A branch-price-and-cut algorithm for the minimum evolution problem

Accepted Manuscript

A Branch-Price-and-Cut Algorithm for the Minimum Evolution Problem

Daniele Catanzaro, Roberto Aringhieri, Marco Di Summa, Raffaele Pesenti

PII: S0377-2217(15)00120-4
DOI: 10.1016/j.ejor.2015.02.019
Reference: EOR 12788

To appear in: *European Journal of Operational Research*

Received date: 13 October 2013
Revised date: 3 February 2015
Accepted date: 10 February 2015

Please cite this article as: Daniele Catanzaro, Roberto Aringhieri, Marco Di Summa, Raffaele Pesenti, A Branch-Price-and-Cut Algorithm for the Minimum Evolution Problem, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.02.019

Highlights

- We study the Minimum Evolution problem (MEP), which arises in Computational Biology.
- We study the polyhedral combinatorics of the MEP.
- We present an exact solution approach for the MEP.
- We study relationships between the MEP and the Balanced Minimum Evolution Problem.
- We show the statistical consistency of the MEP.

A Branch-Price-and-Cut Algorithm for the Minimum Evolution Problem

Daniele Catanzaro* Roberto Aringhieri† Marco Di Summa‡ Raffaele Pesenti§

February 14, 2015

Abstract

We investigate the *Minimum Evolution Problem* (MEP), an $\mathcal{NP}$-hard network design problem arising from computational biology. The MEP consists in finding a weighted unrooted binary tree having n leaves, minimal length, and such that the sum of the edge weights belonging to the unique path between each pair of leaves is greater than or equal to a prescribed value. We study the polyhedral combinatorics of the MEP and investigate its relationships with the Balanced Minimum Evolution Problem. We develop an exact solution approach for the MEP based on a nontrivial combination of a parallel branch-price-and-cut scheme and a non-isomorphic enumeration of all possible solutions to the problem. Computational experiments show that the new solution approach outperforms the best mixed integer linear programming formulation for the MEP currently described in the literature. Our results give a perspective on the combinatorics of the MEP and suggest new directions for the development of future exact solution approaches that may turn out useful in practical applications. We also show that the MEP is statistically consistent.

Keywords: network design, combinatorial inequalities, branch-price-and-cut, mixed integer programming, symmetry breaking, tree encoding, tree isomorphism, computational biology.

1 Introduction

Molecular phylogenetics studies the hierarchical evolutionary relationships among species, or *taxa*, by means of molecular data such as DNA, RNA, amino acids, or codon sequences. These relationships are usually described through a weighted tree, called a *phylogeny* (see Figure 1), whose leaves represent the observed taxa, internal vertices represent the intermediate ancestors, edges represent the estimated evolutionary relationships, and edge weights represent evolutionary processes (e.g., evolutionary rates, expected number of mutations, and so on) between pairs of taxa (Felsenstein, 2004). As usual in most of the literature, in this paper the leaves of a phylogeny are labeled in order to identify the given taxa, whereas the internal vertices are unlabeled.

Phylogenies provide fundamental information in the analysis of many fine-scale genetic data. For this reason, the use of molecular phylogenetics has become more and more frequent (and sometimes indispensable) in a multitude of research fields such as systematics, medical research, drug discovery, epidemiology, and population dynamics (Pachter and Sturmfels, 2007). For example, the use of molecular phylogenetics was of considerable assistance in predicting evolution of human influenza A (Bush et al., 1999), understanding the relationships between the virulence and the genetic evolution of HIV (Ross and Rodrigo, 2002; Ou et al., 1992), identifying emerging viruses such as SARS (Marra et al., 2003), recreating and investigating ancestral proteins (Chang and Donoghue, 2000), designing neuropeptides causing smooth muscle contraction (Bader et al., 2001), and relating geographic patterns to macroevolutionary processes (Harvey et al., 1996). More recently, phylogenies have been used to study the evolutionary relationships of the genetic factors

*Louvain School of Management, Université Catholique de Louvain, Chausse de Binche 151, bte M1.01.01, 7000 Mons, Belgium. Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Voie du Roman Pays 34, L1.03.01, B-1348, Louvain-la-Neuve, Belgium.

†Dipartimento di Informatica, Università di Torino, Corso Svizzera 135, I-10149 Torino, Italy.

‡Dipartimento di Matematica, Università degli Studi di Padova, Via Trieste 63, I-35121 Padova, Italy.

§Department of Management, Università Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venezia, Italy.

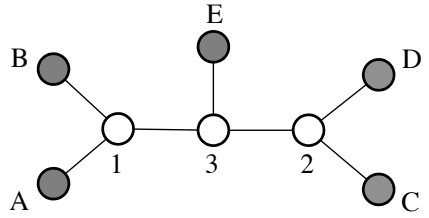

Figure 1: A phylogeny with five taxa (A, B, C, D, E) and three internal vertices. Though the internal vertices of a phylogeny are unlabeled, here we attach numbers to them for ease of reference. Edge weights are not shown.

involved in common human diseases (Pennington et al., 2006; Sridhar et al., 2007, 2008; Misra et al., 2011). Similarly, phylogenies have been also employed to reconstruct a plausible progression of carcinomas over time (Riester et al., 2010; Subramanian et al., 2013) by using, in particular, single-cell sampled data from affected individuals (Catanzaro et al., 2013; Chowdhury et al., 2013). In this context, phylogenies allowed the classification of tumor cells in subfamilies characterized by specific evolutionary traits. This classification might enable a better understanding of cellular atypia over time and, on a long run, suggest new therapeutical approaches for tumor pathologies. A recent survey concerning the practical uses of molecular phylogenetics can be found in Beerenwinkel et al. (2015).

The internal vertices of a phylogeny of n taxa represent speciation events occurred throughout evolution of the observed taxa and are usually constrained to have degree three. This constraint has both a biological foundation (see Felsenstein, 2004) and a mathematical motivation, as it proves helpful when formalizing the evolutionary process of taxa. In fact, it does not introduce oversimplifications, as any m-ary tree can be transformed into a phylogeny by adding "dummy" vertices and edges, e.g., see Figure 2. Moreover, as observed in Felsenstein (2004), the degree constraint helps both in quantifying a-priori the number of edges and internal vertices of a phylogeny ($2n-3$ and $n-2$, respectively), otherwise hard to determine, and in counting the overall number of possible phylogenies for a set of n taxa, (i.e., $(2n-5)!! = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-5)$). The large number of possible phylogenies for a set of n taxa entails the use of an estimation criterion to select a phylogeny from among plausible alternatives.

The literature on molecular phylogenetics proposes a number of possible criteria to estimate phylogenies from molecular data. Apart from particular cases (e.g., Bayesian Inference, Huelsenbeck et al., 2001, 2002), these criteria can be quantified and expressed in terms of objective functions, giving rise to families of optimization problems, whose general paradigm can be stated as follows:

Problem 1 (The Phylogeny Estimation Problem (PEP))**.** *Given a set Γ of n taxa, find a phylogeny T^* that solves the problem*

$$\begin{array}{rl} \textit{optimize} & \Lambda(T) \\ \textit{s.t.} & \Omega(\Gamma, T) \geq 0 \\ & T \in \mathcal{T}, \end{array}$$

where $\mathcal{T}$ is the set of all possible phylogenies of Γ, $\Lambda : \mathcal{T} \to \mathbb{R}$ is a function modeling the selected criterion of phylogeny estimation, and $\Omega : \Gamma \times \mathcal{T} \to \mathbb{R}^m$, for some $m \geq 1$, is a function correlating the set Γ to a phylogeny T.

A specific PEP is completely characterized by defining the functions Λ and Ω. A phylogeny T^* that optimizes Λ and satisfies the constraints described by Ω is referred to as *optimal*.

A possible version of PEP is *Minimum Evolution* (ME) (see Felsenstein, 2004, p. 161) which consists in minimizing the length of a phylogeny (i.e., the sum of its edge weights) with respect to a given measure of dissimilarity among taxa (see Page and Holmes, 1998; Felsenstein, 2004; Gascuel, 2005). In the context of ME, a phylogeny of Γ is defined to be optimal if it satisfies the following requirements:

(i) it has the shortest length, i.e., the minimum sum of edge weights;

(ii) it has nonnegative edge weights and, for each pair of distinct taxa $i, j \in \Gamma$, the sum of the weights of the edges belonging to the (unique) path from i to j in T^* is not smaller than the given measure of the dissimilarity between i and j.

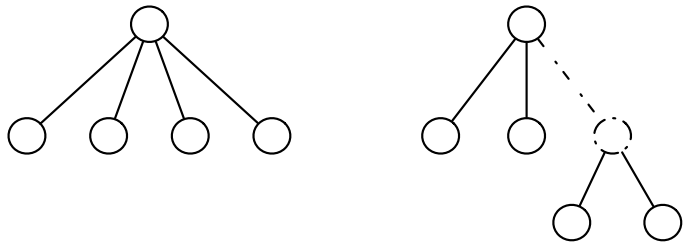

Figure 2: The 4-ary tree (on the left) can be transformed into a phylogeny by adding a dummy vertex and a dummy edge (dashed, on the right).

We refer the interested reader to Farach et al. (1995) for a biological justification of constraint (ii). Here we just observe that if only (i) and the nonnegativity of the weights were imposed, then the problem would be trivial, as all weights would be set to zero.

A number of versions of ME have been proposed in the literature, mainly differing from one another by the way in which the edge weights are estimated. Examples include Kidd and Sgaramella-Zonta (1971); Beyer et al. (1974); Rzhetsky and Nei (1992, 1993); Felsenstein (2004); Gascuel (2005). One of the earliest edge weight estimation models was proposed by Waterman et al. (1977) and can be stated as follows:

Problem 2 (The Minimum Evolution Problem (MEP))**.** *Given a set Γ of n taxa and values $d_{ij} \geq 0$ for all pairs of taxa $i, j \in \Gamma$ $(i \neq j)$, find a phylogeny $T^* \in \mathcal{T}$ that solves the problem*

$$\min_{T \in \mathcal{T}} \quad \mathcal{L}(T) = \sum_{e \in E(T)} w_e \tag{1}$$

$$s.t. \quad \sum_{e \in P(i,j)} w_e \geq d_{ij} \qquad \forall\ i, j \in \Gamma : i \neq j \tag{2}$$

$$w_e \geq 0 \qquad \forall\ e \in E(T) \tag{3}$$

where $E(T)$ denotes the set of edges of a phylogeny $T \in \mathcal{T}$, w_e is the weight of edge $e \in E(T)$, and $P(i, j)$ is the unique path in T connecting the distinct taxa $i, j \in \Gamma$.

