Mathematical Models to Reconstruct Phylogenetic Trees Under the Minimum Evolution Criterion

Daniele Catanzaro and Martine Labbé
Graphes et Optimisation Mathématique, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium

Raffaele Pesenti
Dipartimento di Matematica Applicata, Universitá Ca' Foscari, Dorsoduro 3246 - 30123, Venice, Italy

Juan-José Salazar-González
Departamento de Estadística, Investigación Operativa y Computación, Universidad de La Laguna, Av. Astrofísico Francisco Sánchez, 38271 La Laguna, Tenerife, Spain

A basic problem in molecular biology is to rebuild phylogenetic trees (PT) from a set of DNA or protein sequences. Among different criteria used for this purpose, the minimum evolution criterion is an optimality based criterion aiming to rebuild PT characterized by a minimal length. This problem is known to be $\mathcal{NP}$-hard. We introduce in this article some mixed integer programming models, and we also study possible cuts and lower bounds for the optimal value. So far, the number of sequences that can be involved in optimal phylogenetic reconstruction is still limited to 10. © 2008 Wiley Periodicals, Inc. NETWORKS, Vol. 53(2), 126–140 2009

Keywords: phylogenetic estimation; minimum evolution criterion; network design; branch–and–bound

1. INTRODUCTION

A phylogenetic tree (PT) models the phylogeny of an observed set of species (*taxa*) and of their unobserved common ancestors. Its leaves represent the observed taxa, its internal vertices represent the common ancestors, and each edge weight represents the *evolutionary distance* between the pair of vertices connected by the associated edge.

In this article we aim at determining the PT of a set Γ of n taxa given the knowledge of a $n \times n$ symmetric *distance matrix* $\mathbf{D} = \{d_{ij}\}$ of estimated evolutionary distances between each pair i, j of taxa in Γ. In doing so, we require that the PT is a *binary phylogenetic X-tree* (see e.g. [25]): it is unrooted and it has all its internal vertices with degree three. Trees with such a structure are also known as *3-Cayley trees*, *trivalent trees* or *boron trees*. In addition, we make use of the *Minimum Evolution* (ME) *criterion* [14]. ME affirms that, when unbiased estimates of the distances among the pairs of sequences are available, the tree must be the shortest among all the weighted PTs compatible with the distances in $\mathbf{D}$ (see e.g. [23, 24]), i.e. the cumulative weight through the path connecting two taxa along the tree is not smaller than the estimated distance between these two taxa. The distances between taxa are input data to our work, and they are usually computed on the basis of a given markovian substitution model of molecular evolution (e.g., those described in [5, 9, 11, 13, 15, 17, 21, 26]) or, more rarely, by means of metric models (e.g., those described in [3, 14]). Figure 1 shows an example of distance matrix between ten taxa, and Figure 2 displays a PT for this instance. The numbers in Figure 2 are the edge weights, and they are (together with the tree structure) the output of the problem addressed in this article. Figure 3 shows different weights for the same tree structure of Figure 2, both feasible solutions for the problem associated with Figure 1 and both with identical total weight of 3.58.

To formally state the problem of our interest, a PT is a edge-weighted tree $T = (V, \mathcal{E})$ with V and $\mathcal{E}$ the set of vertices and edges, respectively. In particular, let $V = V_{\text{ex}} \cup V_{\text{in}}$, where V_{ex} is the set of n leaves representing the n taxa in Γ, and V_{in} is the set of $(n-2)$ internal vertices representing the common ancestors of the taxa in Γ. Accordingly, let

Received March 2007; accepted April 2008
Correspondence to: J.-J. Salazar-González; e-mail: jjsalaza@ull.es
Contract grant sponsor: Belgian National Fund for Scientific Research (FNRS)
Contract grant sponsor: Communauté française de Belgique — Actions de Recherche Concertécs (ARC)
Contract grant sponsor: Ministerio de Educación y Ciencia; Contract grant number: MTM2006-14961-C05-03
DOI 10.1002/net.20281
Published online 10 October 2008 in Wiley InterScience (www.interscience.wiley.com).

	B	C	D	E	F	G	H	I	J
A	0.831	0.928	0.831	0.925	0.847	1.321	1.326	1.314	1.121
B	-	0.414	0.013	0.411	0.275	1.296	1.274	1.290	1.166
C	-	-	0.413	0.176	0.441	1.256	1.233	1.264	1.218
D	-	-	-	0.411	0.275	1.291	1.267	1.288	1.160
E	-	-	-	-	0.443	1.255	1.233	1.258	1.212
F	-	-	-	-	-	1.300	1.251	1.269	1.154
G	-	-	-	-	-	-	1.056	1.067	1.348
H	-	-	-	-	-	-	-	0.315	1.456
I	-	-	-	-	-	-	-	-	1.437

FIG. 1. Jukes—Cantor pairwise distance estimates (input).

$\mathcal{E} = \mathcal{E}_{\text{ex}} \cup \mathcal{E}_{\text{in}}$, where $\mathcal{E}_{\text{ex}}$ is the set of n external edges, i.e., edges which have an extreme in a leaf, and $\mathcal{E}_{\text{in}}$ is the set of the remaining $(n-3)$ internal edges. Then an $n(n-1)/2 \times (2n-3)$ matrix $\mathbf{X} = \{x_{ij,e} : i,j \in V_{\text{ex}}\ s.t.\ i < j, e \in \mathcal{E}\}$ can represent a PT as follows: the generic entry $x_{ij,e}$ is equal to 1 if the edge e belongs to the path in the PT from the leaf i to the leaf j, 0 otherwise ([19], p. 550–559). The matrix $\mathbf{X}$ is called *Edge-Path incidence matrix of a Tree* (EPT). Denote $\mathbf{w}$ as the $(2n-3)$ vector of the edge weights of the PT, and $\mathbf{D}^{\triangle}$ as the $n(n-1)/2$ vector whose components are obtained by taking row by row the entries of the strictly upper triangular matrix of $\mathbf{D}$. Using the above notation and given V and $\mathbf{D}$ as input, the optimization problem of our interest is named *Minimum Evolution Problem* (MEP for short) and consists in finding $\mathcal{E}$ (or equivalently $\mathbf{X}$) and $\mathbf{w}$ such that:

Model 1.

$$\min_{\mathbf{X},\mathbf{w}} z = \|\mathbf{w}\|_1$$

$$s.t.\ \mathbf{X}\mathbf{w} \geq \mathbf{D}^{\triangle}$$

$$\mathbf{w} \geq 0$$

$$\mathbf{X} \in \mathcal{X}$$

where $\|\cdot\|_1$ is the $\mathcal{L}^1$ vector-norm and $\mathcal{X}$ is the set of the edge-path incidence matrices for the $(2n-5)!! = (2n-5) \cdot (2n-7) \cdot \ldots 3 \cdot 1$ different possible topological structures for a PT with n leaves ([9], p. 25).

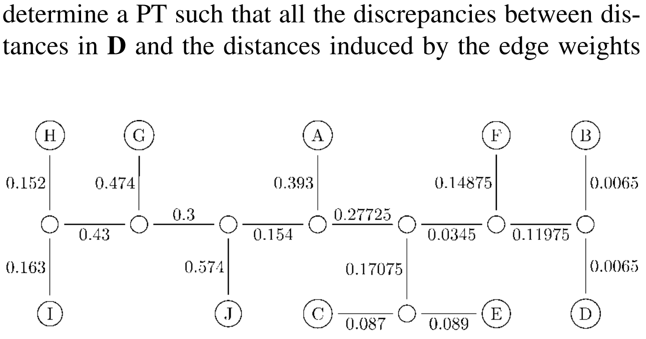

FIG. 2. PT under the ME criterion (output).

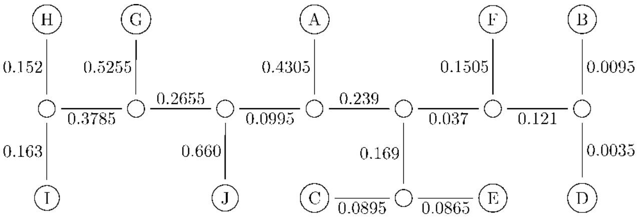

FIG. 3. Alternative edge weights for the same binary tree of Figure 2.

MEP was introduced in [3, 27]. Its solution is a PT with non-negative edge weights which guarantees that the evolutionary distances between each pair of taxa are not less than the estimated ones. Observe that, in general, the linear problem $\mathbf{X}\mathbf{w} = \mathbf{D}^{\triangle}$, $\mathbf{w} \geq 0$ is infeasible (i.e., it is impossible to determine a PT such that all the discrepancies between distances in $\mathbf{D}$ and the distances induced by the edge weights are zero). For this reason, many authors (see e.g. [6] and [9]) interpret ME as we do in this. Some authors require that an optimal PT is a binary tree defined by a matrix $\mathbf{X}$ and associated with a vector $\mathbf{w}$ such that the sum of the square values of the above mentioned discrepancies is minimum. In particular, Rzhetsky and Nei [22] propose a problem that differs from MEP as constraints $\mathbf{X}\mathbf{w} \geq \mathbf{D}^{\triangle}$, $\mathbf{w} \geq 0$ are substituted by the least square optimality conditions:

$$\mathbf{w} = \mathbf{X}^{\dagger}\mathbf{D}^{\triangle}, \tag{1}$$

where $\mathbf{X}^{\dagger}$ is the Moore–Penrose's pseudoinverse of $\mathbf{X}$. Similarly, Beyer et al. [3] and Fitch and Margoliash [10] consider a weighted least squares approach and impose condition:

$$\mathbf{w} = (\mathbf{X}^t\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Q}\mathbf{D}^{\triangle} \tag{2}$$

where $\mathbf{Q}$ is a given strictly positive definite diagonal matrix whose elements q_{ij} represent weights associated with the pair of taxa i and j. Finally, Hasegawa et al. [12] introduce a generalized least squares function and impose condition:

$$\mathbf{w} = (\mathbf{X}^t\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{C}^{-1}\mathbf{D}^{\triangle} \tag{3}$$

where $\mathbf{C}$ is a strictly positive definite symmetric matrix representing the covariance matrix of $\mathbf{D}$.

Unfortunately, MEP is $\mathcal{NP}$-hard as well as all its least square variations (see e.g. [8]). Moreover, these last variants have the additional drawback that some negative weights may occur in their optimal solutions.

In the literature, to the best of authors' knowledge, no solution strategy is known for MEP whereas its least square variations are solved either by a brute force approach, which enumerates all of the possible EPT matrices, or by heuristic approaches, which are efficient but do not guarantee the minimality of the solution found. We remit the reader to [4] for a recent survey on the solution methods.

