogenetics. Oxford University Press, New York, NY.
[144] Semple, C., Steel, M.A., 2004. Cyclic permutations and evolutionary trees. Advances in Applied Mathematics 32, 669–680.
[145] Simonsen, M., Mailund, T., Pedersen, C.N.S., 2008. Rapid neighbour joining. Lecture Notes in Bioinformatics 5251, 113–122.
[146] Sinkhorn, R., 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. Annals of Mathematical Statistics 35, 876–879.
[147] Sridhar, S., Dhamdhere, K., Blelloch, G.E., Halperin, E., Ravi, R., Schwartz, R., 2007. Algorithms for efficient near-perfect phylogenetic tree reconstruction in theory and practice. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4, 561–571.
[148] Sridhar, S., Lam, F., Blelloch, G.E., Ravi, R., Schwartz, R., 2008. Mixed integer linear programming for maximum parsimony phylogeny inference. IEEE/ACM Transactions on Computational Biology and Bioinformatics 5, 323–331.
[149] Stuart, G.W., Moffett, K., Baker, S., 2002a. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18, 100–108.
[150] Stuart, G.W., Moffett, K., Leader, J.J., 2002b. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution 19, 554–562.
[151] Studier, J.A., Keppler, K.J., 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Molecular Biology and Evolution 5, 729–731.
[152] Subramanian, A., Shackney, S., Schwartz, R., 2013. Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10, 1422–1431.
[153] Tavare, S., 1987. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17, 57–86.
[154] Valiente-Banuet, A., Verdú, M., 2013. Plant facilitation and phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44, 347–366.
[155] Vinga, S., 2014. Information theory applications for biological sequence analysis. Briefings in Bioinformatics 15, 376–389.
[156] Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison - A review. Bioinformatics 19, 513–523.
[157] Vinh, L.S., von Haeseler, A., 2005. Shortest triplet clustering: Reconstructing large phylogenies using representative sets. BMC Bioinformatics 6, 1–14.
[158] Volkenstein, M.V., Livshits, M.A., 1989. Speciation and bifurcations. Biosystems 23, 1–5.
[159] Waddell, P.J., Steel, M.A., 1997. 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.
[160] Washburne, A.D., Morton, J.T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A.M., Knight, R., 2018. Methods for phylogenetic analysis of microbiome data. Nature Microbiology 3, 652–661.
[161] Waterman, M.S., Smith, T.F., Singh, M., Beyer, W.A., 1977. Additive evolutionary trees. Journal of Theoretical Biology 64, 199–213.
[162] Wu, B.Y., Chao, K.M., 2004. Spanning trees and optimization problems. Chapman and Hall/CRC, Boca Raton, FL.
[163] Yang, Z., 1994. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39, 105–111.
[164] Yang, Z., 2014. Molecular Evolution: A Statistical Approach. Oxford University Press, Oxford, UK.
[165] Zhou, P., Yang, X.L., Shi, Z.L., *et al.*, 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.
[166] Zielezinski, A., Vinga, S., Almeida, J., Karlowski, W.M., 2017. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology 18, 186.