The generic value d_{ij} in Problem 2 is called the *observed evolutionary distance* between taxa i and j, and represents a given measure of the observed dissimilarity between i and j; the d_{ij}'s satisfy $d_{ij} = d_{ji}$ for all $i, j \in \Gamma$, $i \neq j$. These values are usually computed, e.g., by using one of the models described in Jukes and Cantor (1969); Fitch (1971); Kidd and Sgaramella-Zonta (1971); Beyer et al. (1974); Hasegawa et al. (1981); Kimura (1980); Lanave et al. (1984); Rodriguez et al. (1990); Waddell and Steel (1997); Galtier (2001); Huelsenbeck (2002); Lopez et al. (2002); Felsenstein (2004); Catanzaro et al. (2006).

The objective function (1) models condition (i), while constraints (2)–(3) impose condition (ii). Problem 2 is a particular network design problem (see Johnson et al., 1978; Pop, 2012) with specific degree constraints and unknown edge weights. As observed in Catanzaro (2011), all known versions of ME can be obtained from MEP by imposing further constraints on the edge weights.

The MEP can be solved in polynomial time if the observed evolutionary distances satisfy some specific properties described in Waterman et al. (1977). In contrast, if the observed evolutionary distances are generic, then the MEP is proved to be $\mathcal{NP}$-hard via a reduction from the k-coloring problem (Farach et al., 1995). Moreover, in such a case the MEP cannot be even approximated in polynomial time within ratio n^ϵ, for some $\epsilon > 0$, unless $\mathcal{P} = \mathcal{NP}$ (Farach et al., 1995).

The hardness of the MEP has prevented researchers from finding practical solution techniques even for small ($n \leq 10$) instances. This fact justifies the interest of operations researchers in developing approaches that can exactly solve the problem or approximate its optimal solution. In this context, Catanzaro et al. (2009) recently presented mixed integer programming models for the MEP based on a particular encoding of phylogenies by means of edge-path incidence matrices of trees (see, e.g., Nemhauser and Wolsey, 1999). This encoding allowed the use of ad-hoc block decomposition methods capable of reducing the solution space of the problem. Unfortunately, all the proposed models proved to be unable to solve real instances of the MEP containing more than 8 taxa.

Starting from the theoretical results presented in Catanzaro et al. (2009); Aringhieri et al. (2011) and Catanzaro et al. (2012), in this article we investigate the mathematical foundations of the MEP and its

relationships with another recent version of the PEP: the *Balanced Minimum Evolution Problem* (BMEP), an $\mathcal{NP}$-hard combinatorial optimization problem closely related to the Quadratic Assignment Problem. One of our findings shows that the BMEP provides lower bounds on the MEP. Moreover, we also show that the MEP is statistically consistent, a very desirable property for phylogeny estimation methods (see Section 6). We also develop an exact solution approach for the problem based on a nontrivial combination of a parallel branch-price-and-cut scheme and a non-isomorphic enumeration of all possible phylogenies for a set of n taxa. This particular approach has two main benefits: it allows both to break symmetries in the problem and to solve instances whose size is 62.5% larger than the size of the instances solved on the same machine with the techniques described in Catanzaro et al. (2009). Due to the hardness of the problem, the computational performance of the algorithm is still far from being suitable for practical use. However, the theoretical results provide a new perspective on the combinatorics of the MEP and may suggest new directions for the development of better exact solution approaches that could turn out useful in practical applications.

The rest of the paper is organized as follows. In Section 2 we introduce some notation, definitions and fundamental properties of phylogenies. In Section 3 we investigate a possible solution approach to the MEP based on a particular type of implicit enumeration called *Non-isomorphic Enumerative Approach* (NEA). In Section 4 we develop a branch-price-and-cut algorithm to speed-up some enumerative tasks involved in the NEA. In Section 5 we discuss the computational performance of the NEA on real instances of the problem and compare our results with the current state-of-the-art algorithm for the MEP. Finally, in Section 6 we prove the statistical consistency of the MEP.

2 Notation and fundamental properties of phylogenies

In this section we introduce some notation and briefly recall some fundamental properties of phylogenies that will be useful throughout the article. We mention that these properties allow one to consider the MEP as an optimization problem over the lattice of the phylogenies of Γ. The interested reader is referred to Catanzaro et al. (2012) and Parker and Ram (1996) for a more systematic discussion about these properties.

Given a phylogeny T of Γ and a taxon $i \in \Gamma$, we denote by Γ_i the set $\Gamma \setminus \{i\}$, by V_I the set of $n-2$ internal vertices of T, and by τ_{ij} the *topological distance* between two distinct taxa/internal vertices $i, j \in \Gamma \cup V_I$, i.e., the number of edges belonging to the path from i to j in T. We also assume that Γ is ordered and we use the notation $i < j$, for some distinct i and $j \in \Gamma$, to indicate that taxon i precedes taxon j in Γ. We define the *path-length sequence* $\tau_i = [\tau_{ij} : j \in \Gamma_i]$ to be the sequence of the topological distances relative to the $n-1$ paths from taxon i to each taxon $j \in \Gamma_i$ in T, ordered accordingly to the order of taxa in Γ. For example, for the phylogeny shown in Figure 1, the path-length sequence from taxon 'A' is $\tau_A = [2, 4, 4, 3]$.

Given a phylogeny T of Γ and a taxon $i \in \Gamma$, we denote by $\hat{i}$ the only vertex adjacent to i in T. We say that a given edge is *external* if it joins a taxon i with the corresponding vertex $\hat{i}$, and *internal* otherwise. For example, for the phylogeny shown in Figure 1, if $i =$'A' then $\hat{i} = 1$, hence A-1 is an external edge and e.g., 1-3 is internal. We denote by $\mathbf{D}$ an $n \times n$ symmetric distance matrix whose generic entry d_{ij}, $i, j \in \Gamma$, $i \neq j$, represents the observed evolutionary distance between the corresponding pair of taxa. Finally, we call *unweighted* a phylogeny of Γ for which the edge weights are not specified.

Topological distances in a phylogeny $T \in \mathcal{T}$ are symmetric, that is:

$$\tau_{ij} = \tau_{ji}$$

for all $i, j \in \Gamma \cup V_I$, $i < j$. Topological distances have also less trivial properties that derive from the particular degree constraints of phylogenies. Specifically, we have:

Proposition 1. *(Parker and Ram, 1996) Let Γ be a set of n taxa, and let $i \in \Gamma$. A sequence of positive integers $\tau_i = [\tau_{ij} : j \in \Gamma_i]$ is a path-length sequence (with respect to taxon i) of a phylogeny $T \in \mathcal{T}$ if and only if the entries of τ_i satisfy the following condition:*

$$\sum_{j \in \Gamma_i} 2^{-\tau_{ij}} = \frac{1}{2}. \tag{4}$$

We refer to equation (4) as the *Kraft equality*.

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

Table 1: Number of labeled and unlabeled phylogenies for increasing number of taxa (some values are approximated).

Proposition 2. *(Catanzaro et al., 2012) Let Γ be a set of n taxa. Then, for all the $T \in \mathcal{T}$, the following equality holds:*

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = 2n - 3. \tag{5}$$

3 A non-isomorphic enumerative solution approach to the MEP

We say that two (unweighted) phylogenies $T_1, T_2 \in \mathcal{T}$ are *isomorphic* if there exists a graph isomorphism between T_1 and T_2, i.e., a bijection ϕ from the vertex set of T_1 to the vertex set of T_2 such that two vertices u, v are adjacent in T_1 if and only if $\phi(u), \phi(v)$ are adjacent in T_2. (Note that here we ignore the edge weights of phylogenies and only consider the underlying graph.) Phylogeny isomorphism defines an equivalence relation in $\mathcal{T}$ and we refer to each equivalence class as a single *unlabeled phylogeny*. For example, the phylogeny in Figure 1 belongs to the equivalence class of the unlabeled phylogeny shown in Figure 3.

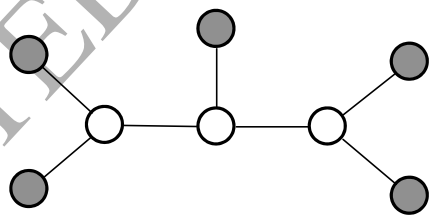

Figure 3: An unlabeled phylogeny of five taxa.

All phylogenies in $\mathcal{T}$ can be generated 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 $f : \Gamma \to L$. This insight is useful to develop a possible exact solution approach for the MEP based on the iteration of the following two operations: (i) choose an unlabeled phylogeny U of Γ, and (ii) assign taxa to leaves of U and weights to its edges in order to minimize (1). This approach, hereafter called *Non-isomorphic Enumerative Approach* (NEA), has a fundamental advantage: 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 1). Hence, the NEA may have a chance to compete with the exact solution approach for the MEP described in Catanzaro et al. (2009), provided that the unlabeled phylogenies of Γ can be efficiently enumerated and the assignment problem can be efficiently solved. In this section we shall investigate both the enumeration and the assignment problems. However, before proceeding we need to introduce some definitions and tools that will prove useful.

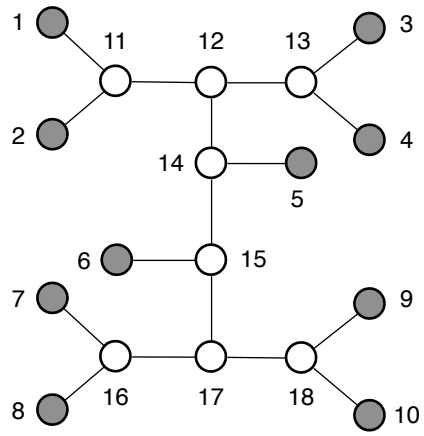

Figure 4: An unlabeled phylogeny with 10 taxa (numbers are attached to nodes for ease of reference).

3.1 Root-vertex and edge isomorphisms

Consider an unlabeled phylogeny U of Γ and let u be an internal vertex of U. Let S_1 and S_2 be two subtrees of U rooted at u. We say that S_1 and S_2 are *root-vertex isomorphic with respect to* u if (i) they are isomorphic, (ii) u is their only vertex in common, and (iii) the leaves of S_1 and S_2 are leaves of U. For example, in the unlabeled phylogeny in Figure 4, the subtree rooted at vertex 12 having $\{1, 2\}$ as leaf set and the subtree rooted at vertex 12 having $\{3, 4\}$ as leaf set are root-vertex isomorphic. Similarly, let $e = uv$ be an internal edge of U, and let S_1 and S_2 be two subtrees of U rooted at u and v, respectively. We say that S_1 and S_2 are *edge-isomorphic with respect to* e if (i) they are isomorphic, (ii) they have no vertex in common, and (iii) the leaves of S_1 and S_2 are leaves of U. For example, in the phylogeny in Figure 4, the subtree rooted at vertex 14 having $\{1, 2, 3, 4, 5\}$ as leaf set and the subtree rooted at vertex 15 having $\{6, 7, 8, 9, 10\}$ as leaf set are edge-isomorphic.

With little abuse of notation with respect to the standard terminology, we call *claw* of U each $K_{1,2}$-subtree of U consisting of two external edges that are incident with a common internal vertex (thus the remaining endpoints are two leaves of U). Note that each claw of U consists of two single-edge root-vertex isomorphic subtrees. For instance, the unlabeled phylogeny shown in Figure 4 has four claws rooted at vertices 11, 13, 16, 18, respectively.