Herewith we introduce some basic mixed integer programming models for MEP. Unfortunately, none of the models provides the optimal PT in a reasonable time when the number of taxa in Γ is greater than 10. Using Xpress to solve all the models, instances containing 10 taxa need more than 3 h to be exactly solved on a Intel Core 2 Duo 2 GHz PC with 2 GB of RAM. For this reason, we also study possible cuts for the proposed models and some possible lower bounds for MEP. So far, the problem of finding an optimal PT for a set Γ with $n > 10$ taxa is still difficult.

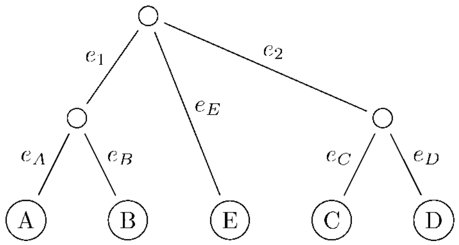

FIG. 4. An example of phylogenetic tree with five taxa (species).

The remainder of the article is organized as follows. Section 2 introduces some mixed-integer programming models to solve MEP. In particular, Section 2.1 studies and exploits the structure of the EPT representation of a PT to develop an *ad hoc* linear model characterized by a polynomial number of integer and continuous variables. The second model, introduced in Section 2.2, is characterized by the smallest polynomial number of integer and continuous variables but by two exponential sets of constraints. The third model shown in Section 2.4 is characterized by a polynomial number of integer and continuous variables but needs a nontrivial nonisomorphic tree generator. Section 3 investigates some combinatorial lower bounds. Section 4 shows the computational results obtained by some of our models on real instances.

2. MIXED-INTEGER PROGRAMMING MODELS

In this section, we introduce possible mixed-integer programming models for MEP.

2.1. Edge-Path Based Model

We consider here models that derive from the linearization of Model 1. Initially, we study the structure and properties of the EPT matrices, subsequently exploited to impose linear conditions on the values of the entries of these matrices. Then, we introduce the linear models.

2.1.1. Structure of the Edge-Path Matrices. There exist many EPT matrices describing the same PT. Indeed, given an EPT matrix $\mathbf{X}$ associated with the PT, any other EPT matrix $\hat{\mathbf{X}}$ obtainable from $\mathbf{X}$ by swapping its rows and/or columns still describes the same PT. Therefore, we will restrict the set of EPT matrices of our interest to obtain a one-to-one correspondence between EPT matrices and PTs. Then, we will show that any EPT matrix presents redundant information, i.e., the knowledge of a relatively small number of entries of an EPT matrix is sufficient to describe fully the associated PT.

Throughout this and the following sections, we assume that the set Γ is ordered lexicographically according to the names of the taxa, and that the names of the taxa in this sequence are A, B, C, and so on. In addition, we indicate each leaf of a PT with the name of the corresponding taxon. To help the reader, throughout this section we illustrate our arguments by an example: we consider a set of five species $\Gamma = \{A, B, C, D, E\}$ and a possible PT in Figure 4.

Let $\mathbf{X}$ be a generic EPT matrix that describes the generic phylogenetic tree T. By definition, $\mathbf{X}$ is a 0/1 matrix that presents a row for each path joining two leaves i, j, and a column for each edge of T. We can assume without loss of generality that the rows of $\mathbf{X}$ are always ordered lexicographically on the basis of the order in Γ. By referring to the PT in Figure 4, the first row of $\mathbf{X}$ describes the path between taxa A and B, the second row the path between taxa A and C, and so on. We can also assume that the first n columns of $\mathbf{X}$ correspond to the external edges of T and that they are sorted according to the order of the taxa at one of their extremes. In Figure 4, the column associated with external edge e_A precedes column associated with external edge e_B. Finally, we assume that the remaining $(n-3)$ columns of $\mathbf{X}$, corresponding to the internal edges of T, are sorted according to a relation defined in the following way. Given a generic external edge e, define $\text{dist}(e)$ to be the topological distance of e from the leaf associated with taxon A. In Figure 4, $\text{dist}(e_1)$ is equal to 2, whereas $\text{dist}(e_2)$ is equal to 3. In addition, define $\text{path}(e)$ the first path, from a lexicographical point of view, to which e belongs. In Figure 4, both $\text{path}(e_1)$ and $\text{path}(e_2)$ return the path between taxa A and C. Then, we impose that the column associated with the internal edge e_1 precedes the column associated with the internal edge e_2 in $\mathbf{X}$ if one of the following two conditions holds: $\text{path}(e_1)$ lexicographically precedes $\text{path}(e_2)$ or $\text{path}(e_1) = \text{path}(e_2)$ and $\text{dist}(e_1) < \text{dist}(e_2)$. This order relation is complete as the lexicographic order is complete in the path set, and in a tree we cannot have $\text{path}(e_1) = \text{path}(e_2)$ and $\text{dist}(e_1) = \text{dist}(e_2)$. In the rest of the article, we assume that all the EPT matrices satisfy the above order relations for their rows and columns. The EPT matrix $\mathbf{X}$ describing the PT in Figure 4 is reported in Figure 5.

We show now that it is possible to decompose any EPT matrix $\mathbf{X}$ into blocks, so that only a subset of the entries of $\mathbf{X}$ are necessary to describe a PT. Decompose a generic edge-path incidence matrix of a tree $\mathbf{X}$ as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{E}^{\otimes} & \mathbf{F}^{\otimes} \end{pmatrix} \tag{4}$$

	e_A	e_B	e_C	e_D	e_E	e_1	e_2
AB	1	1	0	0	0	0	0
AC	1	0	1	0	0	1	1
AD	1	0	0	1	0	1	1
AE	1	0	0	0	1	1	0
BC	0	1	1	0	0	1	1
BD	0	1	0	1	0	1	1
BE	0	1	0	0	1	1	0
CD	0	0	1	1	0	0	0
CE	0	0	1	0	1	0	1
DE	0	0	0	1	1	0	1

FIG. 5. Example of EPT matrix associated with the phylogenetic tree shown in Figure 4.

where **E** is the incidence matrix of the external edges on the paths with an extreme in leaf A, **F** is the incidence matrix of the internal edges on the paths with an extreme in leaf A, $\mathbf{E}^{\otimes}$ is the incidence matrix of the external edges on the paths not involving leaf A, $\mathbf{F}^{\otimes}$ is the incidence matrix of the internal edges on the paths not involving leaf A.

Matrix **E** is $(n-1) \times n$. Since the rows of this matrix are the incidence vectors of the external edges on the $(n-1)$ paths with an extreme in the leaf A, then it is the juxtaposition of a column vector whose entries are all equal to 1 and the $(n-1)\times(n-1)$ identity matrix. As an example, **E** in Figure 5 is the following 4×5 matrix

$$\mathbf{E} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Note also that the structure of matrix **E** is common to all the $(2n-5)!!$ tree-topologies that can be obtained with n leaves.

Matrix $\mathbf{E}^{\otimes}$ is $\frac{(n-1)(n-2)}{2} \times n$. As its rows are the incidence vectors of the external edges on the paths not involving A, its includes redundant information. Indeed, in a tree an edge e belongs to the path between i and j if either e belongs to the path between A and i or to the path between A and j, but not to both paths. Then the generic entry $x_{ij,e}$ of $\mathbf{E}^{\otimes}$ denoting whether the internal edge e belongs to the path between i and j is determined unequivocally from the entries of **E** as follows

$$x_{ij,e} = x_{Ai,e} \otimes x_{Aj,e} \tag{6}$$

where $\otimes$ is the exclusive-or operator. As an example, $\mathbf{E}^{\otimes}$ in Figure 5 is the following 6×5 matrix

$$\mathbf{E}^{\otimes} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{7}$$

By construction also $\mathbf{E}^{\otimes}$ is common to all the tree-topologies that can be obtained with n leaves. Matrix $\mathbf{F}^{\otimes}$ is $\frac{(n-1)(n-2)}{2} \times (n-3)$. Together with **F**, $\mathbf{F}^{\otimes}$ is different for each PT. However, as it happens with $\mathbf{E}^{\otimes}$, matrix $\mathbf{F}^{\otimes}$ includes redundant information. Indeed, the generic entry $x_{ij,e}$ of $\mathbf{F}^{\otimes}$, denoting whether the internal edge e belongs to the path between i and j, is determined unequivocally from the entries of **F** as in (6). As an example, $\mathbf{F}^{\otimes}$ in Figure 5 is the following 6×2 matrix

$$\mathbf{F}^{\otimes} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \tag{8}$$

Finally, matrix **F** is $(n-1) \times (n-3)$. Its rows are the incidence vectors of the internal edges on the paths from the leaf A. Matrix **F** includes the necessary information sufficient to describe the topology of a tree. As an example, **F** in Figure 5 is the following 4×2 matrix

$$\mathbf{F} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}. \tag{9}$$

The values of the entries of **F** are not completely free, i.e., not all 0–1 matrices describe a PT. The following theorem states the conditions under which **F** is feasible:

Theorem 1. *Consider a matrix* **X**, *whose submatrices* **E**, $\mathbf{E}^{\otimes}$, $\mathbf{F}^{\otimes}$ *are structured as previously described. Matrix* **X** *is the EPT matrix of a phylogenetic tree* T *if and only if all the following conditions hold on submatrix* **F**:

1. *It does not include any of the following* 2×2 *submatrices*

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{10}$$

or

$$\mathbf{M}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{11}$$

where the columns and the rows whose intersection defines $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are not necessarily adjacent in* **F**.
2. *Any of its columns has at least two entries equal to 1.*
3. *If columns* r *and* s, *with* $r < s$, *have an entry equal to 1 in the same row then the number of entries equal to 1 in column* r *is greater or equal than the number of entries equal to 1 in column* s.

Proof. Necessity.

1. Let e_1 (respectively e_2) be the edge of T associated with the first (respectively the second) column of $\mathbf{M}_1$ or $\mathbf{M}_2$. The presence of $\mathbf{M}_1$ or $\mathbf{M}_2$ in **F** indicates the contradictory fact that, in the forest T, one can reach edge e_2 from leaf A along two different paths since only one includes e_1 as a predecessor of e_2.
2. Observe that, as T must be a connected tree, e must belong to at least one path from a leaf A. Then, note that if e is an internal edge, then it must belong to at least two paths that join A to two other different taxa. Hence the column of **F** associated with e must include at least two entries equal to 1.
3. If columns r and s have entry equal to 1 in some row, $r < s$, then e_r and e_s belongs to the same path. Then e_r belongs to all paths from A containing e_s, therefore Condition 3 holds.

