The Balanced Minimum Evolution Problem

Daniele Catanzaro* Martine Labbé◇ Raffaele Pesenti†
Juan-José Salazar-González‡

September 28, 2010

Abstract

A phylogeny is an unrooted binary tree that represents the evolutionary relationships of a set of n species. Phylogenies find applications in several scientific areas ranging from medical research, to drug discovery, to epidemiology, to systematics, and to population dynamics. In such applications the available information is usually restricted to the leaves of a phylogeny and is represented by molecular data extracted from the analyzed species, such as DNA, RNA, amino acid or codon fragments. On the contrary, the information about the phylogeny itself is generally missing and is determined by solving an optimization problem, called the Phylogeny Estimation Problem (PEP), whose versions depend on the criterion used to select a phylogeny from among plausible alternatives. In this article we investigate a recent version of the PEP, called the *Balanced Minimum Evolution Problem* (BMEP). We present a mixed integer linear programming model[1] to solve exactly instances of the BMEP and develop branching rules and families of valid inequalities to further strengthen the model. Our results give perspective on the mathematics of the BMEP and suggest new directions on the development of future efficient exact approaches to solution of the problem.

Keywords: network design, combinatorial optimization, lagrangian relaxation, computational biology, balanced minimum evolution, combinatorial inequalities, Kraft equality, Huffman coding.

1 Introduction

Molecular phylogenetics studies the hierarchical evolutionary relationships among species, or *taxa*, by means of molecular data such as DNA, RNA, amino acid 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* the intermediate ancestors, *edges* the estimated evolutionary relationships, and *edge weights* measures of the similarity between pairs of taxa (Catanzaro, 2009).

Phylogenies provide a fundamental information in analysis of many fine-scale genetic data, for this reason their use has become more and more frequent, and sometimes indispensable, in a multitude of research fields such as medical research, drug discovery, epidemiology, or population dynamics (Pachter and Sturmfels, 2007). For example, the use of molecular phylogenetics was of considerable assistance to predict the evolution of human influenza A (Bush et al., 1999), to understand the relationships between the virulence and the genetic evolution of HIV (Ross and Rodrigo, 2002; Ou et al., 1992), to identify emerging viruses as SARS (Marra et al., 2003), to recreate and investigate ancestral proteins (Chang and Donoghue, 2000), to design neuropeptides causing smooth muscle contraction (Bader et al., 2001), or to relate geographic patterns to macroevolutionary processes (Harvey et al., 1996).

The internal vertices of a phylogeny represent speciation events occurred throughout the evolution of the observed taxa and are usually constrained to have degree three. The degree constraint has not necessarily a biological foundation but it proves helpful when formalizing the evolutionary process of the analyzed taxa

*◇Graphes et Optimisation Mathématique (G.O.M.), Computer Science Department, Université Libre de Bruxelles (U.L.B.), Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium. Phone: +32 2 650 5628. Fax: +32 2 650 5970.

†Dipartimento di Matematica Applicata, Universitá Ca' Foscari, Dorsoduro 3246 - 30123, Venice, Italy. Phone: +39 041 2346927. Fax: +39 041 5221756.

‡Departamento de Estadística, Investigación Operativa y Computación, Universidad de La Laguna, Av. Astrofísico Francisco Sánchez, s/n 38271, La Laguna, Tenerife, Spain. Phone: +34 922 318184. Fax: +34 922 318170.

[1]See the online supplement for codes and data.

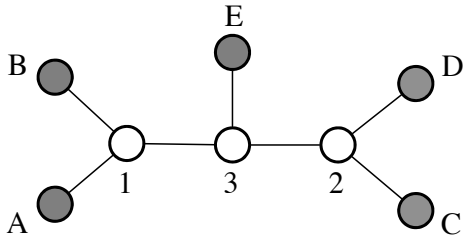

Figure 1: An example of a phylogeny of five taxa (A, B, C, D, E) and three internal vertices (1, 2, 3).