Finally, given an unlabeled phylogeny U of Γ, we denote by $\mathcal{S}$ the set of all unordered pairs of subtrees of U that are either root-vertex isomorphic or edge-isomorphic. For example, the set $\mathcal{S}$ relative to the unlabeled phylogeny shown in Figure 4 includes six pairs of root-vertex isomorphic subtrees, namely the four claws mentioned before and two pairs of isomorphic subtrees rooted at vertices 12 and 17, respectively. Moreover, $\mathcal{S}$ includes also one pair of edge-isomorphic subtrees rooted at the endpoints of edge 14-15.

We remark that for a fixed unlabeled phylogeny U, a given subtree of U can be root-vertex or edge isomorphic to at most one other subtree, with possibly one exception: there might be a node of U that is the root of three root-vertex isomorphic subtrees (see Figure 5). It is easy to see that this happens for at most one node of U.

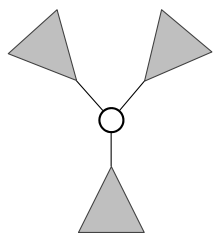

Figure 5: An unlabeled phylogeny containing a node that is the root of three root-vertex isomorphic subtrees. The triangles represent isomorphic binary subtrees.

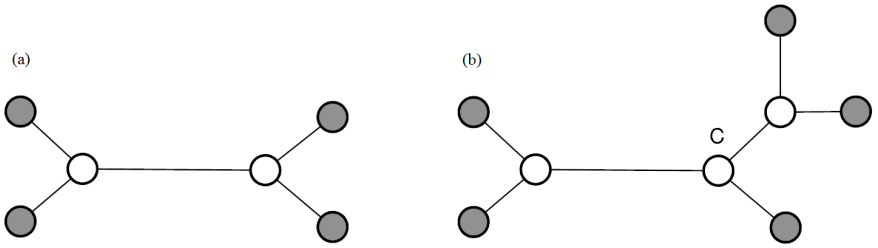

Figure 6: An example of one-to-one enumeration: tree (b) with 8 vertices (5 taxa) generated from a tree (a) with 6 vertices (4 taxa).

3.2 Enumerating unlabeled phylogenies

The NEA enumerates the unlabeled phylogenies through a procedure based on the results on the degree-constrained tree enumeration described in Aringhieri et al. (2003). Specifically, starting from an initial tree, this procedure generates recursively larger and larger trees in such a way that, at each step of the enumeration process, the degree constraint on the internal vertices is satisfied. Aringhieri et al. (2011) showed that, when dealing with unlabeled phylogenies, this enumeration process can be easily performed by recursively substituting a leaf of the current tree with a new claw. For example, the unlabeled phylogeny with 5 taxa shown in Figure 6 can be obtained from the unlabeled phylogeny with 4 taxa by replacing its top-right leaf with a new claw.

The enumeration procedure encodes the trees that it recursively generates in order to maximize the efficiency of the process. Among the possible encodings, see e.g., Read (1972); Kvasnička and Pospichal (1991); Trinajstić et al. (1991); Hansen et al. (1994), the *$\mathcal{CN}$-tuple encoding* (Hansen et al., 1994) seems particularly suitable for unlabeled phylogenies. This kind of encoding is based on the following result.

Proposition 3. *(Jordan, 1869) Let T be a tree with n vertices.*

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

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

   (a) *a unique node v, called the* centroid*, such that all (three or more) subtrees obtained by removing v contain less than k nodes, or*

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

Here, we observe that if an unlabeled phylogeny T has a pair of edge isomorphic subtrees, then the corresponding edge is the bicentroid of the tree. If T has a vertex that is the root of three isomorphic subtrees, then this vertex is the centroid of the tree. In the remainder we refer to the centroid or the bicentroid of a given unlabeled phylogeny simply as *the center of the tree.*

The $\mathcal{CN}$-tuple code of an unlabeled phylogeny U with n taxa (and thus $2n - 2$ vertices) is a string $\langle \alpha_1 \alpha_2 \ldots \alpha_{2n-2} \rangle$ over the alphabet $\Sigma = \{0, 1, 2, 3\}$ that is associated to the center of U and built from U accordingly to a particular recursive law. Specifically, if the center of U is a centroid, then the $\mathcal{CN}$-tuple code of U is recursively defined as follows:

Base case: If U consists of a single vertex, its code is $\langle 0 \rangle$.

Recursive law: If U has more than one vertex, let r be the center of U and let $v_1, \ldots, v_g$, $g \leq 3$, be the vertices adjacent to r in U. Set g as the first character of the code. Remove r and its incident edges from U. Recursively compute the codes $c_1, c_2, \ldots, c_g$ associated to the roots of the subtrees rooted at $v_1, \ldots, v_g$. The code of U is then obtained by concatenating the codes $c_1, c_2, \ldots, c_g$ in such a way that the resulting string is lexicographically maximum.

For instance, let us consider the unlabeled phylogeny U in Figure 6(b). U has a unique center (labeled with c), which is connected to $g = 3$ nodes. By removing the center, we obtain one single-vertex subtree, whose

code is $\langle 0 \rangle$, and two subtrees (U_1 and U_2, say) with three nodes each. For $i = 1, 2$, U_i has a unique center, whose $g = 2$ neighbors are single vertices and thus have code $\langle 0 \rangle$; then, for $i = 1, 2$, the code of U_i is obtained by setting '2' as the first character and then concatenating '0' and '0' in maximum lexicographic order, that is $\langle 200 \rangle$. The final $\mathcal{CN}$-tuple code of U is obtained by setting '3' as the first character and then concatenating '0', '200' and '200' in maximum lexicographic order: $\langle 32002000 \rangle$.

When the center of U is a bicentroid, the $\mathcal{CN}$-tuple code of U is the string that is lexicographically maximum between the ones that are obtained by applying the recursive law to each of the two vertices that constitute the bicentroid. For example the $\mathcal{CN}$-tuple code of the unlabeled phylogeny in Figure 6(a) is $\langle 320000 \rangle$.

The enumerating procedure exploits the $\mathcal{CN}$-tuple encoding to avoid generating isomorphic phylogenies from the same initial subtree U_0. Specifically, it employs the code to detect the presence of isomorphic subtrees in U_0. As an example, in this way the enumerating procedure identifies that tree (a) in Figure 6 contains isomorphic subtrees and then it generates only one of the four isomorphic phylogenies (b) that can be obtained by transforming into a claw any of the four leaves of (a). Other kinds of enumeration procedures could have been implemented but, in the authors' opinion, they would require more sophisticated techniques to prevent the generation of isomorphic unlabeled phylogenies, see, e.g., the techniques described in Avis and Fukuda (1992, 1996); Aringhieri et al. (2003).

3.3 The assignment problem

The second step in the NEA consists in assigning taxa to leaves of an unlabeled phylogeny U and weights to its edges minimizing objective function (1). This task can be formalized as follows. Let U be an unlabeled phylogeny of Γ, denote by $\mathcal{F}$ the set of all $n!$ bijective mappings f of Γ to the set of n leaves of U, and by T_f the phylogeny belonging to the class U such that the taxa are assigned to its leaves by means of the assignment $f \in \mathcal{F}$. Then, NEA solves the following assignment problem:

Problem 3 (Fixed Phylogeny Problem (FPP))**.** *Given an unlabeled phylogeny U of Γ, find an optimal assignment $f^* \in \mathcal{F}$ and weights $w_e^* \geq 0$ that solve the problem*

$$z(U) = \min_{f \in \mathcal{F}} z_f \tag{6a}$$

$$z_f = \min \sum_{e \in E(T_f)} w_e \tag{6b}$$

$$\text{s.t.} \sum_{e \in P(f(i), f(j))} w_e \geq d_{ij} \qquad \forall\ i, j \in \Gamma : i \neq j \tag{6c}$$

$$w_e \geq 0 \qquad \forall\ e \in E(T_f). \tag{6d}$$

From now on, we define two phylogenies T_f and $T_{f'}$, that are feasible solutions of FPP, *equivalent* if each pair i, j of taxa in Γ has the same topological distance $\tau_{ij} = \tau'_{ij}$ on the two phylogenies, and the weights w_e and w'_e of corresponding edges of the paths from i to j on the two phylogenies are equal, $w_e = w'_e$. Then we observe that in general, given an unlabeled phylogeny U of Γ, different assignments in $\mathcal{F}$ may produce equivalent phylogenies. Given an assignment $f \in \mathcal{F}$ and any automorphism ϕ of T_f (i.e., any permutation ϕ of the vertices of T_f, such that the pair of vertices uv forms an edge if and only if the pair $\phi(u)\phi(v)$ also forms an edge), the assignment $f' \in \mathcal{F}$ defined by $f'(i) = \phi(f(i))$ produces a phylogeny $T_{f'}$ equivalent to T_f. As an example, for the unlabeled phylogeny shown in Figure 4, it is easy to realize the equivalence of the phylogenies that differ only for the swapping of the assignments of the taxa between the pair of leaves $\{1, 2\}$, or $\{3, 4\}$, or $\{7, 8\}$, or $\{9, 10\}$. We also observe that the relation "defining an equivalent phylogeny" is trivially an equivalence relation on $\mathcal{F}$. Then we can constrain the search for an optimal assignment f^* to a single representative assignment f for each equivalence class in $\mathcal{F}$.

The next proposition provides the number of automorphisms $\#T$ of a phylogeny T, i.e., the number of elements of $\mathcal{F}$ in each equivalence class. To this end, we recall that by $\mathcal{S}$ we denote the set of unordered pairs of root-vertex or edge isomorphic subtrees of U.

Proposition 4. *Given a phylogeny T belonging to the class of U, $\#T = \frac{3}{4} \cdot 2^{|\mathcal{S}|}$ if T has a node that is the root of three root-vertex isomorphic subtrees, and $\#T = 2^{|\mathcal{S}|}$ otherwise.*

Proof. The proof is by induction on the number of nodes of T. For the induction to work, it is convenient to prove the result not only for phylogenies, but for every tree in which all internal vertices have degree 3 except at most one vertex, which might have degree 2. We remark that all the definitions and results of Section 3.1 apply to this more general case.

The base case of the induction is that of a tree consisting of a single node: in this situation $\mathcal{S} = \emptyset$ and the only automorphism is the identity, thus the result $\#T = 1$ is verified.

Now let T be a tree as above, where the number of nodes is at least 2. Let us first assume that T does not have a bicentroid (therefore there is no pair of edge isomorphic subtrees in T). If there is a (unique) node of degree 2 in T, let v denote such a node; otherwise, let v be the centroid of the tree. Let k be the degree of v (so either $k = 2$ or $k = 3$) and denote by $v_1, \ldots, v_k$ its neighbors. If we remove v from T, we obtain k subtrees $T_1, \ldots, T_k$ (containing $v_1, \ldots, v_k$, respectively).