Sufficiency. Note that the columns of matrix **X** corresponding to **E** and $\mathbf{E}^{\otimes}$ describe a star graph with all the taxa for leaves, hence matrix **X** always represents a connected network. Hereafter assume that a path P_{Aj} from vertex A to vertex $j \in V_{\text{ex}} \setminus \{A\}$ is built in the following way: initially $P_{Aj} = \{e_A\}$; subsequently edge $e_f, f \in \{1, \ldots, n-3\}$, is appended to P_{Aj} if the corresponding entry in the j-th row of matrix **F** is set to 1; finally, edge e_j is added. This construction implies that any two edges of the graph have a vertex in common only if they appear consecutively on the path from taxon A. Since Condition 2 holds there does not exist an internal edge belonging only to a single path $P_{Aj}, j \in V_{\text{ex}} \setminus \{A\}$, i.e., there does not exist an internal vertex incident to an external edge and having degree 2. Analogously, since Condition 3 holds it does not exist an internal vertex incident to only two internal edges. Then proving that **F** represents a PT means to prove that: (i) all the paths described by matrix **F** present no cycle, and (ii) it does not exist an internal vertex having degree greater than 3.

Assume that two distinct paths P_{Ai} and P_{Aj} are given, and that the first q edges contained in these paths are equal. Since the two paths are distinct there exists at least one edge, say e_{q+1}, that is included in P_{Ai} and not in P_{Aj}. If P_{Ai} and P_{Aj} form a cycle this means that they share a common edge e_w, $w > q + 1$, since they diverged, and this fact would lead to the presence of matrix $\mathbf{M}_1$ (or $\mathbf{M}_2$) that is forbidden by Condition 1.

Finally, since the overall number of columns in the matrix **X** is $(2n - 3)$, the sum of the degrees of the vertices involved in the graph is $2(2n - 3)$. Moreover, since the graph is a tree, there exist $2n - 2$ vertices of which n are terminal vertices (ending points of the corresponding paths). The sum of the degrees of the remaining internal vertices is $3(n - 2)$, i.e., the average degree of each internal vertex is 3. Since Conditions 2 and 3 prevent the existence of internal vertices having degree smaller than 3, it follows that the degree of each internal vertex is exactly 3. ■

For each edge $e \in \mathcal{E}_{\text{in}}$ and for each pair $i, j \in V_{\text{ex}} : i < j$, let $x_{ij,e}$ the entries of the submatrices **F** and $\mathbf{F}^{\otimes}$ of the EPT matrix **X**. Then, we can express the relations between the entries **F** and $\mathbf{F}^{\otimes}$ and the conditions of Theorem 1 through a set of linear inequalities. The equivalence of conditions (6) for the generic entry $x_{ij,e}$ of $\mathbf{F}^{\otimes}$ become:

$$\begin{aligned} x_{ij,e} &\le x_{Ai,e} + x_{Aj,e} \\ x_{ij,e} &\le 2 - x_{Ai,e} - x_{Aj,e} \\ x_{ij,e} &\ge x_{Ai,e} - x_{Aj,e} \\ x_{ij,e} &\ge -x_{Ai,e} + x_{Aj,e}. \end{aligned}$$

Condition 1 of Theorem 1, for each pair of rows Ai and Aj, and each pair of columns e and f, with $e < f$, in matrix **F**, becomes:

$$\begin{aligned} x_{Ai,e} - x_{Aj,e} + x_{Ai,f} + x_{Aj,f} &\le 2 \\ -x_{Ai,e} + x_{Aj,e} + x_{Ai,f} + x_{Aj,f} &\le 2. \end{aligned}$$

Condition 2 of Theorem 1, for each column e in submatrix **F**, becomes:

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,e} \ge 2$$

and in particular

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,f} = 2$$

if f is the last column of **F**. Indeed, the last column of **F** can have only two 1's as one of its extreme must always be in common with two external edges.

Condition 3 of Theorem 1 for the first column, say e, of **F** becomes:

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,e} \le n - 2$$

whereas, the same condition for each $j \in V_{\text{ex}} \setminus \{A\}$, and pair of columns r and s in submatrix **F**, $r < s$, becomes:

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,s} \le \sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,r} - 1 + (n - 1)(2 - x_{Aj,r} - x_{Aj,s}).$$

2.1.2. Edge-Path Model.

We are now ready to formulate MEP in a way to exploit the previously cited structure and properties of the EPT matrices. To this end, assume a set Γ of n taxa is given together with the corresponding distance matrix **D**.

For each edge $e \in \mathcal{E}$ and, when necessary, for each pair i, $j \in V_{\text{ex}} : i < j$, we introduce the following decision variables: $x_{ij,e}$ represents a binary entry of either submatrix **F** or $\mathbf{F}^{\otimes}$; w_e represents a non-negative weight of edge e; $v_{ij,e}$ is the minimal (non-negative) weight that e must assume so that the length of the path between i and j is not less than the distance d_{ij}. MEP is equivalent to:

Model 2.

$$\min z = \sum_{e \in \mathcal{E}} w_e \tag{12}$$

$$s.t.\ v_{ij,e} \le w_e \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} \ : \ i < j \tag{13}$$

$$w_i + w_j + \sum_{e \in \mathcal{E}_{\text{in}}} v_{ij,e} \ge d_{ij} \quad \forall\, i,j \in V_{\text{ex}} \ : i < j \tag{14}$$

$$v_{ij,e} \le d_{ij} x_{ij,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} \ : \ i < j \tag{15}$$

$$x_{ij,e} \le x_{Ai,e} + x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall\, i,j \in V_{\text{ex}} \ : \ i < j \tag{16}$$

$$x_{ij,e} \le 2 - x_{Ai,e} - x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \tag{17}$$

$$x_{ij,e} \ge x_{Ai,e} - x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall\, i,j \in V_{\text{ex}} \ : \ i < j \tag{18}$$

$$x_{ij,e} \ge -x_{Ai,e} + x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \tag{19}$$

$$x_{Ai,f} + x_{Aj,f} + x_{Ai,e} - x_{Aj,e} \le 2 \quad \forall e \in \mathcal{E}_{\text{in}}, \forall\, i,j \in V_{\text{ex}} \ : i<j \tag{20}$$

$$x_{Ai,f} + x_{Aj,f} - x_{Ai,e} + x_{Aj,e} \le 2 \quad \forall e \in \mathcal{E}_{\text{in}}, \forall\, i,j \in V_{\text{ex}} \ : i<j \tag{21}$$

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,e} \ge 2 \quad \forall e \in \mathcal{E}_{\text{in}} \tag{22}$$

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,(2n-3)} = 2 \tag{23}$$

$$\sum_{i \in V_{\text{ex}} \setminus \{A\}} x_{Ai,(n+1)} \le n - 2 \tag{24}$$

$$\sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,s} \leq \sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,r} - 1 + (n-1)(2 - x_{Aj,r} - x_{Aj,s})$$
$$\forall r,s \in \mathcal{E}_{\text{in}}, r < s, \forall j \in V_{\text{ex}} \setminus \{A\} \quad (25)$$

$$w_e, v_{ij,e} \geq 0 \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \quad (26)$$

$$x_{ij,e} \in \{0,1\} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j. \quad (27)$$

Objective function (12) imposes that the PT has minimum length. Constraints (13) and (14) require that the length of each path between i and j is greater than or equal to d_{ij}. Constraints (15) impose that the lower bound on the edge e can be applied only if e belongs to the path between i and j. Constraints (16)–(25) impose that entries $\{x_{ij,e}\}$ represent a PT as described in Section 2.1.1.

The model is characterized by a relatively small number $O(n^2)$ of binary variables and $O(n^3)$ of continuous variables. Indeed, variables x are $O(n^3)$, but for imposing conditions of Theorem 1 we need to impose the integer constraints only on the $O(n^2)$ variables corresponding to the entries of **F**. The bound obtained by considering the linear programming relaxation of the above model is in general very poor, mainly for the presence of conditions like (15). Such bound is usually equal to the greatest estimated distance between two taxa, $\hat{d} = \max\{d_{ij}\}$.

We can eliminate the variables v through the Fourier-Motzkin approach ([18], p. 39–46). In such a case, we substitute constraints (13)–(15) with the following ones:

$$w_i + w_j + \sum_{e\in F} w_e + \sum_{e\in\mathcal{E}\setminus F} d_{ij}x_{ij,e} \geq d_{ij}$$
$$\forall F \subseteq \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j, \quad (28)$$

thus leading to another equivalent formulation for MEP.

Let P_{ij} be the set of the edges of a generic path between i and j. For $F = P_{ij}$, (28) imposes that the weight of the path between i and j is not less than d_{ij}. For $F \supset P_{ij}$, (28) imposes a condition weaker than the previous one, because the weight of some extra edge not in the path between i and j is also considered. Finally, when $F \subset P_{ij}$ then $\mathcal{E} \setminus F$ includes at least an edge of P_{ij}, and (28) holds for all the integer values of x, because at least a variable $x_{ij,e}$ is equal to 1 for some $e \in \mathcal{E} \setminus F$. The number of constraints (28) is exponential but implementing an efficient separation procedure is quite trivial: let a solution (x^*, w^*) be given; define the corresponding set $\hat{F}$ as follows $\hat{F} = \{e \in \mathcal{E} : d_{ij}x^*_{ij,e} < w^*_e\}$. If it holds that $w^*_i + w^*_j + \sum_{e\in\mathcal{E}\setminus\hat{F}} w^*_e + \sum_{e\in\hat{F}} d_{ij}x^*_{ij,e} < d_{ij}$ we have found a violated cut; otherwise, all inequalities (28) are satisfied by the current solution.

2.1.3. Strengthening the Edge-Path Model.

In this section, we introduce some properties of the PTs that may turn useful in defining valid inequalities for the description of the set of feasible edge-path incidence matrices used in Model 2.

Let a generic phylogenetic tree T be given and **X** its edge-path incidence matrix characterized by a submatrix **F**. As previously seen, the order relation for the columns of **F** limits the number of 1's in the last column of such a matrix. Generalizing the above result we obtain:

$$\sum_{j\in V_{\text{ex}}\setminus\{A\}} x_{Aj,e} \leq (n-1-e) \quad e = 1,\ldots,(n-3). \quad (29)$$

To prove the above condition, consider a generic internal vertex i, extreme to the internal edge e. In particular, let i be the farthest of the two extremes of e from vertex A. The order relation implies that vertex i can be seen as the root of a binary tree that includes at maximum $(n-3-e)$ internal edges. As a rooted binary tree with $(n-3-e)$ internal edges it has $(n-1-e)$ leaves then (29) holds. Constraints (29) is tight when T is a comb.

Another immediate consequence of the order relation on the columns of **F** is that such a matrix has two identical rows for each column e whose sum of its entries is equal to 2. Let $x_{Ai,e}$ and $x_{Aj,e}$ be the only two entries of e equal to 1. Then, by Condition 1 of Theorem 1 it must hold that $x_{Ai,f} = x_{Aj,f}$ for any other column f preceding e in **F**. On the other hand, by Condition 3 of Theorem 1 it must hold that $x_{Ai,f} = x_{Aj,f} = 0$ for any other column f succeeding e.

Since (12) guarantees that the last column of **F** has exactly two entries equal to 1, **F** has at least two identical rows. On the other hand, note that $x_{Ai,e}$ and $x_{Aj,e}$ are the only two entries of column e equal to 1, if the external edges of leaves i and j are adjacent. Then, **F** can have no more than $\left(\frac{n}{2} - 1\right)$ (respectively $\frac{n-1}{2}$) columns with exactly two entries equal to 1, if n is even (respectively odd).