(see Catanzaro, 2010, p. 10). 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. On the other hand, the degree constraint helps in quantifying a-priori the number of edges and internal vertices of phylogeny T ($(2n-3)$ and $(n-2)$, respectively), otherwise hard to determine. As a drawback, the degree constraint implies that the overall number of possible phylogenies for a set of n taxa is $(2n-5)!!$, being $n!!$ the double factorial of n (Catanzaro, 2010). This fact entails the use of an estimation criterion to select a phylogeny from among plausible alternatives.

Different estimation criteria have been proposed in the literature on phylogenetics (see Catanzaro, 2010). Each criterion adopts its own set of hypotheses whose ability to describe the evolutionary process of taxa determines the gap between the *real* and the *true phylogeny*, i.e., the gap between the phylogeny that describes the real evolutionary process occurred in nature and the phylogeny that one would obtain, under the given the set of hypotheses, if all molecular data from taxa were available (Catanzaro, 2009). The criteria can usually 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. – *The Phylogenetic Estimation Problem (PEP)*

$$\begin{array}{rl} \textit{optimize} & f(T) \\ \textit{s.t.} & g(\Gamma, T) = 0 \\ & T \in \mathcal{T} \end{array}$$

where $\mathcal{T}$ is the set of $(2n-5)!!$ phylogenies of Γ, $f : \mathcal{T} \to \mathbb{R}$ a function modeling the selected criterion, and $g : \Gamma \times \mathcal{T} \to \mathbb{R}$ a function correlating the set Γ to a phylogeny T. The phylogeny T^* that optimizes f and satisfies g is referred to as *optimal*, and if T^* approaches the true phylogeny as the amount of molecular data extracted from taxa increases, the corresponding criterion is said to be *statistically consistent* (Gascuel, 2005). The statistical consistency is an important property in molecular phylogenetics because it measures the ability of a criterion to recover the true (and hopefully the real) phylogeny of the analyzed taxa.

In this article we investigate a recent version of the PEP, firstly introduced by Pauplin (2000) and called the *Balanced Minimum Evolution Problem* (BMEP). Specifically, given a set Γ of n taxa, consider a $n \times n$ symmetric distance matrix $\mathbf{D}$, whose generic entry d_{ij}, $i, j \in \Gamma$, represents a measure of dissimilarity between the corresponding pair of molecular data (Catanzaro, 2009). Then, the BMEP consists of finding a phylogeny T that minimizes the following *length function*

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

where the *topological distance* τ_{ij} represents the number of edges belonging to the path from taxon i to taxon j in T (Catanzaro, 2009).

The optimal solution T^* to the BMEP is known to be statistically consistent (see Desper and Gascuel, 2004), for this reason at least solving exactly the BMEP is highly desirable. Unfortunately, the $\mathcal{NP}$-hardness of the BMEP limits the size of the instances analyzable to the optimum (Fiorini and Joret, 2010). At present, instances of the BMEP containing more than 16 taxa constitute a hard computational challenge. To the best of our knowledge, the only attempts aiming at solving exactly instances of the BMEP are restricted to the

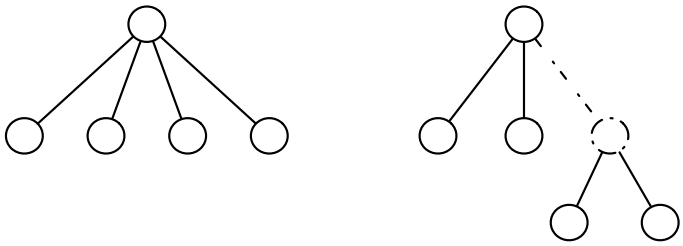

Figure 2: The 4-ary tree (on the left) can be transformed into a phylogeny by adding a dummy vertex and a dummy edge (dashed, on the right).

use of implicit enumeration algorithms such as those recently proposed by Pardi (2009). Specifically, from the combinatorial interpretation of the length function proposed by Semple and Steel (2004), Pardi derived a number of lower bounds for the problem that combined with ingenious speed-up techniques led to an exact algorithm able to tackle instances of the BMEP containing up to 20 taxa.