In the following, as inductive step, we show that $\#T$ is proportional to $\prod_{i\in[k]} \#T_i$, where $[k] = \{1, \ldots, k\}$, and we determine the proportionality constant.

Note that every automorphism ϕ of T satisfies $\phi(v) = v$, as either v is the unique vertex of degree 2 or the unique centroid of T (and both properties are preserved by automorphisms). Note also that there exist automorphisms ϕ of T such that $\phi(v_i) = v_j$ with $i, j \in [k]$ if and only if T_i and T_j are root-vertex isomorphic subtrees of T (or $i = j$). In general, we can say that every automorphism ϕ of T is associated with a specific permutation σ of the set $[k]$ such that $\phi(v_i) = v_{\sigma(i)}$ for $i \in [k]$. Accordingly, we call feasible every permutation σ such that there is an automorphism of T associated with σ.

Denote by $\mathcal{S}_i$ the set of unordered pairs of root-vertex or edge isomorphic subtrees of the class U_i that includes T_i, for $i \in [k]$, and observe that the properties of v can be linked to the number of root-vertex or edge isomorphic subtrees of T and the number of automorphisms of T:

(a) if v is not the root of any two root-vertex isomorphic subtrees, then $|\mathcal{S}| = |\mathcal{S}_1| + \cdots + |\mathcal{S}_k|$ and there is only one feasible permutation σ, the identical one ($\sigma(i) = i$ for all $i \in [k]$), as every automorphism ϕ of T must satisfy $\phi(v_i) = v_i$ for $i \in [k]$;

(b) if v is the root of exactly two root-vertex isomorphic subtrees, let us say T_i and T_j, then $|\mathcal{S}| = |\mathcal{S}_1| + \cdots + |\mathcal{S}_k| + 1$ and there are only two feasible permutations, the identical one and the one that permutes only i with j, as every automorphism ϕ of T must satisfy either $\phi(v_i) = v_i$ for $i \in [k]$, or $\phi(v_i) = v_j$, $\phi(v_j) = v_i$, and $\phi(v_l) = v_l$ for $l \in [k] \setminus \{i, j\}$;

(c) if v is the root of exactly three root-vertex isomorphic subtrees (thus, in particular, $k = 3$), then $|\mathcal{S}| = |\mathcal{S}_1| + \cdots + |\mathcal{S}_k| + 3$ and all the $k! = 6$ permutations of $[k]$ are feasible.

Let t be the number of feasible permutations. In the three cases analyzed above, the value of t is 1, 2, and 6, respectively.

We now claim that $\#T = t \times \prod_{i\in[k]} \#T_i$. To see this, note that if ϕ is an automorphism of T and σ is the associated permutation of $[k]$, then $\sigma^{-1} \circ \phi$ is an automorphism of T associated with the identity permutation; vice versa, if ϕ' is any automorphism of T associated with the identity permutation and σ is a feasible permutation, then $\phi' \circ \sigma$ is an automorphism of T associated with σ. It follows that $\#T$ is equal to t times the number of automorphisms of T associated with the identity permutation. Now, an automorphism of T is associated with the identity permutation if and only if its restriction to T_i is an automorphism of T_i, for $i \in [k]$; vice versa, any choice of automorphisms of $T_1, \ldots, T_k$ leads to an automorphism of T associated with the identity permutation. It follows that the number of automorphisms of T associated with the identity permutation is exactly $\prod_{i\in[k]} \#T_i$, and thus $\#T = t \times \prod_{i\in[k]} \#T_i$ as claimed.

Note that each T_i is a tree with all internal vertices of degree 3, except for v_i, whose degree is 2 (here we use the fact the v is defined as the only vertex of degree 2 in T if there is such a vertex, and as the centroid only when there is no vertex of degree 2 in T). This easily implies that no vertex of T_i is the root of three root-vertex isomorphic subtrees in T_i. Then, by induction, $\#T_i = 2^{|\mathcal{S}_i|}$, hence $\#T = t \times 2^{|\mathcal{S}_1|} \times \cdots \times 2^{|\mathcal{S}_k|}$. The conclusion now follows by using (a)–(c) and recalling that t is equal to 1 (resp., 2, 6) in case (a) (resp., (b), (c)).

This concludes the proof for the case of a tree T with no bicentroid. Now assume that T has a bicentroid $e = uv$. Note that in this case all nodes of T have degree 3 (otherwise by removing e we would obtain two subtrees with a different number of nodes, a contradiction). We can reduce this case to the one analyzed above by subdividing the edge e, i.e., replacing edge $e = uv$ with a path of length two of the form u, z, v, where z is a new node. It is easy to check that neither $|\mathcal{S}|$ nor $\#T$ changes. □

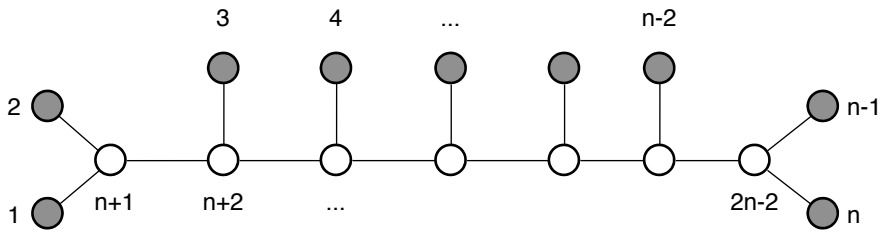

Figure 7: An example of unlabeled phylogeny with only two claws.

We remark that the above proof also shows that every automorphism of a phylogeny T can be derived as composition of automorphisms of T associated with the pairs of subtrees in $\mathcal{S}$. This will prove useful later.

By the above proposition, only $\frac{n!}{2^{|\mathcal{S}|}}$ or $\frac{4}{3} \cdot \frac{n!}{2^{|\mathcal{S}|}}$ out of $n!$ assignments need to be considered for each unlabeled phylogeny. Assuming $n \geq 4$, we can derive an estimation of $|\mathcal{S}|$ by using the following trivial properties:

1. an unlabeled phylogeny has at least as many pairs of root-vertex isomorphic subtrees as the number of its claws;

2. an unlabeled phylogeny has at most one pair of edge-isomorphic subtrees; moreover, this situation occurs only if the unlabeled phylogeny possesses a bicentroid.

Then the following proposition holds:

Proposition 5. *Every unlabeled phylogeny U of n taxa satisfies $3 \leq |\mathcal{S}| \leq n$.*

Proof. We have $|\mathcal{S}| \geq 3$, as any unlabeled phylogeny has at least three claws, except for the unlabeled phylogenies in which all $n-2$ internal nodes are adjacent to a leaf (see Figure 7). For this latter kind of phylogenies, $|\mathcal{S}| = 3$ holds, since they have two claws and, if they have an odd number of internal vertices, they contain two root-vertex isomorphic subtrees rooted at the centroid, and otherwise, they contain two edge-isomorphic subtrees with respect to the bicentroid.

To see that $|\mathcal{S}| \leq n$, assume first that U does not have a vertex that is the root of three vertex isomorphic subtrees. Then, since only the internal vertices can be roots of root-vertex isomorphic subtrees, and there is at most one pair of edge-isomorphic subtrees, we have $|\mathcal{S}| \leq n-1$. Finally, suppose that U has a vertex v that is the root of three root-vertex isomorphic subtrees. Then v is the centroid of U, thus there is no pair of edge-isomorphic subtrees. Since v is the root of three pairs of isomorphic subtrees, and every other internal vertex is the root of at most one pair of isomorphic subtrees, we conclude that $|\mathcal{S}| \leq n$. □

3.4 A mixed integer programming formulation for the fixed phylogeny problem

In the light of the results discussed in the previous sections, we now introduce a mixed integer programming model to solve the FPP (Problem 3) for a fixed unlabeled phylogeny U of Γ having L as set of leaves. To this end, let us assume without loss of generality that $L = \{1, \ldots, n\}$ and $\Gamma = \{1, \ldots, n\}$ (so they are both totally ordered sets). Given an unlabeled phylogeny U of Γ, we define $w_e, e \in E(U)$, as a nonnegative continuous variable representing the edge weight of $e \in E(U)$. Moreover, we use the following decision variables:

$$x_i^p = \begin{cases} 1 & \text{if taxon } p \text{ is assigned to leaf } i \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, p \in \Gamma, \forall\, i \in L$$

and

$$y_{ij}^{pq} = \begin{cases} 1 & \text{if } p \text{ is assigned to } i \text{ and } q \text{ to } j \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, p, q \in \Gamma : p \neq q, \forall\, i, j \in L : i < j.$$

Note the correspondence between the above variables and the desired assignment f^*: $x_i^p = 1$ if and only if $f^*(p) = i$, and $y_{ij}^{pq} = 1$ if and only if $f^*(p) = i$ and $f^*(q) = j$. Then a formulation for the FPP is the following:

Formulation 1.

$$z(U) = \min \sum_{e \in E(U)} w_e \tag{7a}$$

$$s.t. \sum_{e \in P(i,j)} w_e \geq d_{pq}(y_{ij}^{pq} + y_{ij}^{qp}) \qquad \forall\, i,j \in L : i < j,\ \forall\, p,q \in \Gamma : p < q \tag{7b}$$

$$\sum_{\substack{p,q \in \Gamma \\ p \neq q}} y_{ij}^{pq} = 1 \qquad \forall\, i,j \in L : i < j \tag{7c}$$

$$\sum_{\substack{i,j \in L \\ i < j}} (y_{ij}^{pq} + y_{ij}^{qp}) = 1 \qquad \forall\, p,q \in \Gamma : p \neq q \tag{7d}$$

$$y_{ij}^{pq} \leq x_i^p \qquad \forall\, i,j \in L : i < j,\ \forall\, p,q \in \Gamma : p \neq q \tag{7e}$$

$$y_{ij}^{pq} \leq x_j^q \qquad \forall\, i,j \in L : i < j,\ \forall\, p,q \in \Gamma : p \neq q \tag{7f}$$

$$\sum_{i \in L} x_i^p = 1 \qquad \forall\, p \in \Gamma \tag{7g}$$

$$\sum_{p \in \Gamma} x_i^p = 1 \qquad \forall\, i \in L \tag{7h}$$

$$w_e \geq 0 \qquad \forall\, e \in E(U) \tag{7i}$$

$$x_i^p \in \{0,1\} \qquad \forall\, p \in \Gamma,\ \forall\, i \in L \tag{7j}$$

$$y_{ij}^{pq} \in \{0,1\} \qquad \forall\, p,q \in \Gamma : p \neq q,\ \forall\, i,j \in L : i < j. \tag{7k}$$

Constraints (7b) impose that, for each pair of distinct taxa $p, q \in \Gamma$, the sum of the weights of the edges in $P(i,j)$ is greater than or equal to the observed evolutionary distance d_{pq} if $P(i,j)$ joins taxa p and q. Equations (7c) impose that for a given pair of taxa (i,j) there is a unique y_{ij}^{pq}-variable set to 1. Symmetrically, equations (7d) ensure that for each pair of taxa (p,q) there exists exactly one y_{ij}^{pq}-variable set to 1. Inequalities (7e) impose that $x_i^p = 1$ whenever $y_{ij}^{pq} = 1$ for some leaf j and taxon q, and (7f) impose a similar condition. Constraints (7g)–(7h) ensure that the x_i^p variables model a one-to-one taxon-leaf assignment. Finally, constraints (7i)–(7k) impose the nonnegativity of the edge weights and integrality of variables x_i^p and y_{ij}^{pq}, respectively.