By definition of the EPT matrix, **F** may not have more than one null row. Such a situation occurs if the external edge with an extreme in leaf A is adjacent to another external edge and to an internal edge.

Let us now introduce constraints that bound the number of 1's present in **F**. To this end, denote as i the internal vertex extreme to the external edge of leaf A. Observe that the number of 1's in each generic row of **F** associated with the path between leaves A and j is, by definition, equal to the length, expressed in term of number of internal edges that included in the path between i and j. Consequently, **F** includes the minimal number of 1's if it is associated with a balanced binary tree rooted in i, whereas **F** includes the minimal number of 1's if it is associated with a degenerate binary tree rooted in i, whose internal edges describe a single path of length $(n-3)$, and the PT is a comb (see Fig. 6). Hence, the following condition holds:

$$(n-1)\lceil log_2(n-1)\rceil - 2^{\lceil log_2(n-1)\rceil}$$
$$\leq \sum_{e\in V_{\text{in}}} \sum_{j\in V_{\text{ex}}\setminus\{A\}} x_{Aj,e} \leq \frac{n^2-3n}{2}. \quad (30)$$

The second inequality in (30) can also be seen as a trivial consequence of (29).

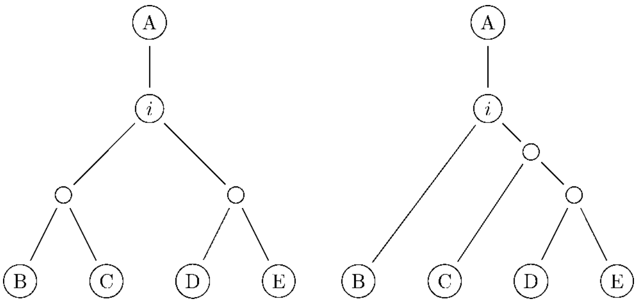

FIG. 6. An example of balanced phylogenetic tree and comb seen from vertex i, ancestor of taxon A.

2.2. All-Paths Model

In this section, we introduce a model that identifies an optimal phylogenetic tree T for the set Γ of taxa as a minimum spanning tree satisfying some side constraints of the graph described as follows. Let $G = (V, E)$ be a graph where the set $V = V_{\text{ex}} \cup V_{\text{in}}$, and the edges in E induce a clique on the $(n-2)$ internal vertices and a fan from each internal vertex to all the external vertices. By definition, the set E has cardinality $\frac{(n-2)(n-3)}{2} + (n-2)n = 3\frac{(n-1)(n-2)}{2}$, and includes no edge joining two external vertices.

For each edge $e \in E$, we introduce the following decision variables: y_e binary variable equal to 1 if edge e belongs to T, w_e non-negative weight of edge e.

The following notation is used. For a subset $S \subseteq V$, $E(S) \subseteq E$ denotes the subset of edges induced by the vertices in S; $\delta(i) \subseteq E$ denotes the subset of edges incident in the vertex $i \in V$; for $\hat{E} \subseteq E$, $y(\hat{E}) = \sum_{e \in \hat{E}} y_e$; $P_{ij} \subseteq E$ denotes the set of edge of a generic (simple) path joining the external vertices i and j; $\mathcal{P}_{ij} = \{P_{ij}\}$ denotes the set of all possible P_{ij} in G. Then, we can formulate MEP as follows:

Model 3.

$$\min z = \sum_{e \in E} w_e \tag{31}$$

$$s.t.\ y(E(S)) \leq |S| - 1 \quad \forall S \subset N \tag{32}$$

$$y(E(V)) = 2n - 3 \tag{33}$$

$$y(\delta(i)) = 3 \quad \forall i \in V_{\text{in}} \tag{34}$$

$$w_e \leq \hat{d} y_e \quad \forall e \in E \tag{35}$$

$$\sum_{e \in P_{ij}} (w_e + d_{ij}(1 - y_e)) \geq d_{ij} \quad \forall P_{ij} \in \mathcal{P}_{ij}, \forall i,j \in V_{\text{ex}} : i < j \tag{36}$$

$$y_e \in \{0, 1\} \quad \forall e \in E \tag{37}$$

$$w_e \geq 0 \quad \forall e \in E. \tag{38}$$

Constraints (32)–(34) and (37) impose that variables y_e define a tree. Constraints (35) impose that the weight of a generic edge e is positive only if edge e is an edge of the tree. Recall $\hat{d} = \max\{d_{ij}\}$. Finally, constraints (36) impose that the sum of the weights belonging to a path P_{ij} on the tree T is not less than d_{ij}.

This model is characterized by a relatively small number $O(n^2)$ of binary and continuous variables, and by an exponential number of constraints. Two separation procedures are therefore necessary. To exclude solutions violating the packing constraints (32) it is sufficient to apply the procedure described in [2], whereas to exclude solutions violating constraints (36) it is sufficient to apply the Floyd-Warshall's algorithm in a direct network having leaf A as source vertex.

The bound obtained by solving the linear programming relaxation of the above model can be very poor, not only due to the conditions such as (35), but also because it does not exclude half-integer fractional solutions with respect to the edge variables. For example, the convex combination, with weight $1/2$, of the incident vectors of the two complementary PT's in Figure 7 is a feasible solution for the linear relaxation of (31)–(38).

2.3. Flow Model

This section shows a different model where the non-negative weights are represented as unknown potentials to ensure the minimum distance between each pair of leafs.

As before, each edge $e \in E$ is associated with a 0–1 decision variable y_e which is 1 when edge e belongs to the phylogenetic tree T, and also with a continuous decision variable w_e which represents the non-negative weight of edge e. In addition, for each $i \in V_{\text{ex}}$ and $j \in V$, we consider u_{ij} the unknown potential representing the length from i to j $(i \neq j)$. Then, we can formulate MEP as follows:

Model 4.

$$\min z = \sum_{e \in E} w_e \tag{39}$$

$$s.t.\ y(E(S)) \leq |S| - 1 \quad \forall S \subset N \tag{40}$$

$$y(E(V)) = 2n - 3 \tag{41}$$

$$y(\delta(i)) = 3 \quad \forall i \in V_{\text{in}} \tag{42}$$

$$u_{ij} \geq d_{ij} \quad \forall i,j \in V_{\text{ex}}, i \neq j \tag{43}$$

$$u_{ik} - u_{il} \leq w_{(k,l)} + \hat{d}(1 - y_{(k,l)}) \quad \forall i \in V_{\text{ex}}, \forall k \in V_{\text{ex}}, \forall l \in V_{\text{in}} \tag{44}$$

$$y_e \in \{0, 1\} \quad \forall e \in E \tag{45}$$

$$w_e \geq 0 \quad \forall e \in E \tag{46}$$

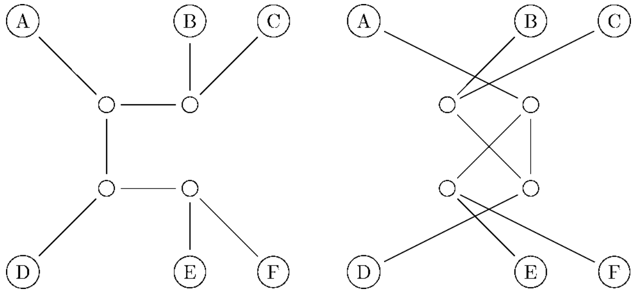

FIG. 7. An example of two complementary phylogenetic trees, see text for more details.

where the big-M value is $\hat{d} = \max\{d_{ij}\}$. Constraints (44) ensure that the length increment from k to l is not larger than $w_{(k,l)}$, while constraints (43) force the length of a path from i to j to be at least d_{ij}.

The motivation for the name of Model 4 is based on the fact that one can get it by considering a directed graph and by introducing a multi-commodity flow formulation. Indeed, in addition to the y and w variables, one may consider a flow variable f_{ij}^{k} representing a flow that i sends to k from node j if (i,j) is an arc, $i \in V$ and $k \in V_{\text{ex}}$. For a given k, this flow should cumulate the weight of edge $\{i,j\}$, and to this end one needs to impose that the flow going out from j must be at least the flow going out from i plus $w_{(i,j)}$. Then Model 4 is obtained by simply replacing the flow going out from i by u_{ik}.

A disadvantage of this model is the existence of many equivalent solutions leading to the same PT structure. Indeed, each external vertex is associated to a different taxon, but the internal vertices are identical elements. This implies that a PT structure is associated to many solutions, each one determined by a different permutation of the internal vertices. To avoid this kind of symmetry drawback Model 4 may be strengthened with variable fixings like, for example, $y_{A,1} = y_{1,2} = y_{2,3} = 1$ where 1, 2 and 3 represent three fixed vertices in V_{in}. However, as it will be pointed out in Section 4, the lower bound of the strengthened model is still weak. The next section introduces another approach to overcome the symmetry drawback.

2.4. *Fixed Tree Model*

The model proposed in this section is based on the following observation: the number of nonisomorphic unlabeled PTs grows exponentially with the number n of leaves [20]. However, such a number is still relatively small for $20 < n < 30$. Therefore, as a possible approach, we can decompose the MEP: we enumerate the nonisomorphic trees and, for each of them, we determine the optimal assignment of taxa to the tree leaves. However, note that the taxa assignment problem is probably still $\mathcal{NP}$-hard and the procedure that generates all the nonisomorphic trees is not trivial to be implemented.

In the following, we consider that the structure (and hence the edge-path incidence matrix $\mathbf{X}$) of the desired phylogenetic tree $T = (V, \mathcal{E})$ is given. We must determine the mapping of the taxa in Γ to the tree leaves. Then, for each taxon $i \in \Gamma$, leaf $r \in V_{\text{ex}}$, and edge $e \in \mathcal{E}$, we introduce the following decision variables: y_{ir} binary variable equal to 1 if taxon i is assigned to leaf r; w_e non-negative weight of e. The subproblem of interest becomes:

Model 5.

$$\min z = \sum_{e \in \mathcal{E}} w_e \tag{47}$$

$$s.t. \sum_{e \in \mathcal{E}} x_{rs,e} w_e \geq d_{ij}(y_{ir} + y_{is} + y_{js} + y_{jr} - 1)$$
$$\forall\, r, s \in V_{\text{ex}} : r < s,\ \forall\, i,j \in \Gamma : i < j \tag{48}$$

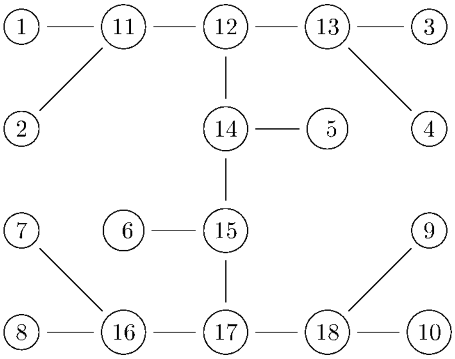

FIG. 8. An example of an unlabelled phylogenetic tree with ten taxa.

$$\sum_{i} y_{ir} = 1, \quad \forall r \in V_{\text{ex}} \tag{49}$$

$$\sum_{r} y_{ir} = 1, \quad \forall i \in \Gamma \tag{50}$$

$$w_e \geq 0 \quad \forall e \in \mathcal{E} \tag{51}$$

$$y_{ir} \in \{0, 1\} \quad \forall i \in \Gamma, \forall r \in V_{\text{ex}}. \tag{52}$$

Constraints (48) impose that the weight of the path between leaves r and s on the tree T is not less than d_{ij}, when taxa i and j are assigned to such vertices. Constraints (49), (50) impose that each taxon is assigned to a single leaf and vice versa.