In this article we present an alternative and competitive exact approach to solution of the BMEP based on mixed integer linear programming. Specifically, we investigate the properties of the topological distances in order to provide a valid polynomial size formulation for the problem. Moreover, we develop families of strengthening valid inequalities, branching rules, and lower bounds to improve the performances of the formulation. Our results give perspective on the mathematics of the BMEP and suggest new directions on the development of future efficient exact approaches to solve this problem.

2 Notations and Properties of the Topological Distances

We investigate here some properties of the topological distances that will turn out useful to describe a possible valid formulation for the BMEP. Before that, we introduce some preliminary definitions that will prove useful throughout the paper.

Similarly to Parker and Ram (1996), by a *sequence*, we mean an ordered collection of nonnegative real values such as $\mathbf{x} = [x_1, x_2, \ldots, x_m]$, $x_j \in \mathbb{R}_{0^+}$. Repetition of values in the sequence is permitted: the values x_j need not be distinct. The *length* of this sequence is m and for simplicity we also refer to the set of such sequences with the vector notation $\mathbb{R}^m_{0^+}$.

Given a phylogeny T of Γ and a taxon $i \in \Gamma$, we denote Γ_i as the set $\Gamma \setminus \{i\}$, V as the set of $(n-2)$ internal vertices, and we define *path-length sequence* $\tau_i = [\tau_{ij} : j \in \Gamma_i]$ as the sequence of the topological distances relative to the $(n-1)$ paths from taxon i to each taxon $j \in \Gamma_i$ in T. Moreover, we define $\tau = [\tau_i : i \in \Gamma]$ as *path-length sequence collection* of the topological distances in T. For example, considered the phylogeny showed in Figure 1, the path-length sequence from taxon 'A' is $\tau_A = [2, 3, 4, 4]$ and the path-length sequence collection is $\tau = [\tau_A, \tau_B, \tau_C, \tau_D, \tau_E] = [[2, 3, 4, 4], [2, 3, 4, 4], [3, 3, 3, 3], [4, 4, 3, 2], [4, 4, 3, 2]]$.

We denote $\mathcal{T}$ as the set of all possible phylogenies for Γ, Θ as the set of path-length sequence collections τ associated to the phylogenies in $\mathcal{T}$, and, for each taxon $i \in \Gamma$, Θ_i as the set of all path-length sequences τ_i associated to the phylogenies in $\mathcal{T}$. Given a phylogeny T of Γ and a taxon $i \in \Gamma$, we denote $\mathbf{d}_i$ as the distance vector $\{d_{ij} : j \in \Gamma_i\}$ and $\hat{i}$ as the only vertex adjacent to i in T. For example, considered the phylogeny showed in Figure 1, if $i =$'A' then $\hat{i} = 1$. We assume that Γ is ordered and we use the notation $i < j$, for some i and $j \in \Gamma$, to mean that taxon i precedes taxon j in Γ. Moreover, we write $j = i + 1$ to mean that j immediately follows i in Γ.

We introduce now the main properties that characterize the topological distances of the phylogenies in $\mathcal{T}$. Since phylogenies are non-oriented graphs, the simplest property can be stated as follows:

$$\tau_{ij} = \tau_{ji} \tag{2}$$

for all $i, j \in \Gamma$, $i < j$. We refer to equation (2) as *the symmetry equality.*

A nontrivial property on the topological distances can be derived by from the analogies between phylogenies and *Huffman trees* (see Parker and Ram, 1996). Specifically, Huffman trees are rooted binary trees used in coding theory to represent symbols belonging to an alphabet $\hat{\Gamma}$. The leaves of a Huffman tree correspond to the symbols in $\hat{\Gamma}$, and the whole tree is usually described by means of path-length sequences $\rho = [\rho_j : j \in \hat{\Gamma}]$

from the root to each symbol $j \in \hat{\Gamma}$. Hence, given a phylogeny T of Γ and a taxon $i \in \Gamma$, if we disregard the edge $(i, \hat{i})$ in T, the remaining tree can be seen as a Huffman tree rooted in $\hat{i}$ and coding the symbols in Γ_i. Thus, the following proposition holds:

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

$$\sum_{j \in \Gamma_i} 2^{-\tau_{ij}} = \frac{1}{2}. \tag{3}$$

A direct consequence of the Kraft equality is that the BMEP is polynomially solvable if $d_{ij} = d$, $d \in \mathbb{R}_{0^+}$, for all $i, j \in \Gamma$. In fact, in this case, the Kraft equality implies that all phylogenies in $\mathcal{T}$ have the same length $\mathcal{L}(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} d_{ij}/2^{\tau_{ij}} = d \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} 1/2^{\tau_{ij}} = dn/2$. Hence, any phylogeny in $\mathcal{T}$ is an optimal solution to the BMEP.

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

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = (2n - 3). \tag{4}$$

Proof. From Pauplin (2000), we know that, for any phylogeny T with edgeset $\mathcal{E}(T)$ and for any set of edge weights $\{w_e : e \in \mathcal{E}\}$, the following condition holds: $\sum_{e \in \mathcal{E}} w_e = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \delta_{ij} 2^{-\tau_{ij}}$, where δ_{ij} is equal to the sum of the weights w_e along the path from taxon i to taxon j, for all $i \in \Gamma$ and $j \in \Gamma_i$. When setting $w_e = 1$, for all $e \in \mathcal{E}(T)$, we obtain $\delta_{ij} = \tau_{ij}$ and the statement follows. □

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

Proposition 3. *(Triangular inequalities) Let Γ be a set of n taxa. Then, for all the $T \in \mathcal{T}$, the following inequalities hold:*

$$\tau_{ik} + \tau_{kj} \geq \tau_{ij} + 2, \quad \forall i, j, k \in \Gamma \tag{5}$$

Proof. Let $P(i, j)$ the set of edges of a phylogeny T defining the path from taxon i to taxon j. As T is a tree, the following equality holds $P(i, j) = (P(i, k) \cup P(k, j)) \setminus (P(i, k) \cap P(k, j))$ (see Catanzaro et al., 2009). Then, since $P(i, k) \cap P(k, j) \supseteq \{k, \hat{k}\}$, it holds that $\tau_{ij} = |P(i, j)| = |(P(i, k)| + |P(k, j)| - 2|(P(i, k) \cap P(k, j))| = \tau_{ik} + \tau_{kj} - 2|(P(i, k) \cap P(k, j))| \leq \tau_{ik} + \tau_{kj} - 2$. □

With an abuse of notation, let us extend the definition of a path-length sequence collection also to the internal vertices of a phylogeny $T \in \mathcal{T}$. Then, the following property holds:

Proposition 4. *(Four-point condition, Buneman (1974)) Let Γ be a set of n taxa, and let $i, j, q, t \in \Gamma \cup V$, $i \neq j \neq q \neq t$. Then, any phylogeny $T \in \mathcal{T}$ contains no triangle and satisfies the following condition:*

$$\tau_{ij} + \tau_{qt} \leq \max\{\tau_{iq} + \tau_{jt}, \tau_{it} + \tau_{jq}\}. \tag{6}$$

Equation (6) derives from a restriction of a more general property relative to additive matrices described in Buneman (1974). Proposition 1 completely characterizes the path-length sequences that belongs to Θ_i, i.e., it states that the integrity of the topological distances τ_{ij} and the Kraft equality (3) are necessary and sufficient conditions for a sequence τ_i to belong to Θ_i. Similarly, it is easily seen that conditions (3) and (6) completely characterize the path-length sequence collections that belong to Θ.