Thanks to the presence of constraints (7c)–(7d), we can replace inequalities (7b) with the following stronger inequalities:

$$\sum_{e \in P(i,j)} w_e \geq \sum_{\substack{p,q \in \Gamma \\ p \neq q}} d_{pq}(y_{ij}^{pq} + y_{ij}^{qp}) \qquad \forall\, i,j \in L : i < j. \tag{8}$$

Furthermore, inequalities (7e)–(7f) could be replaced by

$$\sum_{\substack{q \in \Gamma \\ q \neq p}} y_{ij}^{pq} = x_i^p \qquad \forall\, p \in \Gamma,\ \forall\, i,j \in L : i < j \tag{9}$$

$$\sum_{\substack{p \in \Gamma \\ p \neq q}} y_{ij}^{pq} = x_j^q \qquad \forall\, q \in \Gamma,\ \forall\, i,j \in L : i < j, \tag{10}$$

thus obtaining a strengthened linear-programming relaxation. However, extensive preliminary tests run on a number of variations of the above model showed that, under our branch-price-and-cut approach described in Section 4.1, it is convenient to write the constraints in the form (7e)–(7f) and add equations (9)–(10) as cuts.

We also observe that the integrality requirements on the y_{ij}^{pq}-variables can be relaxed, as the following proposition shows.

Proposition 6. *Constraints* (7k) *can be relaxed to*

$$y_{ij}^{pq} \geq 0 \qquad \forall\, p,q \in \Gamma : p \neq q,\ \forall\, i,j \in L : i < j. \tag{11}$$

Proof. We show that variables y_{ij}^{pq} take integer values in any feasible solution to the problem in which constraints (7k) are replaced by (11). Observe that constraints (11) and (7e)–(7f) imply

$$0 \leq y_{ij}^{pq} \leq \min\{x_i^p, x_j^q\} \quad \forall\, p, q \in \Gamma : p \neq q, \forall\, i, j \in L : i < j, \tag{12}$$

hence, because of (7j), $y_{ij}^{pq} > 0$ only if $x_i^p = x_j^q = 1$. In this case, (7g)–(7h) imply that $x_i^{p'} = x_j^{q'} = 0$ for all $p' \neq p$ and $q' \neq q$. Then, by (7e)–(7f), $y_{ij}^{p'q'} = 0$ whenever $(p', q') \neq (p, q)$. Equation (7c) then implies that $y_{ij}^{pq} = 1$. □

3.4.1 Strengthening Formulation 1

A first family of valid cuts for our formulation is given below.

Proposition 7. *For every leaf $i \in L$ and distinct taxa $p, q \in \Gamma$, constraint*

$$\sum_{\substack{j \in L \\ i<j}} y_{ij}^{pq} + \sum_{\substack{j \in L \\ i>j}} y_{ji}^{qp} = x_i^p \tag{13}$$

is valid for Formulation 1.

Proof. The proof of Proposition 6 shows that in any feasible solution $y_{ij}^{pq} = x_i^p x_j^q$ for all $p, q \in \Gamma$ such that $p \neq q$ and $i, j \in L$ such that $i < j$. With this in mind, one immediately sees that if $x_i^p = 0$ then both sides of (13) are equal to zero and the equation is satisfied, and if $x_i^p = 1$, then there exists a unique $j \in L \setminus \{i\}$ such that $x_j^q = 1$, thus the equation is again satisfied. □

By exploiting the fundamental properties of phylogenies and the integrality of variables x_i^p and y_{ij}^{pq}, some valid inequalities can be developed to strengthen Formulation 1. Specifically, the following propositions hold:

Proposition 8. *The equalities*

$$\sum_{q \in \Gamma_p} \sum_{\substack{j \in L \\ i<j}} \left(y_{ij}^{pq} + y_{ij}^{qp}\right) 2^{-\tau_{ij}} = \frac{1}{2} \qquad \forall\, i \in L, \forall\, p \in \Gamma \tag{14}$$

$$\sum_{\substack{p,q \in \Gamma \\ p \neq q}} \sum_{\substack{i,j \in L \\ i<j}} (y_{ij}^{pq} + y_{ij}^{qp}) \tau_{ij} 2^{-\tau_{ij}} = 2n - 3 \tag{15}$$

are valid for Formulation 1.

Proof. The equalities follows by observing that in any feasible solution to the FPP the Kraft equality (4) and equality (5) hold. □

Proposition 9. *Given an unlabeled phylogeny U of Γ, the inequality*

$$\sum_{e \in E(U)} w_e \geq \sum_{\substack{p,q \in \Gamma \\ p \neq q}} \sum_{\substack{i,j \in L \\ i<j}} (y_{ij}^{pq} + y_{ij}^{qp}) d_{ij} 2^{-\tau_{ij}} \tag{16}$$

is valid for Formulation 1.

Proof. From Semple and Steel (2004) it follows that for every phylogeny T belonging to the class U,

$$\mathcal{L}(T) = \sum_{e \in E(T)} w_e = \sum_{\substack{i,j \in L \\ i \neq j}} \delta_{ij} 2^{-\tau_{ij}}, \tag{17}$$

where δ_{ij} is equal to the sum of the weights of the edges belonging to the path from leaf i to leaf j in T. Then, as $E(T) = E(U)$, constraints (6c) give

$$\sum_{e \in E(U)} w_e = \sum_{\substack{i,j \in L \\ i \neq j}} \delta_{ij} 2^{-\tau_{ij}} \geq \sum_{\substack{i,j \in L \\ i \neq j}} d_{ij} 2^{-\tau_{ij}} = \sum_{\substack{p,q \in \Gamma \\ p \neq q}} \sum_{\substack{i,j \in L \\ i<j}} (y_{ij}^{pq} + y_{ij}^{qp}) d_{ij} 2^{-\tau_{ij}}. \tag{18}$$

□

It is worth noting that Proposition 9 not only gives a valid inequality for the MEP, but it also provides a relationship between the *Balanced Minimum Evolution Problem* (BMEP, Pauplin (2000); see also Catanzaro et al., 2012) and the MEP. Specifically, given a set Γ of n taxa and its corresponding distance matrix $\mathbf{D}$, the BMEP consists in finding a phylogeny T that minimizes the following objective function (Pauplin, 2000):

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} d_{ij} 2^{-\tau_{ij}}.$$

Proposition 9 states that the optimal solution to the BMEP provides a lower bound for the MEP. This will be useful when proving the statistical consistency of the MEP in Section 6. Moreover, since it was observed by Fiorini and Joret (2012) that the optimal value of the BMEP is at least half the length of the shortest Hamiltonian cycle in the complete graph with vertex set Γ and weights d_{ij}, Proposition 9 immediately implies the following result.

Proposition 10 (Cycle inequality)**.** *Let G be the weighted complete graph with vertex set Γ and weights d_{ij} for $i, j \in \Gamma$, $i \neq j$. Denote by C^* a shortest Hamiltonian cycle in G and by $L(C^*)$ its length. Then the inequality*

$$\sum_{e\in E(U)} w_e \geq \frac{1}{2} L(C^*) \tag{19}$$

is valid for Formulation 1.

The inclusion of inequality (19) in Formulation 1 can be performed by solving the Traveling Salesman Problem (TSP) on G. Though the TSP is an $\mathcal{NP}$-hard problem, it can be solved very efficiently for quite large values of n (see, e.g., D'Ambrosio et al., 2011); in particular, it can be solved almost instantaneously for the values of n of our instances using Concorde (Applegate et al., 2001).

3.4.2 Reducing the model size by symmetry breaking

A standard branch-and-bound algorithm tends to perform poorly when solving Formulation 1, due to the presence of the equivalent assignments discussed in Section 3.3. Removing those assignments reduces the size of the model and speeds up the overall solution time of the FPP. To do so, we enumerate a single representative assignment for each class of equivalent assignments. Recalling that every automorphism of a given unlabeled phylogeny is the composition of automorphisms associated with the pairs in $\mathcal{S}$, it is sufficient to require that, for each pair of subtrees $\{S_1, S_2\} \in \mathcal{S}$, taxa assigned to S_1 lexicographically precede taxa assigned to S_2. This can be implemented by imposing that, for each pair of isomorphic subtrees $\{S_1, S_2\} \in \mathcal{S}$, the following constraints hold:

$$y_{ij}^{pq} = 0 \quad \forall\, p, q \in \Gamma : q < p \tag{20}$$

where i and $j = \phi(i)$ are two reference vertices in S_1 and S_2, respectively (with ϕ being the isomorphism between S_1 and S_2). As an example, if we consider the phylogeny in Figure 7, all equivalent assignments can be excluded by imposing the following constraints:

$$y_{12}^{pq} = 0 \quad \forall\, p, q \in \Gamma : q < p \tag{21a}$$

$$y_{n,n-1}^{pq} = 0 \quad \forall\, p, q \in \Gamma : q < p \tag{21b}$$

$$y_{1,n-1}^{pq} = 0 \quad \forall\, p, q \in \Gamma : q < p. \tag{21c}$$

Imposing constraints (20) requires finding the tree symmetries discussed in Section 3. Specifically, given an unlabeled phylogeny U, one can detect the presence of a pair of root-vertex isomorphic subtrees in U by checking whether the $\mathcal{CN}$-tuple code of U contains two consecutive identical substrings c_1, c_2, where the first character of c_1 and the first character of c_2 correspond to nodes with the same predecessor. For example, the phylogeny in Figure 6(b) has code $\langle 32002000 \rangle$; the two substrings $\langle 200 \rangle$ denote the presence of a pair of root-vertex isomorphic subtrees rooted at the centroid of U, while each of the first two pairs of consecutive 0s indicates the presence of a claw. Edge-isomorphic subtrees in an unlabeled phylogeny can be detected in a similar way.

Finding the tree symmetries for a fixed unlabeled phylogeny does not require any further computational overhead and can be implemented in a very easy way by using the algorithms described in Aringhieri et al. (2003).

4 Improving the performance of the NEA

The large number of y_{ij}^{pq}-variables in Formulation 1 tends to slow down the resolution of the FPP and constitutes a bottleneck for the NEA. Therefore, it is worth investigating possible strategies to decrease the number of variables involved in the formulation. To this end, in this section we shall develop a branch-price-and-cut approach for the FPP. Moreover, we shall investigate possible heuristic strategies to speed up the overall performance of the NEA.