A standard branch-and-bound algorithm performs very poorly in solving Model 5 due to poor linear-programming relaxations. In fact, by removing the integrality condition on variables y_{ir}, the right-hand side of constraints (48) usually assumes non positive values. In addition, Model 5 does not exploit any topological information (such as the fact that T is a PT) whereas it presents many equivalent solutions due to the usual presence of symmetries in the tree structure.

Much better performances are obtained when constraints are added to prevent the existence of the equivalent solutions. As an example, if we consider the tree in Figure 8 and we assume that the set Γ of taxa is ordered, we can impose that: (i) the first taxon can be assigned only to leaf 1 or to leaf 5; (ii) the taxon assigned to leaf 1 must precede the taxon assigned to leaf 2 and repeat the same argument for leaves 3 and 4, 7 and 8, and 9 and 10; (iii) the taxon assigned to leaf 1 must precede the taxon assigned to leaf 3 and repeat the same argument for leaves 7 and 9. In particular, if we assume that the first taxon is assigned to leaf 1 (i.e., $y_{A1} = 1$) we obtain:

Model 6.

$$\min z = \sum_{e \in \mathcal{E}} w_e \tag{53}$$

$$s.t. \sum_{e \in \mathcal{E}} x_{1s,e} w_e \geq d_{Aj} y_{js} \quad \forall s \in V_{\text{ex}} \setminus \{1\}, \forall j \in \Gamma \setminus \{A\} \tag{54}$$

$$\sum_{e\in\mathcal{E}} x_{rs,e}w_e \geq d_{ij}(y_{jr}+y_{is}+y_{js}+y_{ir}-1)$$

$$\forall r,s \in V_{\text{ex}} : r < s, \forall i,j \in \Gamma \quad (55)$$

$$\sum_i y_{ir} = 1 \quad \forall r \in V_{\text{ex}} \quad (56)$$

$$\sum_r y_{ir} = 1 \quad \forall i \in \Gamma \setminus \{A\} \quad (57)$$

$$y_{A1} = 1 \quad (58)$$

$$y_{i4} \leq \sum_{j=B}^{i-1} y_{j3} \quad \forall i \in \Gamma \setminus \{A,B\} \quad (59)$$

$$y_{i10} \leq \sum_{j=B}^{i-1} y_{j9} \quad \forall i \in \Gamma \setminus \{A,B\} \quad (60)$$

$$y_{i8} \leq \sum_{j=B}^{i-1} y_{j7} \quad \forall i \in \Gamma \setminus \{A,B\} \quad (61)$$

$$y_{i9} \leq \sum_{j=B}^{i-1} y_{j7} \quad \forall i \in \Gamma \setminus \{A,B\} \quad (62)$$

$$w_e \geq 0 \quad \forall e \in \mathcal{E} \quad (63)$$

$$y_{ir} \in \{0,1\} \quad \forall i \in \Gamma, \forall r \in V_{\text{ex}}. \quad (64)$$

In solving Model 6, we do not obtain significant gains from a computational time point of view if we impose that constraints (56) and (57) define SOS sets. Then we can add some cuts to (53)–(64); in particular, a possible family of cuts is the following: for each leaf r the sum of the weights of the paths rooted in r is not smaller than $\sum_{j\neq i} d_{ij}$ if i is assigned to r, i.e.,

$$\sum_{s\in V_{\text{ex}}\setminus\{r\}} \sum_{e\in\mathcal{E}} x_{rs,e}w_e \geq \sum_{i\in\Gamma}\left(y_{ir}\sum_{j\neq i} d_{ij}\right) \quad \forall r \in \mathcal{V}_{\text{ex}}. \quad (65)$$

We can also dynamically add some cuts on the basis of the observation that the sum of each subset of k left-hand side of constraints (54) and (55) must be at least equal to the sum of the first k shortest distances d_{ij} between taxa. In particular for any subset $\mathcal{F} \subseteq \{(i,j) : i,j \in \Gamma,\ i < j\}$ we have:

$$\sum_{(r,s)\in\mathcal{F}} \sum_{e\in\mathcal{E}} x_{rs,e}w_e \geq \sum_{t=1}^{|\mathcal{F}|} d^t, \quad (66)$$

where d^t is the t-th shortest distance between two taxa. Constraints (66) are exponential in number, but the separation algorithm is trivial. Given any fractional solution, order constraints (54) and (55) on the basis of values assumed by their right-hand sides. Then, for any k, consider the first k constraints (54) and (55) and check whether the sum of their left-hand sides is smaller than $\sum_{t=1}^{k} d^t$. If this condition holds true, the pairs of taxa associated with such constraints define a set $\mathcal{F}$. Both cuts (65) and (66) increase significantly the value of the linear relaxation of Model 6. However, they do not reduce the overall time required by the branch-and-bound (branch-and-cut) algorithm to find the optimal solution.

3. LOWER BOUNDS

In this section we introduce some lower bounds for MEP based on the following nonlinear formulation:

Model 7.

$$\min z = \sum_{e\in\mathcal{E}} w_e \quad (67)$$

$$w_i + w_j + \sum_{e\in\mathcal{E}_i} w_e x_{ij,e} \geq d_{ij} \quad \forall i,j \in V_{\text{ex}} : i < j \quad (68)$$

$$x_{ij,e} \leq x_{Ai,e} + x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \quad (69)$$

$$x_{ij,e} \leq 2 - x_{Ai,e} - x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \quad (70)$$

$$x_{ij,e} \geq x_{Ai,e} - x_{Aj,e} \quad \forall e \in \mathcal{E},\ \forall\, i,j \in V_{\text{ex}} : i < j \quad (71)$$

$$x_{ij,e} \geq -x_{Ai,e} + x_{Aj,e} \quad \forall e \in \mathcal{E}, \forall i,j \in V_{\text{ex}} : i < j \quad (72)$$

$$x_{Ai,f} + x_{Aj,f} + x_{Ai,e} - x_{Aj,e} \leq 2 \quad \forall e \in \mathcal{E}_{\text{in}}, \forall i,j \in V_{\text{ex}} : i<j \quad (73)$$

$$x_{Ai,f} + x_{Aj,f} - x_{Ai,e} + x_{Aj,e} \leq 2 \quad \forall e \in \mathcal{E}_{\text{in}}, \forall i,j \in V_{\text{ex}} : i < j \quad (74)$$

$$\sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,e} \geq 2 \quad \forall e \in \mathcal{E}_{\text{in}} \quad (75)$$

$$\sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,(2n-3)} = 2 \quad (76)$$

$$\sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,(n+1)} \leq n-2 \quad (77)$$

$$\sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,s} \leq \sum_{i\in V_{\text{ex}}\setminus\{A\}} x_{Ai,r} - 1 + (n-1)(2 - x_{Aj,r} - x_{Aj,s})$$

$$\forall r,s \in \mathcal{E}_{\text{in}} : r < s, \forall j \in V_{\text{ex}} \setminus \{A\} \quad (78)$$

$$w_e \geq 0 \quad \forall e \in \mathcal{E} \quad (79)$$

$$x_{ij,e} \in \{0,1\} \quad \forall i,j \in V_{\text{ex}} : i < j, \forall e \in \mathcal{E}. \quad (80)$$

Let z^* be the optimal solution of the above program. The following inequality holds:

$$z^* \leq \frac{n}{2}\hat{d} \quad (81)$$

where $\hat{d} = \max\{d_{ij}\}$. It is easy to verify that the solution $w_e = \frac{\hat{d}}{2}$, for all $e \in \mathcal{E}_{\text{ex}}$ and $w_e = 0$, for $e \in \mathcal{E}_{\text{in}}$, is feasible whichever EPT matrix $\mathbf{X}$ is chosen. Indeed, all the paths between two leaves have weight equal to $\hat{d}$ in a tree with such weights on the external edges. Later, in this section, we prove that the above bound is tight.

3.1. *Bounds by Surrogate Relaxation of Model 7*

We obtain a first set of lower bounds by exploiting the surrogate relaxation of some constraints (68) in Model 7.

A first trivial lower bound is $z_E := \hat{d}$. We obtain this bound by eliminating constraints (68) except one for which the right-hand side is $d_{ij} = \hat{d}$. There exist some instances for which such a lower bound is tight, e.g., when $d_{Aj} = \hat{d}$ for all $j \in \Gamma \setminus \{A\}$ and $d_{ij} = 0$ otherwise. However, in general, $\hat{d}$ is a poor lower bound.

Now, assume that the surrogate multipliers for constraints (68) are all equal to 1. Then, in Model 7, we substitute the relaxed constraints with the single condition:

$$\sum_{e \in \mathcal{E}} u_e^T w_e \geq \Delta \tag{82}$$

where we define $\Delta = \sum_{i \in V_{\text{ex}}} \sum_{j \in V_{\text{ex}}: i<j} d_{ij}$ and $u_e^T = \sum_{i \in V_{\text{ex}}} \sum_{j \in V_{\text{ex}}: i<j} x_{ij,e}$ if $e \in \mathcal{E}_{\text{in}}$, $(n-1)$ otherwise, i.e., if $e \in \mathcal{E}_{\text{ex}}$. Observe that u_e^T represents the number of paths including edge e on the tree T. Then the maximum value that u_e^T can assume is $\frac{n^2}{4}$ if n is an even number, $\frac{n^2-1}{4}$ otherwise. This situation occurs when the edge e defines a cut that partitions the tree vertices into sets of cardinality $\frac{n}{2}$ if n is even, $\frac{n+1}{2}$ and $\frac{n-1}{2}$ otherwise.

Assume hereafter n even for the sake of simplicity and note that $\frac{n^2}{4} \geq n-1$ for $n \geq 2$. The proposed surrogate relaxation of Model 7 has optimal solution, and hence lower bound of the original problem

$$z_{RS(1)} = \frac{4\Delta}{n^2}. \tag{83}$$

Such a solution is associated with a tree with all edge weights equal to 0, except for the edge that partitions the vertices in two sets with the same cardinality. Note that

$$z_{\text{RS}(1)\Delta} = \frac{4\Delta}{n^2} \leq \frac{4}{n^2}\frac{n(n-1)}{2}\hat{d} = 2\left(1 - \frac{1}{n}\right)\hat{d} \tag{84}$$

where the inequality becomes equality when $d_{ij} = \hat{d}$ for all $i,j \in \Gamma$.