An interesting question is whether the restriction of the four-point condition to Γ instead of $\Gamma \cup V$, together with conditions (2), (3), and (4) suffice to completely characterize the path-length sequence collections in Θ. At present we known that these conditions are necessary and independent even when we restrict our attention to integral sequences. For example, given five taxa, a sequence collection τ whose path-length sequences are $\tau_i = [3, 3, 3, 3]$, for all $i \in \Gamma$, satisfies (2), (3), and (6), but not (4). Hence, τ cannot be associated to any phylogeny T of 5 taxa. We have also experienced that conditions (2), (3), (4), and (6) are sufficient to guarantee that a sequence collection τ belongs to Θ whenever $|\Gamma| \leq 15$. This fact led us to suspect that these conditions could also be in general sufficient, however we do not have a formal proof of this conjecture so far.

3 A MIP Formulation for the BMEP

The fundamental properties of the topological distances discussed in the previous section suggest as possible approach to solution of the BMEP the use of mathematical programming. In this section we shall develop a possible polynomial size mixed integer linear programming model for the BMEP. Moreover, we shall also present a number of valid inequalities to further strengthen such a model.

Consider the following binary decision variables

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

where $L = \{1, 2, 3, \ldots, (n-1)\}$. Similarly, consider the following set of binary decision variables introduced to linearize the max function in (6):

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

Then, we can formulate the BMEP in terms of the following mixed integer programming model:

Formulation 1. *Path-Length-4 point (PL4)*

$$\min z = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} d_{ij} \left(\sum_{k \in L \setminus \{1\}} 2^{-k} x_{ij}^k \right) \tag{7a}$$

$$s.t. \quad \sum_{k \in L} x_{ij}^k = 1 \qquad \forall\ i \neq j \in \Gamma \cup V \tag{7b}$$

$$x_{ji}^k = x_{ij}^k \qquad \forall\ i < j \in \Gamma \cup V,\ k \in L \tag{7c}$$

$$\sum_{j \in \Gamma_i} \sum_{k \in L \setminus \{1\}} 2^{-k} x_{ij}^k = \frac{1}{2} \qquad \forall\ i \in \Gamma \tag{7d}$$

$$\sum_{k \in L \setminus \{1\}} k 2^{-k} \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} x_{ij}^k = (2n-3) \tag{7e}$$

$$\sum_{k \in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k \in L} k(x_{iq}^k + x_{jt}^k) + (2n-2) y_{ijqt} \qquad \forall\ i \neq j \neq q \neq t \in \Gamma \cup V \tag{7f}$$

$$\sum_{k \in L} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k \in L} k(x_{it}^k + x_{jq}^k) + (2n-2)(1 - y_{ijqt}) \qquad \forall\ i \neq j \neq q \neq t \in \Gamma \cup V \tag{7g}$$

$$x_{ij}^1 = 0 \qquad \forall\ i \neq j \in \Gamma \tag{7h}$$

$$\sum_{i,j \in \Gamma \cup V, i \neq j} x_{ij}^1 = (2n-3) \tag{7i}$$

$$\sum_{j \in V} x_{ij}^1 = 1 \qquad \forall\ i \in \Gamma \tag{7j}$$

$$\sum_{j \in \Gamma \cup V, i \neq j} x_{ij}^1 = 3 \qquad \forall\ i\ \in\ V \tag{7k}$$

$$x_{ij}^1 + x_{il}^1 + x_{lj}^1 \leq 2 \qquad \forall\ i \neq j \neq l \in V \tag{7l}$$

$$x_{ij}^k + 1 \geq x_{il}^{(k-1)} + x_{lj}^1 \qquad \forall\ i \neq j \in \Gamma,\ l \in V, k \in L \setminus \{1, n-1\} \tag{7m}$$

$$x_{ij}^k + x_{ij}^{(k-2)} + 1 \geq x_{il}^{(k-1)} + x_{lj}^1 \qquad \forall\ i \neq j \neq\ l \in \Gamma \cup V, k \in L \setminus \{1, 2, n-1\} \tag{7n}$$

$$x_{ij}^k \in \{0, 1\} \qquad \forall\ i, j \in \Gamma \cup V, k \in L \tag{7o}$$

$$y_{ijqt} \in \{0, 1\} \qquad \forall\ i \neq j \neq q \neq t \in \Gamma \cup V. \tag{7p}$$