4.1 A branch-price-and-cut approach for the fixed phylogeny problem

A possible strategy to speed up the resolution of the FPP consists in using a branch-price-and-cut approach in which the y_{ij}^{pq}-variables are dynamically added to Formulation 1. Specifically, at each node of the search tree we solve the corresponding linear program by means of a column generation technique, as described below. Subsequently, we strengthen the linear-programming bound by adding inequalities (9)–(10) and the valid inequalities described in Section 3.4.1. If the node is not pruned, we perform a branching by fixing the value of the most fractional x_i^p variable (i.e., the variable with maximum distance from the closest integer). Finally, we reiterate the procedure by considering the resulting subproblems. We remark that at each node of the search tree we compute a bound for a subproblem by removing all strengthening inequalities previously added to the formulation and by including all the y_{ij}^{pq}-variables introduced during the previous iterations. The choice of ignoring the valid inequalities when performing column generation was due to the large number of preliminary tests carried out on different variants of the model. The strategy described here turned out to be the one providing the best trade-off between solution time and tightness of the bound.

Given a generic node of the search tree of the FPP, consider Formulation 1 and replace inequalities (7b) by the stronger constraints (8); moreover, relax all of the integrality constraints on the decision variables. We assign dual variables B_{ij} to inequalities (8), dual variables C_{ij} constraints to (7c), ..., dual variables H_i to constraints (7h). The feasible region of the dual problem reads as follows:

$$\sum_{\substack{i,j\in L: i<j \\ e\in P(i,j)}} B_{ij} \leq 1 \qquad \forall\, e \in E(U) \tag{22a}$$

$$\sum_{\substack{j\in L \\ j>i}} \sum_{\substack{q\in\Gamma \\ q\neq p}} E_{ij}^{pq} + \sum_{\substack{j\in L \\ j<i}} \sum_{\substack{q\in\Gamma \\ q\neq p}} F_{ji}^{qp} + G_p + H_i \leq 0 \qquad \forall\, i \in L, \forall\, p \in \Gamma \tag{22b}$$

$$d_{pq} B_{ij} + C_{ij} + D_{pq} - E_{ij}^{pq} - F_{ij}^{pq} \leq 0 \qquad \forall\, i,j \in L : i<j, \forall\, p,q \in \Gamma : p \neq q \tag{22c}$$

$$B_{ij} \geq 0,\ E_{ij}^{pq} \geq 0,\ F_{ij}^{pq} \geq 0, \qquad \forall\, i,j \in L : i<j, \forall\, p,q \in \Gamma : p \neq q. \tag{22d}$$

We start with a reduced primal problem restricted to the following variables: variables w_e for all $e \in E(U)$; variables x_i^p for all $i \in L$ and $p \in \Gamma$; variables y_{ij}^{ij} for all $i,j \in L$ such that $i<j$. These variables are sufficient to make the formulation feasible for the FPP. Note that none of inequalities (22a) or (22b) can be violated by the dual solution of the reduced problem, as all primal variables corresponding to these constraints are included in the reduced linear program. Thus we only have to check if the current dual solution satisfies all inequalities (22c). We also note that if a variable y_{ij}^{pq} is not part of the current reduced problem, then the corresponding constraints (7e)–(7f) are redundant, and thus we can ignore the corresponding dual variables E_{ij}^{pq} and F_{ij}^{pq}. On the other side, if a variable y_{ij}^{pq} is part of the current reduced problem, then the dual constraint (22c) associated with indices i,j,p,q is of course satisfied.

We conclude that, in order to verify whether there is a violated dual constraint, it is enough to check whether there exist indices $i,j \in L, i<j$, and $p,q \in \Gamma, p \neq q$, such that the current dual solution satisfies

$$d_{pq} B_{ij} + C_{ij} + D_{pq} > 0.$$

This can be achieved by enumerating all possible quadruples of indices i,j,p,q —it seems unlikely that a less trivial pricing oracle can be devised. Though this pricing oracle requires $\Theta(n^4)$ operations, the overall computation time needed to solve the linear relaxation of Formulation 1 with this approach turned out to be much smaller than the time needed to solve directly the original model.

4.2 Speeding up the NEA

In order to speed up the overall performance of the NEA, one can reduce the number of unlabeled phylogenies of Γ to be considered during the enumeration process. This task can be accomplished by performing a preprocessing phase consisting of the following subtasks: (i) computing a first primal upper bound for the problem; (ii) computing a lower bound for each possible unlabeled phylogeny of Γ (e.g., by solving the linear relaxation of Formulation 1); (iii) sorting in nondecreasing order the unlabeled phylogenies of Γ according to their lower bounds; and finally (iv) pruning those unlabeled phylogenies whose lower bound is greater then the primal bound previously computed. The fact of keeping ordered the list of the unpruned unlabeled phylogenies allows a further potential dynamical pruning whenever a better primal bound is determined. The efficiency of this strategy can be maximized by parallelizing the resolution of the assignment problem, i.e., by solving in parallel the FPPs corresponding to multiple unpruned unlabeled phylogenies.

Finding a good primal bound for the MEP is a critical step to reduce the number of unlabeled phylogenies of Γ to be considered. In order to do so, we exploit the insights provided by Proposition 10. Specifically, let G be the weighted complete graph with vertex set Γ and weights d_{ij} for $i, j \in \Gamma$, $i \neq j$. Let C^* be a shortest Hamiltonian cycle in G and denote by $L(C^*)$ its length. Fix any unlabeled phylogeny U of Γ and denote by $T(C^*)$ the labeled phylogeny obtained from U by proceeding as follows: select at random a leaf of U and assign the taxa to the leaves of U according to the order given by C^*, starting from the selected leaf and proceeding clockwise. Now, finding weights w_e^* for the edges of $T(C^*)$ that minimize the length of the phylogeny is a linear programming problem (as it is Problem 2 for a fixed T) that can be solved in a negligible time. Then we have the following primal bound for the MEP:

Proposition 11. *For a fixed unlabeled phylogeny U of Γ, let $(\bar{w}_e)_{e \in E(U)}$ be the optimal weights for the edges of U. Then*

$$\sum_{e \in E(U)} \bar{w}_e \leq \sum_{e \in E(T(C^*))} w_e^*.$$

The bound provided by Proposition 11 is not tight, but it can be improved by (i) applying the heuristic assignment to all unlabeled phylogenies of Γ, and (ii) using local search. In this context, it is worth noting that for a fixed unlabeled phylogeny of Γ, a first improvement on the heuristic assignment of the taxa in C^* to the leaves of U can be obtained by shifting the starting leaf during the assignment process. Moreover, a further improvement can be achieved applying 2-OPT (Glover and Kochenberger, 2003) on the best shift found. We experienced in preliminary numerical experiments that the primal bounds obtained by using this approach are usually below 3% from the optimum.

5 Numerical experiments

In order to evaluate the efficiency of our exact solution approach for the MEP, we tested the performance of the NEA on nine real aligned DNA datasets already used in Catanzaro et al. (2009, 2012), namely: "Primates12", a dataset of 12 sequences, 898 characters each from primates mitochondrial DNA; "SeedPlant25", a dataset of 25 sequences of 19784 characters each from pinoles; "RbcL55/1314", a dataset of 55 sequences, 1314 characters each of the rbcL gene; "Rana64 /1976", a dataset of mitochondrial DNA containing 64 taxa of 1976 characters each from ranoid frogs; "M17/2550", "M43/2086", "M18/8128", "M82/2062", "M62/3768", five datasets of respectively 17 sequences of 2550 characters each from insects, 43 sequences of 2086 characters each from mammals, 18 sequences of 8128 characters each from cetacea, 82 sequences of 2062 characters each from fungi, and finally 62 sequences of 3768 characters each from hyracoidae. Table 2 provides a summary of the characteristics of considered datasets.

From each dataset we have extracted the first 20 taxa (or all taxa if $n < 20$) and built the associated $n \times n$ distance matrix by using the General Time Reversible (GTR) model of DNA sequence evolution (Lanave et al., 1984; Rodriguez et al., 1990). Moreover, from each distance matrix we have extracted the corresponding k-th leading principal submatrices, $k \in \{10, \ldots, \max\}$, where max is 12 for Primates12, 17 for M17, 18 for M18, and 20 for the remaining datasets, generating therefore an overall number of 167 real instances of the MEP. Datasets and corresponding distance matrices can be found at `http://di.unito.it/datasetsmep2`.

We implemented both the NEA and its improved version based on the Branch-Price-and-Cut approach (BPC-NEA) in ANSI C++ by using Xpress Optimizer libraries v18.10.00. The experiments ran on a Pentium 4, 3.2 GHz, equipped with 2 GByte RAM and operating system Gentoo release 7 (kernel linux 2.6.17). During

Dataset	Taxa	Sites	Taxonomy	Dataset	Taxa	Sites	Taxonomy
Primates12	12	898	primates	M17	17	2550	insects
SeedPlant25	25	19784	pinoles	M18	18	8128	cetacea
RbcL55	55	1314	rbcL gene	M43	43	2086	mammals
Rana64	64	1976	ranoids	M62	62	3768	hyracoidae
				M82	82	2062	fungi

Table 2: Description of the datasets.

the runtime of the NEA, we activated Xpress automatic cuts, Xpress pre-solve strategy, and Xpress primal heuristic to generate the first upper bound for the FPP. In contrast, during the runtime of the BPC-NEA we deactivated Xpress automatic cuts, pre-solve strategy, and primal heuristic. When computing the overall solution time taken by both versions of the NEA to solve a specific dataset, we disregarded the overhead added by the unlabeled enumeration routine. The reason for this choice is twofold: first, such overhead is negligible as shown in Table 3, and second, the unlabeled enumeration can be performed offline once for all, as the unlabeled phylogenies are independent of the specific dataset.

In order to measure the performance of both versions of the NEA, we considered, as a reference, the performance of the mixed integer programming model introduced in Catanzaro et al., 2009, at present the only exact algorithm for MEP available in the literature. We refer to such a model as the *EPT model*, as it is based on a particular representation of a phylogeny in terms of the *Edge-Path incidence matrix of a Tree* (EPT), see, e.g., Nemhauser and Wolsey, 1999. In order to have a sound comparison of the results, we implemented the model in Catanzaro et al., 2009, on our machine.

Table 4 shows the computational results for the given datasets. The columns relative to the EPT model show respectively the running time (expressed in seconds) taken to solve a generic instance of the MEP, the number of branches needed, and the gap (expressed in percentage) i.e., the difference between the optimal value found and the value of the linear relaxation at the root node of the search tree, divided by the optimal value. When "24 hours" appears in the columns "Time", this is to highlight that the run relative to a specific instance took longer than 24 hour with respect to a specific approach. The columns "NEA Time" and "BPC-NEA Time" show the running time (expressed in seconds) taken by the NEA and the BCP-NEA, respectively, to solve a generic instance of the MEP. Finally, the last column provides the linear relaxation gap, i.e., the percentage value relative to the difference between the optimal value to the MEP and the value of the poorest linear relaxation in the list of the unlabeled phylogenies, divided by the optimal value.