Unfortunately, we are not able to determine analytically other results for different surrogate relaxations of Model 7. Consider as an example, the possibility of eliminating constraints (68) except the ones associated with paths rooted in the same leaf, say i. In this case the optimal solution of the relaxed problem is a tree where the weight of the external edge with an extreme in i is equal to $\max_j\{d_{ij}\}$ and all the other edge weights are equal to 0.

3.2. Combinatorial Bounds

Throughout this section we make reference to an auxiliary graph defined as follows. Let $H = (\Gamma, \mathcal{A})$ be a graph, with $\mathcal{A} = \{(i,j) : i,j \in \Gamma,\ i<j\}$ and where the weight of generic edge $(i,j) \in \mathcal{A}$ is equal to the distance d_{ij}.

3.2.1. Path Packing Bound. Let the set Γ include n taxa, with n an even number. Consider a possible phylogenetic tree T and denote P_{ij} as the set of the edges of the path between the leaves i and j. We define a set Λ^T as a packing of paths when $P_{ij} \cap P_{rs} = \emptyset$ for a pair of paths P_{ij} and P_{rs} in Λ^T. In particular, observe that no pair of paths in a packing have an extreme vertex (a leaf) in common.

For any packing Λ^T, it holds trivially that

$$\sum_{e \in \mathcal{E}} w_e \geq \sum_{P_{ij} \in \Lambda^T} \sum_{e \in P_{ij}} w_e. \tag{85}$$

Then, z^* is always greater than the sum of lengths of the paths in a packing.

In the following Theorem 2 we prove the existence of a packing with cardinality equal to $\frac{n}{2}$. We also observe that the sum of lengths of the paths in such a packing must be greater than or equal to the minimum 1-weighted matching on the graph $H = (\Gamma, \mathcal{A})$. Then, we obtain the following lower bound for z^*:

$$z^* \geq z_C = d^{(1)} + d^{(2)} + \ldots + d^{(n/2)} \tag{86}$$

where $d^{(1)}, d^{(2)}, \ldots, d^{n/2}$ are weight of edges of the optimal matching, being $d^{(1)} \leq d^{(2)} \leq \ldots \leq d^{n/2}$. Note that if $d_{ij} = \hat{d}$ for all $i,j \in V_{\text{ex}}$, (86) becomes $z^* \geq \frac{n}{2}\hat{d}$ and this proves that inequality (81) is tight.

The following definitions are necessary for Theorem 2. Consider a generic tree T, define a path P_{ij} as *simple* if it does not include an internal edge. Define an internal edge as *appended* if after the removal of some other edge of T one of its extreme has degree 1. Finally, let us call *edge shrinking* the operation that substitutes two generic adjacent edges $e_1 = (i,j)$ and $e_2 = (j,k)$, with the single edge $e = (i,k)$, and removes vertex j with degree 2. If e_1 and e_2 are internal edges, define the new edge e as internal, otherwise define e as external.

Theorem 2. *If n is an even number and T is a PT, there exists a packing Λ^T which includes $\frac{n}{2}$ paths.*

Proof. As T is a PT, its internal vertices have degree three. Then at least one simple path on T exists, say P_{ij}. Then define a new tree $\hat{T}$ obtained from T as follows: remove the edges in P_{ij}; recursively remove all the appended edges and all the vertices that remain not connected; finally recursively shrink all the edges with an extreme of degree equal to 2. The new tree $\hat{T}$ is still a connected PT and has 2 leaves less than T. Redefine $\hat{T}$ as T and iterate the above sequence of operations. After $\frac{n}{2}$ iteration obtain an empty tree and stop. At each iteration, the procedure identifies a simple path corresponding to a path (not necessarily simple) on the original tree. The identified $\frac{n}{2}$ paths define a packing by construction. ■

The above theorem turns out to be useful even for determining another lower bound. Observe that, if $d_{ij} > 0$, $\forall\ i,j$ there are at least $\frac{n}{2}$ edges with $w_e > 0$ as there is a packing with $\frac{n}{2}$ paths. Consider again condition (82) and remember that u_e^T represents the number of paths including edge e on the tree T. Then a lower bound for z^* is given by the optimal solution of the following formulation:

Model 8.

$$\min_{T, w_e} z_D = \sum_{e \in \mathcal{E}} w_e$$

$$\sum_{e \in \mathcal{E}} u_e^T w_e \ge \Delta$$

$$\{w_e\} \in \Omega$$

where Ω is the set of vectors $\mathbf{w} = \{w_e\}$ with at least $\frac{n}{2}$ components strictly greater than 0 defining $\frac{n}{2}$ paths of length greater than or equal to $d^{(1)}, d^{(2)}, \ldots, d^{(\frac{n}{2})}$, respectively. Given a generic phylogenetic tree T, let us indicate by $u^{(k)}$ the k-th greatest value u_e^T and $w^{(k)}$ the associated weight. The following theorem hold:

Theorem 3. *For any given phylogenetic tree T, there exists an optimal solution to Model 8 such that*

$$w^{(1)} = \tilde{d} \ge w^{(2)} = d^{(\frac{n}{2}-1)} \ge w^{(3)} = d^{(\frac{n}{2}-2)} \ge \ldots \ge w^{(\frac{n}{2})} = d^{(1)} > w^{(\frac{n}{2}+1)} = \cdots = w^{(2n-3)} = 0. \quad (87)$$

where

$$\tilde{d} = \begin{cases} d^{(\frac{n}{2})} & \text{if } \sum_{k=1}^{\frac{n}{2}} u^{(k)} d^{(\frac{n}{2}-k+1)} \ge \Delta \\ \frac{\Delta - \sum_{k=1}^{\frac{n}{2}} u^{(k)} d^{(\frac{n}{2}-k+1)}}{u^{(1)}} & \text{otherwise.} \end{cases} \quad (88)$$

Proof. Observe that any optimal solution of Model 8 has $z_D \ge d^{(1)} + d^{(2)} + \ldots + d^{(\frac{n}{2})}$ and that the solution proposed in the problem statement is feasible. Then, such a solution is trivially optimal if $\tilde{d} = d^{(\frac{n}{2})}$, since the associated value of z_D is equal to $d^{(1)} + d^{(2)} + \ldots + d^{(\frac{n}{2})}$. On the other hand, observe that $\sum_{k=1}^{\frac{n}{2}} u^{(k)} d^{(\frac{n}{2}-k+1)}$ is the maximum value that we can have with a solution with $\frac{n}{2}$ components respectively equal to $d^{(1)}, d^{(2)}, \ldots, d^{(\frac{n}{2})}$ and all the remaining ones equal to 0. We obtain such a value when $w^{(1)} = d^{(\frac{n}{2})}, w^{(2)} = d^{(\frac{n}{2}-1)}, \ldots, w^{(\frac{n}{2})} = d^{(1)}$. If such a solution is not feasible we have to increase at least one component. As $u^{(1)}$ is greater than any other term u_e^T and as we want to minimize the sum of the edge weights, the best choice is to increase the weight of the associated edge to $w^{(1)} = \frac{\Delta - \sum_{k=1}^{\frac{n}{2}} u^{(k)} d^{(\frac{n}{2}-k+1)}}{u^{(1)}}$ in order to obtain $\sum_{k=1}^{\frac{n}{2}} u^{(k)} w^{(k)} = \Delta$. ■

The above theorem holds for any fixed T. Now, we must find some upper bounds on the values of $u^{(k)}$ that hold for any phylogenetic tree T. This situation is faced in the next Theorem 4.

Let us first observe that given a generic tree $T = (V, \mathcal{E})$, and a generic cut-edge e defining a partition $S - V_{\text{ex}} \setminus S$ on leaves of T, the number of paths using e on T is given by $|S||V_{\text{ex}} \setminus S|$. Let us denote as $u^T = \{u_e^T : e \in \mathcal{E}\}$ the vector with components u_e^T ordered such that $u_e^T \le u_f^T$ if $e < f$. Finally, let us define a phylogenetic tree C as a *comb* with $(2n-3)$ edges (for short, $comb_{(2n-3)}$) if the induced subtree on the internal vertices forms a path graph with topological length $(n-3)$ (see, Fig. 9).

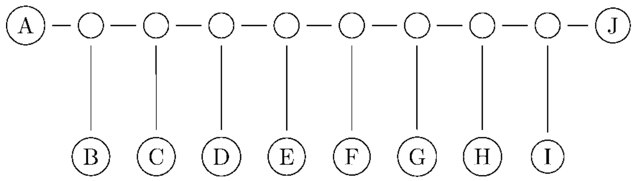

FIG. 9. An example of comb with ten taxa.

Theorem 4. *Consider PTs with n leaves. Let C be a comb, then we have $u^T \le u^C$ for any tree T.*

Proof. Initially note that the values of vector u^T, respectively u^C, do not depend on the leaf labels. Then all the trees T, respectively C, with the same topological structure share the same vector u^T, respectively u^C. Let $C = (V, \mathcal{E})$ be a comb with n leaves. The total number of vertices in C has cardinality $|V| = (2n-2)$ [25], therefore $|V|$ is always an even number. Then, by Jordan's Theorem [1], if T is a tree with $m = 2k$ vertices then there exists (i) a unique vertex, called the *centroid*, such that all three incident subtrees are composed of less than k vertices, or (ii) an unique edge, called the *bicentroid*, such that the incident two subtrees are composed exactly of k vertices. In other words, both centroid and bicentroid are symmetric points for the tree C. Hereafter, let us use the convention of enumerating the internal edges starting from the centroid (bicentroid). If C is characterized by a centroid then such a symmetric point divides the tree in three parts: one subtree consisting of one leaf, and two subtrees consisting of $\left(\frac{n-1}{2}\right)$ leaves. Moreover, the two internal edges adjacent to the centroid, by splitting C in two parts, will be crossed by $\left(\frac{n-1}{2}\right)\left(\frac{n-1}{2}+1\right)$ paths. The internal edges immediately following the internal edges adjacent to the centroid will be crossed by $\left(\frac{n-1}{2}-1\right)\left(\frac{n-1}{2}+2\right)$ paths. In general, by recursively applying such enumeration the k-th internal edges far away from the centroid are crossed by $\left(\frac{n-1}{2}-k\right)\left(\frac{n-1}{2}+1+k\right)$ paths, $k = 0, \ldots, \left(\frac{n-1}{2}-2\right)$. Analogously, when C is characterized by a bicentroid the k-th internal edges far away from the bicentroid are crossed by $\left(\frac{n}{2}-k\right)\left(\frac{n}{2}+k\right)$ paths, $k = 0, \ldots, \left(\frac{n}{2}-2\right)$. Therefore the vector u^C is:

$$[u_0^C, u_0^C, u_1^C, u_1^C, \ldots, u_k^C, u_k^C] \quad (89)$$

where the generic component u_k^C is equal to $\left(\frac{n-1}{2}-k\right)\left(\frac{n-1}{2}+1+k\right)$ if C is characterized by a centroid, and $\left(\frac{n}{2}-k\right)\left(\frac{n}{2}+k\right)$ if C is characterized by a bicentroid. Let us call the odd components of the vector u^C *left*-components, and the even components *right*-components. If Theorem 4 is false, then there exist at least one tree T different from C and an associated vector u^T such that $u^T > u^C$ at least for one component, i.e.,

$$\exists\, j : 0 \le j \le \max\left\{\frac{n-1}{2} - 2, \frac{n}{2} - 2\right\} \text{ and } u_j^T > u_j^C. \quad (90)$$