Constraints (7b) impose that variables τ_{ij} assume exactly one value in L. Constraints (7c) impose the symmetry equalities (2). Constraints (7d) impose the Kraft equalities (3). Constraint (7d) imposes the third equality (4). Constraints (7f) and (7g) impose the four-point inequalities (6). Constraints (7h)-(7n) describe the structure of a phylogeny. Specifically, constraint (7h) imposes that no edge exists between taxa in Γ. Constraint (7i) imposes that exactly $(2n-3)$ edges be present in a phylogeny. Constraints (7j) and

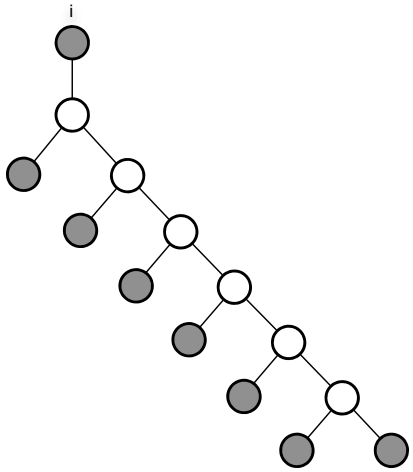

Figure 3: An example of the most imbalanced phylogeny for $n = 8$.

(7k) impose the degree constraint on vertices of a phylogeny. Constraints (7l) prevent triangles. Finally, constraints (7m) and (7n) link edge variables (x_{ij}^k, $k = 1$) to path variables (x_{ij}^k, $k \geq 2$).

Interestingly, alternative exponential size formulations for the BMEP can be obtained either by removing the four-point inequalities and imposing the standard anti-cycle constraints or by using a column generation approach similar to the one proposed by Fischetti et al. (2002) for the minimum routing cost tree. However, preliminary tests showed that these formulations perform worse than PL4, for this reason we do not describe them in this article.

3.1 Strengthening Valid Inequalities

By exploiting the integrality of variables x_{ij}^k a number of valid inequalities can be developed to strengthen PL4.

Proposition 5. *The inequality*

$$\sum_{j \in \Gamma_i} x_{ij}^{(n-1)} \leq 2 \sum_{j \in \Gamma_i} x_{ij}^k \quad \forall\ i \in \Gamma,\ k \in L \setminus \{1, (n-1)\} \tag{8}$$

is valid for PL4.

Proof. For a fixed phylogeny $T \in \mathcal{T}$ and taxon $i \in \Gamma$, either there exist exactly two paths in T from taxon i having length $(n-1)$ or none. When $\sum_{j \in \Gamma_i} x_{ij}^{(n-1)} = 0$, the inequality (8) reduces to $\sum_{j \in \Gamma_i} x_{ij}^k \geq 0$ which is trivially valid. When $\sum_{j \in \Gamma : j \neq i} x_{ij}^{(n-1)} = 2$, the inequality (8) reduces to $\sum_{j \in \Gamma_i} x_{ij}^k \geq 1$ which is valid again as the presence of at least one path of length $(n-1)$ implies the presence of a path of length $(n-2)$, $(n-3)$, and so on. □

Definition 1. *Given a set Γ of n taxa and a taxon $i \in \Gamma$, a phylogeny $\overline{T} \in \mathcal{T}$ is said* a most imbalanced phylogeny with respect to *i if $\overline{T}$ includes two paths from i having length $(n-1)$.*

As an example, Figure 3 shows the most imbalanced phylogeny for $n = 8$.

Proposition 6. *The inequality*

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} x_{ij}^{(n-1)} \leq 8, \tag{9}$$

$n \geq 4$, is valid for PL4.