As a general trend, we observe that the EPT model is unable to tackle instances of the MEP containing more than 10 taxa within 24 hours. The solution time is usually very high, ranging from 1.3 hours to 6 hours, and the gap is always above 30%. In contrast, the NEA provides a better performance both in terms of solution time and gap. Specifically, the solution time for 10 taxa ranged from 137 seconds to a maximum of 333 seconds, with gap usually not above 11%; in some cases (see datasets Rana64 and M62) the value of the gap approaches 1%, demonstrating tightness of the bounds provided by Formulation 1. Unfortunately, the solution time trend of the NEA shows that its performance dramatically deteriorate as the number of taxa increases. In fact, already with 11 taxa the solution time ranges from 944 seconds to 1.64 hours. This behavior is mainly due to the large number of y_{ij}^{pq}-variables involved in Formulation 1, which slows down the simplex algorithm and, more in general, the whole solution time of the FPP. In these circumstances, the BPC-NEA shows its advantages. Specifically, although slower than the NEA below 10 taxa, the BCP-NEA turns out to be the faster approach when tackling instances containing 11 or more taxa. Although the BCP-NEA was able to solve SeedPlant25 and Rana64 when considering 14 taxa, as a general trend the instances containing more than 13 taxa constitute the current limitation for the BPC-NEA.

The performance of the NEA approach depends on the number of taxa and the structure of the distance matrix $\mathbf{D}$. As we have already noted, the number of taxa directly affects the number of y_{ij}^{pq}-variables involved in Formulation 1, which in turn slows down the resolution of the FPP. In this context, we tried to develop an alternative model to the branch-and-price approach described in Session 4.1, which consists in substituting the four-index y_{ij}^{pq}-variables with the following three-index arc variables:

$$\omega_e^{pq} = \begin{cases} 1 & \text{if the arc } e \text{ belongs to the path that joins taxa } p,\, q \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, p, q \in \Gamma : p \neq q,\, \forall\, e \in E(U).$$

Number of Taxa	Number of Unlabeled Phylogenies	Time (sec.)	Number of Taxa	Number of Unlabeled Phylogenies	Time (sec.)
20	11020	0.07	24	254371	2.80
21	23846	0.16	25	565734	8.40
22	52233	0.39	26	1265579	24.537
23	114796	1.01	27	2841632	71.496

Table 3: Computing time necessary for enumerating the unlabeled phylogenies for a given number of taxa.

Unfortunately, we experienced very poor performance for such a model, mainly due to its linear relaxation bound, which is significantly smaller than that of the BPC-NEA and quite close to that of the EPT model.

With regard to the structure of the distance matrix $\mathbf{D}$, we observed that two main characteristics of $\mathbf{D}$ may worsen the NEA performance, namely (i) the uniformity of the entries and (ii) the "non additivity" of the entries, i.e., the tendency of the entries to not satisfy the *additive property* (Waterman et al., 1977)

$$d_{zi} + d_{kj} \leq \max\{d_{zj} + d_{ik}, d_{kz} + d_{ij}\}$$

for any four distinct taxa $i, j, k, z \in \Gamma$. Concerning the uniformity of the entries, the worst possible scenario for the NEA consists in choosing a distance matrix $\mathbf{D}$ whose non-diagonal entries are equal to a given value d. In this case, any phylogeny whose edge weights are

$$w_e^* = \begin{cases} \frac{d}{2} & \text{if } e \text{ is an external edge} \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, e \in E(U)$$

is optimal and has length $\frac{nd}{2}$. In this situation, the NEA has to enumerate all possible phylogenies in $\mathcal{T}$ before stopping, with consequent worsening of the performance.

Concerning the additive property, Waterman et al. (1977) showed that it plays a central role in solving the MEP. In fact, the authors proved that the entries of $\mathbf{D}$ can be seen as the sum of the weights of the edges of a phylogeny if and only if $\mathbf{D}$ is additive. When this situation occurs, the optimal solution to MEP is unique and it can be found in $O(n^2)$ by using a constructive greedy called the *Sequential Algorithm*. In particular, the Sequential Algorithm starts from an initial partial phylogeny of Γ, i.e., a phylogeny having a number of taxa $n' < n$, and determines the corresponding weights by using specific formulae described in Waterman et al. (1977) valid only if $\mathbf{D}$ is additive. Subsequently, the algorithm picks a taxon from among those that are not in the current partial phylogeny and constructs a new partial phylogeny with $n' + 1$ taxa by using again the previous mentioned formulae. The algorithm iterates the constructive step until the unique phylogeny of Γ is obtained. Interestingly, the literature reports on faster versions of the Sequential Algorithms, the best of which is characterized by a computational complexity $O(n \log n)$ (see Culberson and Rudnicki, 1989). Describing these versions is out of the scope of the article. We refer the interested reader to Sattah and Tversky (1977) and Culberson and Rudnicki (1989) for further information.

Unfortunately, the observed evolutionary matrices are usually not additive, as they are generally estimated by means of specific Markov substitution models of molecular evolution, such as those described in Felsenstein (2004); Hasegawa et al. (1981); Jukes and Cantor (1969); Kimura (1980); Lanave et al. (1984); Rodriguez et al. (1990); Waddell and Steel (1997) or, more rarely, by means of metric models, such as those described in Beyer et al. (1974); Kidd and Sgaramella-Zonta (1971). These estimation models provide evolutionary matrices whose entries are nonnegative and symmetric, but that not necessarily satisfy the triangle inequality. Thus, as the satisfaction of the triangle inequality is a necessary condition for additivity (Waterman et al., 1977; Semple and Steel, 2003), these matrices are usually not additive. This fact prevents the use of the Sequential Algorithm even to approximate the optimal solution to the problem, as it requires the additivity of $\mathbf{D}$ to compute the edge weights. In this context, we observed that the "closer" $\mathbf{D}$ is to an additive matrix, the easier (faster) solving the MEP. Hence, in a certain sense, the degree of additivity of a matrix can be considered as a qualitative measure of the goodness of the MEP as a phylogenetic estimation paradigm.

Dataset	Taxa	EPT Time (sec.)	EPT Branches	EPT Gap (%)	NEA Time (sec.)	BPC-NEA Time (sec.)	BPC-NEA Gap (%)
Primates12	10	4696.57	853090	38.52	**136.37**	306.78	3.43
	11	24 hours	n.a.	n.a.	943.26	**656.73**	2.28
	12	n.a.	n.a.	n.a.	4121.51	**3234.96**	3.08
SeedPlant25	10	5799.99	968639	30.45	**150.93**	259.16	9.81
	11	24 hours	n.a.	n.a.	1555.61	**1454.86**	9.34
	12	n.a.	n.a.	n.a.	8532.24	**5500.48**	9.20
	13	n.a.	n.a.	n.a.	24 hours	**40421.29**	9.56
	14	n.a.	n.a.	n.a.	n.a.	**291241.36**	8.96
	15	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
RbcL55	10	20137.01	12342575	62.64	253.21	1254.95	4.77
	11	24 hours	n.a.	n.a.	1162.42	2561.08	4.09
	12	n.a.	n.a.	n.a.	13713.91	**10424.21**	4.33
	13	n.a.	n.a.	n.a.	24 hours	**160546.89**	5.78
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Rana64	10	12695.21	2636771	54.73	227.90	**221.47**	1.91
	11	24 hours	n.a.	n.a.	1105.15	**596.58**	2.04
	12	n.a.	n.a.	n.a.	6632.99	**4328.9**	3.4
	13	n.a.	n.a.	n.a.	24 hours	**38243.53**	4.98
	14	n.a.	n.a.	n.a.	n.a.	**198152.90**	5.54
	15	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M17	10	18926.40	11638956	57.34	201.81	798.89	3.67
	11	24 hours	n.a.	n.a.	2398.24	4324.14	3.53
	12	n.a.	n.a.	n.a.	23598.71	**21037.27**	3.61
	13	n.a.	n.a.	n.a.	24 hours	**117682.58**	4.53
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M18	10	38907.91	10034561	56.64	296.37	1538.53	10.59
	11	24 hours	n.a.	n.a.	4448.06	12990.5	10.32
	12	n.a.	n.a.	n.a.	84531.39	**59220.61**	8.94
	13	n.a.	n.a.	n.a.	24 hours	**101099.87**	8.08
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M43	10	17946.91	10614369	69.99	204.27	562.55	2.97
	11	24 hours	n.a.	n.a.	1514.56	1852.55	3.01
	12	n.a.	n.a.	n.a.	24 hours	**16688.27**	3.14
	13	n.a.	n.a.	n.a.	n.a.	**107301.53**	2.52
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M62	10	10652.51	5623142	50.49	143.70	203.39	1.12
	11	24 hours	n.a.	n.a.	1100.82	**1001.76**	1.47
	12	n.a.	n.a.	n.a.	9783.43	**4575.81**	1.92
	13	n.a.	n.a.	n.a.	24 hours	**35567.51**	2.03
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M82	10	28231.91	7069863	64.66	332.49	1070.73	9.63
	11	24 hours	n.a.	n.a.	5930.71	8116.01	10.02
	12	n.a.	n.a.	n.a.	49555.67	**35981.18**	10.39
	13	n.a.	n.a.	n.a.	24 hours	**243479.21**	9.62
	14	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table 4: Numerical results obtained by the EPT model, the NEA, and the BPC-NEA on the analyzed datasets.

6 On the consistency of the MEP

There is no general way to empirically validate a candidate phylogeny for a set of molecular sequences extracted from taxa, as their *real phylogeny*, i.e., the real evolutionary process undergone by the sequences, is generally unknown. Hence, the literature proposes a number of criteria to select one phylogeny from among plausible alternatives. Each criterion usually adopts its own set of evolutionary hypotheses. The ability of such hypotheses to describe the real evolutionary process of taxa determines the gap between the real and the *true phylogeny*, i.e., the phylogeny that one would obtain under the same set of hypotheses if all molecular data from taxa were available (Felsenstein, 2004). If the optimal phylogeny of a given phylogenetic estimation criterion approaches (according to a given metric) the true phylogeny as the amount of molecular data from taxa increases, the corresponding criterion is said to be *statistically consistent* (Felsenstein, 2004).

The statistical consistency is a desirable property in molecular phylogenetics because it measures the ability of a criterion to recover the true (and hopefully the real) phylogeny of the considered taxa. The literature on molecular phylogenetics shows that in the context of distance matrix methods (see Felsenstein, 2004) a number of versions of the PEP are statistically consistent, such as the minimum evolution method under the least squares edge weight estimation model (see Rzhetsky and Nei, 1993) and the BMEP (Desper and Gascuel, 2004). In this section we show that MEP is statistically consistent as well by using a different approach from the ones used e.g., in Rzhetsky and Nei (1993); Gascuel et al. (2001); Denis and Gascuel (2003); Desper and Gascuel (2004); Gascuel (2005).