Assume, without loss of generality, that u_j^T is a left-component. If T is different from C, than its diameter is necessarily smaller than the corresponding one of C, at least for one internal edge; in other words there must exist an internal edge $\hat{j}$ on the left (as the right part did not change) of the centroid (bicentroid) that is no longer on the longest path from the centroid to the far leaf. As the tree T is connected, internal edge $\hat{j}$ necessarily links two leaves, and therefore, $u_{\hat{j}}^T = 2(n-2)$, i.e., the number of paths crossing it is the same of the internal edges linking the leaves away from the centroid (bicentroid), so $u_{\hat{j}}^T \leq u_k^T$, $0 \leq k < \left(\frac{n-1}{2} - 2\right) \left(0 \leq k < \left(\frac{n}{2} - 2\right)\right)$. Now as all the right and left components of T are the same for C except j then $\hat{j} = j$, and therefore we have that $u_j^C \leq u_j^T = u_{\hat{j}}^T \leq u_k^T = u_k^C$, $0 \leq k < \left(\frac{n-1}{2} - 2\right) \left(0 \leq k < \left(\frac{n}{2} - 2\right)\right)$ that leads to a contradiction of (90). ■

An immediate consequence of the above theorems is the following lower bound on z^*:

$$z^* \geq z_{DC} = \tilde{d} + d^{(2)} + d^{(3)} + \ldots + d^{(\frac{n}{2})} \qquad (91)$$

where $\tilde{d}$ is defined as in (88), when the values $u^{(k)}$ are the ones defined on a comb with n leaves. We obtain analogous results when the number n of taxa in the set Γ is odd. The main difference is that the bound is $z^* \geq z_{DC} = \tilde{d} + d^{(2)} + d^{(3)} + \ldots + d^{(\frac{n-1}{2})}$.

3.2.2. TSP Bound The following bound derives from the observations in [16]. Consider a generic phylogenetic tree T. We can always define a tour along its edges so that we visit each leaf once and each edge twice. Let i_k, $k = 1, 2, \ldots, n$ be the generic leaf and $\sigma = i_1, i_2, \ldots, i_n$ be the sequence that defines the order in which we visit the leaves in the tour. For each pair of consecutive leaves (i_k, i_{k+1}) in σ, it holds trivially that the length of the path on T between i_k and i_{k+1} is greater than or equal to $d_{i_k i_{k+1}}$. We note also that σ induces the Hamiltonian cycle $(i_1, i_2), (i_2, i_3), \ldots, (i_n, i_1)$ on the graph H. Consequently, we obtain the following bound on z^*:

$$z^* \geq z_{tsp} = \frac{l_{tsp}}{2} \qquad (92)$$

where l_{tsp} is the length of the shortest Hamiltonian cycle on H. Trivially, (92) is a stronger bound than (86) as $d^{(1)} + d^{(2)} + \ldots + d^{(\frac{n}{2})} \leq \frac{l_{tsp}}{2}$. The bound (92) is tight when all the distances d_{ij} are equal.

3.2.3. Triangular bound. Consider a generic phylogenetic tree T and any set of three leaves $\{i_1, i_2, i_3\}$. Define as j the only vertex common to the three paths respectively between i_1 and i_2, i_2 and i_3, i_3 and i_1. In addition, defines w_{ij} as the sum of the edge weights belonging to the path from vertex i to j. For any phylogenetic tree T, it holds that:

$$\sum_{e \in \mathcal{E}} w_e \geq w_{i_1 j} + w_{i_2 j} + w_{i_3 j} \geq w(i_1, i_2, i_3)$$

$$= \frac{d_{i_1 i_2} + d_{i_2 i_3} + d_{i_3 i_1}}{2} = \left\{ \begin{array}{ll} \min & w_{i_1 j} + w_{i_2 j} + w_{i_3 j} \\ & w_{i_1 j} + w_{i_2 j} \geq d_{i_1 i_2} \\ & w_{i_1 j} + w_{i_3 j} \geq d_{i_1 i_3} \\ & w_{i_2 j} + w_{i_3 j} \geq d_{i_2 i_3} \\ & w_{i_1 j}, w_{i_2 j}, w_{i_3 j} \geq 0 \end{array} \right\} \qquad (93)$$

Then a bound for z^* is

$$z^* \geq z_{3B} \qquad (94)$$

where $z_{3B} = \max_{\{i_1, i_2, i_3\}} w(i_1, i_2, i_3)$.

We can also see that the above bound is stating that the length of the PT on Γ must be greater than the length of the PT on any $S \subseteq \Gamma$, with cardinality of S equal to 3. From this perspective, we can generalize trivially the above bound, for any class of subsets of Γ of fixed cardinality $k < n$. In particular, we have $z^* \geq z_{kB}$ with $z_{kB} = \max_{S:|S|=k} w(S)$ where $w(S)$ is the length of the PT on the subset S. Unfortunately, the complexity of computing z_{kB} increases exponentially with k. Additionally, it is very likely that this bound is useless in practice due to its simplicity.

3.2.4. A Relaxed Problem: The Degree-Constrained Spanning Tree Problem. A way of improving the bound from a linear-programming relaxation of a mathematical model is by finding new valid inequalities for binary trees. In this line, we find of interest studying the convex hull of all binary trees.

A first approach would consider an undirected graph $G = (V, E)$ where V is partitioned into V_{in} and V_{ex} with $|V_{\text{in}}| = |V_{\text{ex}}| - 2$. The edge set E is also partitioned in E_{in} (all edges connecting vertices in V_{in}) and E_{ex} (all edges with one vertex in V_{in} and the other vertex in V_{ex}). The approach analyzes the binary trees having vertices in V_{in} with degree 3 and vertices in V_{ex} with degree 1. A mathematical description for this set of solutions may be written by introducing a 0-1 variable x_e for each $e \in E$. Then:

$$\begin{array}{rl} x(\delta(i)) = 3 & \text{for all } i \in V_{\text{in}} \\ x(\delta(i)) = 1 & \text{for all } i \in V_{\text{ex}} \\ x(E(S)) \leq |S| - 1 & \text{for all } S \subset V \\ x(E) = |V| - 1 & \\ x_e \in \{0, 1\} & \text{for all } e \in E. \end{array}$$

In order to obtain valid inequalities for the convex hull of these solutions it is worth analyzing the projection into the variables x_e with $e \in E_{\text{in}}$. The projected solutions correspond

TABLE 1. Experimental results relative to instances containing 8 taxa.

Instance	Optimum	Nodes	Time (s.)	LP-root
Primates12/898	0.808395	12000	62.781	0.36723
RbcL55/1314	0.758649	23963	110.703	0.313898
Rana64/1976	0.252526	19983	92.125	0.119746
M17/2550	0.317527	23451	111.734	0.1323
M43/2086	0.559108	30021	134.562	0.17506
M18/8128	0.288401	20141	97.25	0.160755
M82/2062	0.475001	17880	88.14	0.173246
M62/3768	0.694578	24027	114.063	0.31458
SeedPlant25/19784	0.444784	22493	106.031	0.190279

to trees in V_{in} where all vertices have degree 1, 2, or 3. A mathematical model for these tree structures is

$$x(\delta(i)) \leq 3 \quad \text{for all } i \in V_{\text{in}} \tag{95}$$

$$x(E(S)) \leq |S| - 1 \quad \text{for all } S \subset V_{\text{in}} \tag{96}$$

$$x(E_{\text{in}}) = |V_{\text{in}}| - 1 \tag{97}$$

$$x_e \in \{0, 1\} \quad \text{for all } e \in E_{\text{in}}. \tag{98}$$

When the value 3 in (95) is replaced by a positive integer parameter b_i, constrains (95)–(98) define the solution set of the so-called *degree-constrained spanning tree problem* in the graph $G_{\text{in}} = (V_{\text{in}}, E_{\text{in}})$. This problem has been studied by several authors in the literature (see, e.g., Salles da Cunha and Lucena [7]).

As observed in [7], it is possible to derive valid inequalities for the solutions of (95)–(98) in a similar way as done with the 2-matching, the comb and the clique-tree inequalities for the classical Travelling Salesman Problem. The procedure is named {0, 1/2} Chvátal-Gomory derivation and it consists of adding some inequalities in (95) and (96), dividing by 2, using the linear relaxation of (98) (i.e., $0 \leq x_e \leq 1$) and rounding down the right-hand side. Through this procedure one can get the *2-matching constraints* when $b_i = 3$:

$$x(E(H)) + x(T) \leq |H| + \left\lfloor \frac{|H| + |T|}{2} \right\rfloor \tag{99}$$

for each vertex subset $H \subset V_{\text{in}}$ (named *handle*), and each subset $T \subset E_{\text{in}}$ (named *teeth*) of edges with one vertex in H and the other vertex outside H. Clearly, they are useful only when $|H| + |T|$ is an odd number. Another example of valid inequalities for the degree-constrained spanning tree problem with $b_i = 3$ are the following *comb inequalities*:

$$x(E(H)) + \sum_{i=1}^{k} x(E(T_i)) \leq |H| + \sum_{i=1}^{k} (|T_i| - 1) + \left\lfloor \frac{|H| - k}{2} \right\rfloor \tag{100}$$

where $H, T_1, \ldots, T_k \subset V_{\text{in}}$ such that $H \cap T_i \neq \emptyset$, $T_i \setminus H \neq \emptyset$ and $T_i \cap T_j = \emptyset$ for all $i, j = 1, \ldots, k$ $(i < j)$. To be useful in a relaxation (at least) $|H| - k$ must be odd.

4. COMPUTATIONAL RESULTS

We have considered various real aligned DNA datasets for the numerical experiments: "Primates12/898," a dataset of 12 sequences, 898 characters each from primates mtDNA; "RbcL55/1314," a dataset of 55 sequences, 1314 characters each of the rbcL gene; "Rana64/1976," a dataset of mtDNA 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 2,550 characters each from insects, 43 sequences of 2,086 characters each from mammals, 18 sequences of 8,128 characters each from cetacea, 82 sequences of 2,062 characters each from fungi, and 62 sequences of 3,768 characters each form hyracoidae; finally, "SeedPlant25/19784," a dataset of 25 sequences of 19784 characters each from pinoles. From each dataset we have extracted the first 8, 10 and 12 taxa and built the associated 8 × 8, 10 × 10 and 12 × 12 distance matrices by using the Jukes-Cantor model of DNA sequence evolution in which all the gaps were treated as 'N' (see [13]). The datasets and the corresponding distance matrices can be provided by the authors upon request.