Before proceeding, we introduce some notation and definitions. For a given phylogeny T, we will use T to denote both T and its *support*, i.e., the unweighted tree underlying T. Recall that when we disregard the edge weights of T, we use terminology *unweighted phylogeny*. Note however that an unweighted phylogeny is labeled. We denote by T^* the true phylogeny of Γ and by $\mathbf{\Delta} = \{\delta_{ij}\}_{i,j} = \{\sum_{e \in P^*(i,j)} w_e^*\}_{i,j}$ the corresponding true distance matrix, where $P^*(i,j)$ denotes the path from i to j in T^* and w_e^* denotes the weight of edge e in T^*. Let $\mathbf{D}^l = \{d_{ij}^l\}_{i,j}$, for $l \in \mathbb{N}$, be the observed distance matrices, where superscript 'l' means that its entries are estimated from molecular data of length l. Similarly to Gascuel (2005), we say that $\{\mathbf{D}^l\}$ is a consistent estimate of $\mathbf{\Delta}$ if the greater the amount of molecular data of taxa we have (e.g., the longer the DNA sequences), the closer $\mathbf{D}^l$ is to $\mathbf{\Delta}$; more formally, $\lim_{l\to\infty} \|\mathbf{D}^l - \mathbf{\Delta}\|_\infty = 0$. For each l, let T^l be an optimal phylogeny for the MEP when the input distance matrix is $\mathbf{D}^l$, and let w_e^l be the weight of edge e in T^l. We say that T^l converges to T^* if the following conditions are satisfied:

1. there exists $\bar{l}$ such that, for $l \geq \bar{l}$, the support of T^l coincides with the support of T^*;
2. for every edge e of T^*,
$$\lim_{\substack{l\to\infty \\ l\geq\bar{l}}} w_e^l = w_e^*.$$

The MEP is said to be statistically consistent if whenever $\{\mathbf{D}^l\}$ is a consistent estimate of $\mathbf{\Delta}$, T^l converges to T^*.

Proposition 12. *The MEP is statistically consistent.*

Proof. For a given distance function $\mathbf{D}$ and an unweighted phylogeny T, we denote by $\mathcal{L}_{\mathbf{D}}(T)$ the minimum weight of a phylogeny whose support is T. Furthermore, we denote by $\mathcal{B}_{\mathbf{D}}(T)$ the length of a phylogeny whose support is T, computed under the BMEP objective function:

$$\mathcal{B}_{\mathbf{D}}(T) = \sum_{\substack{i,j\in\Gamma \\ i\neq j}} d_{ij} 2^{-\tau_{ij}}. \tag{23}$$

Since Δ is an additive matrix, then T^* is the unique optimal phylogeny under the BMEP when the input distance function is Δ (see Desper and Gascuel (2004)). Thus there exists $\varepsilon > 0$ such that

$$\mathcal{B}_{\mathbf{\Delta}}(T) \geq \mathcal{B}_{\mathbf{\Delta}}(T^*) + \varepsilon \tag{24}$$

for every unweighted phlyogeny $T \neq T^*$. Moreover, if we fix an unweighted phylogeny T, then, since $\{D^l\}$ is a consistent estimate of Δ and (23) is a continuous function of $\mathbf{D}$, there exists l_1 such that $\mathcal{B}_{\mathbf{D}^l}(T) \geq \mathcal{B}_{\mathbf{\Delta}}(T) - \varepsilon/2$ for every $l \geq l_1$.

Now fix an unweighted phylogeny $T \neq T^*$. Then, for $l \geq l_1$,

$$\mathcal{L}_{\mathbf{D}^l}(T) \geq \mathcal{B}_{\mathbf{D}^l}(T) \geq \mathcal{B}_{\mathbf{\Delta}}(T) - \varepsilon/2 \geq \mathcal{B}_{\mathbf{\Delta}}(T^*) + \varepsilon/2 = \mathcal{L}_{\mathbf{\Delta}}(T^*) + \varepsilon/2, \tag{25}$$

where the first inequality follows from (18), the second inequality holds because $l \geq l_1$, the third inequality follows from (24), and the equation holds because (18) is satisfied at equality for T^* (as Δ is the true distance matrix corresponding to the true phylogeny T^*).

On the other hand, since, for a fixed T, $\mathcal{L}_{\mathbf{D}}(T)$ is a continuous function of $\mathbf{D}$, there exists $l_2 \geq l_1$ such that $\mathcal{L}_{\mathbf{D}^l}(T^*) < \mathcal{L}_{\mathbf{\Delta}}(T^*) + \varepsilon/2$ for $l \geq l_2$. Together with (25), this implies that, for $l \geq l_2$, the unique optimal solution to the MEP with input distance matrix D^l has support T^*. Thus the first condition in the definition of convergence given above holds with $\bar{l} = l_2$. The second condition follows from the continuity of $\mathcal{L}_{\mathbf{D}}(T)$ when $T = T^*$ is fixed. □

Acknowledgment

RP acknowledges support from the FNRS for the "bourse missions scientifiques". Part of this work was developed while RP was visiting the Graphs and Mathematical Optimization Unit of the ULB. MDS was supported by the "Progetto di Eccellenza 2008–2009" of the "Fondazione Cassa di Risparmio di Padova e Rovigo". The authors also thank Patrick Mardulyn, Fabio Pardi, and Olivier Gascuel for helpful discussions and Rosa Maria Lo Presti for the datasets she provided.

This article is the fruit of a discussion between Mike Steel (who first proposed the term MEP) and DC started during the INI Programme PLG Workshop "Future Directions in Phylogenetic Methods and Models" in December 2007.

References

D. Applegate, R. Bixby, V. Chvátal, and W. Cook. Concorde, a solver for the traveling salesman problem. Software available at `http://www.math.princeton.edu/tsp/concorde.html`, 2001.

R. Aringhieri, P. Hansen, and F. Malucelli. Chemical trees enumeration algorithms. *4OR*, 1:67–83, 2003.

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

D. Avis and K. Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete Compututational Geometry*, 8:295–313, 1992.

D. Avis and K. Fukuda. Reverse search for enumeration. *Discrete Applied Mathematics*, 65:21–46, 1996.

D. A. Bader, B. M. E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In *SPIE ITCom 4528*, pages 159–168. SPIE, Denver, CO, 2001.

N. Beerenwinkel, R. F. Schwartz, M. Gerstung, and F. Markowetz. Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, in press, 2015.

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

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

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

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

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

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

D. Catanzaro, S. E. Schackney, and R. Schwartz. Classifying the progression of ductal carcinoma from single-cell sample data: A case study. Technical report, G.O.M. - Computer Science Department - Université Libre de Bruxelles (U.L.B.), 2013.

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

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

J. Culberson and P. Rudnicki. A fast algorithm for constructing trees from distance matrices. *Information Processing Letters*, 30:215–220, 1989.

C. D'Ambrosio, A. Lodi, and S. Martello. Combinatorial traveling salesman problem algorithms. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

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

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

M. Farach, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155–179, 1995.

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

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

W. M. Fitch. Rate of change of concomitantly variable codons. *Journal of Molecular Evolution*, 1:84–96, 1971.

N. Galtier. Maximum-likelihood phylogenetic analysis under covarion-like model. *Molecular Biology and Evolution*, 18:866–873, 2001.

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

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

F. Glover and G. A. Kochenberger. *Handbook of metaheuristics*. Kluwer Academic Publishers, Boston, MA, USA, 2003.

P. Hansen, B. Jaumard, C. Labatteux, and M. Zeng. Coding chemical trees with the centered n-tuple code. *Journal of Chemical Information and Computer Science*, 34:782–790, 1994.

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

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

J. P. Huelsenbeck. Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution*, 19: 698–707, 2002.

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

J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

D. S. Johnson, J. K. Lenstra, and A. H. G. Rinnooy Kan. The complexity of the network design problem. *Networks*, 8:279–285, 1978.

C. Jordan. Sur les assemblages des lignes. *Journal für die reine und angewandte Mathematik*, (70):185–190, 1869.

T. H. Jukes and C.R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York, 1969.

K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics*, 23:235–252, 1971.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

V. Kvasnička and J. Pospichal. Canonical indexing and constructive enumeration of molecular graphs. *Journal of Chemical Information and Computer Science*, 56:1777–1802, 1991.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

P. Lopez, D. Casane, and H. Philippe. Heterotachy: An important process of protein evolution. *Molecular Biology and Evolution*, 19:1–7, 2002.

M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper. The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404, 2003.

N. Misra, G. E. Blelloch, R. Ravi, and R. Schwartz. Generalized buneman pruning for inferring the most parsimonious multi-state phylogeny. *Journal of Computational Biology*, 18(3):445–57, 2011.

G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, New York, 1999.

C. Y. Ou, C. A. Ciesielski, G. Myers, C. I. Bandea, C. C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. Maclnnes, J. W. Curran, and H. W. Jaffe. Molecular epidemiology of HIV transmission in a dental practice. *Science*, 256(5060): 1165–1171, 1992.

L. Pachter and B. Sturmfels. The mathematics of phylogenomics. *SIAM Review*, 49(1):3–31, 2007.

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

D. S. Parker and P. Ram. The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees. *SIAM Journal on Computing*, 28(5):1875–1905, 1996.

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

G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz. Reconstructing tumor phylogenies from heterogeneous single-cell data. *Journal of Bioinformatics and Computational Biology*, 5(2a):407–427, 2006.

P. C. Pop. *Generalized network design problems: Modeling and optimization.* De Gruyter, Berlin, 2012.

R. C. Read. *Graph theory and computing.* Academic Press, New York, 1972.

M. Riester, C. Stephan-Otto Attolini, R. J. Downey, S. Singer, and F. Michor. A differentiation-based phylogeny of cancer subtypes. *PLoS Computational Biology*, 6(e1000777), 2010.

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

H. A. Ross and A. G. Rodrigo. Immune-mediated positive selection drives human immunodeficency virus type 1 molecular variation and predicts disease duration. *Journal of Virology*, 76(22):11715–11720, 2002.

A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares generalized least-squares and minimum evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35:367–375, 1992.

A. Rzhetsky and M. Nei. Theoretical foundations of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10:1073–1095, 1993.

S. Sattah and A. Tversky. Additive similarity trees. *Psychometrika*, 42:319–345, 1977.

C. Semple and M. Steel. *Phylogenetics.* Oxford University Press, New York, 2003.

C. Semple and M. Steel. Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32 (4):669–680, 2004.

S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi, and R. Schwartz. Algorithms for efficient near-perfect phylogenetic tree reconstruction in theory and practice. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):561–571, 2007.

S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi, and R. Schwartz. Mixed integer linear programming for maximum parsimony phylogeny inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3):323–331, 2008.

A. Subramanian, S. Shackney, and R. Schwartz. Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(6):1422–1431, 2013.

N. Trinajstić, S. Nicolić, J. V. Knop, W. R. Müller, and K. Szymanski. *Computational chemical graph theory.* Ellis Horwood, New York, 1991.

P. J. Waddell and M. A. Steel. General time-reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. *Molecular Phylogenetics and Evolution*, 8: 398–414, 1997.

M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64:199–213, 1977.