We have written the models in Section 2 using Xpress Mosel 2 and then solved using Xpress Optimizer v18.10.00 on a Intel Core 2 Duo 2 GHz PC with 2 GB of RAM. The running time was limited to a maximum of 3 hours. Only Model 2 was able to be solved to optimality on instances with 8 taxa. The reason for the negative behavior of Models 3 and 4 are due to the poor lower bound obtained from the linear programming relaxation of these models. No model was able to be optimally solved when the input was an instance with 10 or 12 taxa. Tables 1 and 2 show the results related to Model 2 when considering the first 8 and 10 taxa, respectively,

TABLE 2. Experimental results relative to instances containing 10 taxa.

Instance	Upper Bound	Gap	Nodes	Time (s.)	LP-root	Lower Bound
Primates12/898	0.956201	0.315992	598022	limit	0.36723	0.654049
RbcL55/1314	0.894156	0.40575	625847	limit	0.317028	0.531352
Rana64/1976	0.300153	0.460636	567069	limit	0.123747	0.191688
M17/2550	0.45882	0.549675	660760	limit	0.169231	0.247471
M43/2086	0.658711	0.41475	609101	limit	0.175061	0.296634
M18/8128	0.327488	0.446972	653624	limit	0.160755	0.191662
M82/2062	0.590379	0.367236	593089	limit	0.184847	0.326496
M62/3768	0.789958	0.361365	613470	limit	0.320779	0.499857
SeedPlant25/19784	0.52135	0.455467	584834	limit	0.190279	0.283892

TABLE 3. Relations between the lower bounds relative to instances containing 8 taxa.

Instance	Surrogate	Path packing	TSP	Triangular	Xor (Model 2)
Primates12/898	0.6610	0.350736	0.784060	0.282499	0.497178
RbcL55/1314	0.565016	0.496721	0.716420	0.167815	0.418909
Rana64/1976	0.215543	0.110754	0.243161	0.026148	0.151045
M17/2550	0.238140	0.204457	0.304904	0.121614	0.157379
M43/2086	0.315110	0.444207	0.542261	0.136412	0.216111
M18/8128	0.289359	0.129041	0.224816	0.147264	0.160755
M82/2062	0.311843	0.341281	0.448301	0.144258	0.224777
M62/3768	0.566244	0.436944	0.676495	0.290162	0.40562
SeedPlant25/19784	0.342502	0.286313	0.412442	0.125477	0.203073

of each benchmark dataset. Column with label "Nodes" displays the number of branch-and-bound nodes explored by the optimizer. Columns with labels "Upper Bound" and "Lower Bound" represent the best value when the optimizer was stopped. Column with label "Gap" is computed as the difference between the two bounds, divided by the upper bound. Observe that the best bound for 10 taxa cannot be smaller than the optimum value for 8 taxa. Therefore, for example, the effective error of the upper bound for "Primates12/898" is at most 0.154557, i.e., half of the gap indicated in Table 2.

To better illustrate this behavior let us consider the instance in Figure 1. When considering only the first eight taxa (from A to H), Model 2 was optimally solved in 51 s and Model 4 was optimally solved in 75 s. The optimal value of the linear-programming relaxation (at the root node) was 1.326 for both models, and the optimal solution value was 2.7565. Although the initial lower bound was the same, the effect of fixing variables required exploring less branch-and-bound nodes in Model 2 than in Model 4. When considering the instance with all taxa, Model 2 was able to find an optimal solution in about 12 h. This solution is the one represented in Figure 2 and its objective value is 3.58. The initial lower bound was 1.456 both when using Model 2 and when using Model 4. After 1 h running, the gap from Model 4 is two times the gap from Model 2, and in both cases the upper bound was 3.58. Figure 3 shows the feasible solution found when using Model 4 after 1 h. After 12 h, the lower bound from Model 4 was still 1.456. On the basis of our computational experiments, Model 2 outperforms in practice the other models presented in this article. A possible explanation for this behavior is the more accurate way of escaping from the symmetry drawback of Model 4 by using the EPT representation.

We have experienced a dependency of the performances of Model 2 on the entries of the distance matrix. More precisely, we have observed that Model 2 performs better when the similarity among the pairwise DNA sequences is higher, i.e., when the entries of the distance matrix tend to be similar. Tables 3 and 4 show the relations between the lower bounds described in Section 3 for instances containing 8 and 10 taxa, respectively. Beyond considering the surrogate, the path packing, the TSP and the triangular bounds, we have also considered the lower bound obtained by solving Model 2 without constraints (16)–(19), indicated in column with label "Xor (Model 2)". All the lower bounds, except the Xor bound, are computed in less than one second. The Xor bound requires at most 11 s and at most 1 h for instances of 8 and 10 taxa, respectively. The reason of this is due to the existence of big-M values in constraints (15). Finally, as general trend, the TSP bound outperforms the others which emphasizes the results found in Section 3.2.2.

5. CONCLUSION

Given *n* species, the problem of building a PT with minimum length is at the core of Phylogenetics. This optimization problem aims at building a spanning tree whose leaves are the observed species and whose internal vertices represent common ancestors. Each ancestor should be connected to three other vertices of the tree. In addition, weights should be assigned to the edges of the tree such that the length of the path connecting any two species should be not smaller than the estimated evolutionary distance between the same two species. The estimated evolutionary distances are input

TABLE 4. Relations between the lower bounds relative to instances containing 10 taxa.

Instance	Surrogate	Path packing	TSP	Triangular	Xor (Model 2)
Primates12/898	0.661014	0.423117	0.896109	0.295882	0.506888
RbcL55/1314	0.570650	0.454993	0.829118	0.061195	0.420474
Rana64/1976	0.222745	0.115827	0.286244	0.106148	0.198181
M17/2550	0.304616	0.263832	0.4398	0.169231	0.216111
M43/2086	0.315110	0.52713	0.635056	0.134765	0.160755
M18/8128	0.289359	0.145937	0.24461	0.110596	0.227894
M82/2062	0.332725	0.434189	0.539623	0.1031	0.408719
M62/3768	0.577402	0.487633	0.761811	0.320779	0.156715
SeedPlant25/19784	0.342502	0.314803	0.463182	0.188446	0.203073

data computed by technical procedures from sets of DNA or protein sequences.

The paper has presented a mathematical programming model based on the representation of the trees by edge-path incidence matrices. Another three models have also been proposed based on the standard representation of a tree by the set of edges of a complete graph. According to our experiments, the best approach is based on solving the first model. Still, this optimization problem is very challenging for exact methods since instances with 10 taxa or more are solved only heuristically. Different lower bounds have been also discussed in this work.

Acknowledgments

The authors thank Julien Guglielmini for the datasets provided, and Luis Gouveia for interesting discussions which improved this article.

REFERENCES

[1] R. Aringhieri, P. Hansen, and F. Malucelli, Chemical trees enumeration algorithms, 4OR 1 (2003), 67–83.

[2] M.O. Ball, T.L. Magnanti, C.L. Monma, and G.L. Nemhauser, Network models, Vol. 7: Handbooks in operations research and management science, Elsevier Science Publishing Company, Amsterdam, The Netherlands, 1995.

[3] W.A. Beyer, M. Stein, T. Smith, and S. Ulam, A molecular sequence metric and evolutionary trees, Math Biosci 19 (1974), 9–25.

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

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

[6] L.L. Cavalli-Sforza and A.W.F. Edwards, Phylogenetic analysis: Models and estimation procedures, Am J Human Genetics 19 (1967), 233–257.

[7] A. Salles da Cunha and A. Lucena, Lower and upper bounds for the degree-constrained minimum spanning tree problem, Networks 50 (2007), 55–66.

[8] W.H.E. Day, Computational complexity of inferring phylogenies from dissimilarity matrices, Bull Math Biol 49 (1987), 461–467.

[9] J. Felsenstein, Inferring phylogenies, Sinauer Associates, Sunderland, UK, 2004.

[10] W.M. Fitch and E. Margoliash, Construction of phylogenetic trees, Science 155 (1967), 279–284.

[11] M. Hasegawa, H. Kishino, and T. Yano, Evolutionary trees from dna sequences: a maximum likelihood approach, J Mol Evol 17 (1981), 368–376.

[12] M. Hasegawa, H. Kishino, and T. Yano, Dating the human-ape splitting by a molecular clock of mitochondrial dna, J Mol Evol 22 (1985), 160–174.

[13] T.H. Jukes and C.R. Cantor, "Evolution of protein molecules, Mammalian protein metabolism," H. N. Munro (Editor), Academic Press, New York, 1969, pp. 21–123.

[14] K.K. Kidd and L.A. Sgaramella-Zonta, Phylogenetic analysis: concepts and methods, Am J Human Genetics 23 (1971), 235–252.

[15] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nulceotide sequences, J Mol Evol 16 (1980), 111–120.

[16] C. Korostensky and G.H. Gonnet, Using the traveling salesman problem algorithms for evolutionary tree construction, Bioinformatics 16 (2000), 619–627.

[17] C. Lanave, G. Preparata, C. Saccone, and G. Serio, A new method for calculating evolutionary substitution rates, J Mol Evol 20 (1984), 86–93.

[18] R.K. Martin, Large scale linear and integer optimization: a unified approach, Kluwer Academic Publishers, Norwell, Massachusetts, 1999.

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

[20] E.M. Rains and N.J.A. Sloane, On cayley's enumeration of alkanes (or 4-valent trees), J Integer Sequences 2 (1999), 99.1.1.

[21] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina, The general stochastic model of nucleotide substitution, J Theoretical Biol 142 (1990), 485–501.

[22] A. Rzhetsky and M. Nei, A simple method for estimating and testing minimum evolution trees, Comput Appl Biosci 10 (1992), 409–412.

[23] A. Rzhetsky and M. Nei, Statistical properties of the ordinary least-squares, generalized least-squares, and minimum evolution methods of phylogenetic inference, J Mol Evol 35 (1992), 367–375.

[24] A. Rzhetsky and M. Nei, Theoretical foundations of the minimum evolution method of phylogenetic inference, Mol Biol Evol 10 (1993), 1073–1095.

[25] C. Semple and M. Steel, Phylogenetics, Oxford University Press, New York, NY, 2003.

[26] 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, Mol Phylogenetics Evol 8 (1997), 398–414.

[27] M.S. Waterman, T.F. Smith, M. Singh, and W.A. Beyer, Additive evolutionary trees, J Theoretical Biol 64 (1977), 199–213.