Proof. It is easy to see that inequality (9) is trivially valid for PL4 as any imbalanced phylogeny of four or more taxa presents exactly eight paths having length $(n-1)$ and any other phylogeny present no paths of length $(n-1)$. □

Proposition 7. *The inequality*

$$x^2_{ij} - 1 \leq x^k_{iq} - x^k_{jq} \leq 1 - x^2_{ij}, \quad \forall\ i,j,q \in \Gamma,\ \forall\ k\ \in L \setminus \{1\}, \tag{10}$$

is valid for PL4.

Proof. When $x^2_{ij} = 0$ inequalities (10) are trivially valid for PL4. When $x^2_{ij} = 1$, taxa i and j are adjacent to the same internal vertex, hence $\tau_{iq} = \tau_{jq}$ for all $q \in \Gamma_i \cap \Gamma_j$, thus (10) is again valid. □

Proposition 8. *The inequality*

$$\sum_{k\,=\,\max\{2,\,|m-l|\}}^{m+l-2} x^k_{iq} + 1 \geq x^m_{ij} + x^l_{jq}, \quad \forall\ i,j,q \in \Gamma,\ \forall\ m,l \in L \setminus \{1\} \tag{11}$$

is valid for PL4.

Proof. By (7b), the left-hand-side of (11) can assume only values 1 or 2. When the left-hand-side of (11) is equal to 2, (11) is trivially valid for PL4. When the left-hand-side of (11) is equal to 1, τ_{iq} is either greater than $m+l-2$ or less then $|m-l|$. Then, by the triangular inequalities, at most one between x^m_{ij} and x^l_{jq} can be equal 1, thus (11) is again valid for PL4. □

Proposition 9. *Let $q \in \mathbb{N}$, $q \geq 2$. Then, if $n > 2^{q-1}+1$, the following inequality*

$$\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij} \leq 2^{q-1} - 1 \quad \forall i \in \Gamma \tag{12}$$

is valid for PL4.

Proof. Multiplying the Kraft equality by 2^q we obtain that

$$\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij} = 2^{q-1} - \sum_{j\in\Gamma_i}\sum_{k=q+1}^{n-1} 2^{q-k} x^k_{ij} \leq 2^{q-1}.$$

Note that in any feasible solution either $\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij} = 0$ or $\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij} \neq 0$. In the former case (12) is valid for PL4. In the latter case, if $n > 2^{q-1}+1$, it holds that $\sum_{j\in\Gamma_i}\sum_{k=q+1}^{n-1} 2^{q-k} x^k_{ij} > 0$ otherwise we would have a contradiction of (3). Hence, as the coefficients of $\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij}$ are integers and 2^{q-1} is integer we have that $\sum_{j\in\Gamma_i}\sum_{k=2}^{q} 2^{q-k} x^k_{ij} \leq 2^{q-1} - 1$ and (12) holds valid again. □

Proposition 10. *Fixed a taxon $i \in \Gamma$, the following inequalities are valid for PL4:*

1. $\sum_{k\in L}(k-1)x^k_{ij} \geq 1$, *for all $j \in \Gamma_i$ and $n \geq 6$;*
2. $\sum_{k\in L}(k-1)x^k_{ij_1} + \sum_{h\in L}(h-1)x^h_{ij_2} \geq 3$, *for all $j_1, j_2 \in \Gamma_i$, $j_1 \neq j_2$, and $n \geq 7$;*
3. $\sum_{j\in\Gamma_i}\sum_{k\in L}(k-1)x^k_{ij} \leq \frac{n(n-1)}{2}$;
4. $\sum_{j\in\Gamma_i}\sum_{k\in L}(k-1)x^k_{ij} \geq 2c(r+1) + (n-1-2c)r$, *for all $r \geq 1$, $0 < c < 2^r$, and $n = 2^r + c + 1$;*
5. $\sum_{h=1}^{h-1} F_{h-1}\sum_{k\in L}(k-1)x^k_{ij_h} \geq F_{n+3} - 3$, $n \geq 4$, *where F_h is the h-th element of the Fibonacci sequence, with the convention that $F_0 = 1$, and j_h indicates the h-th taxon in Γ_i according any arbitrary sorting of taxa in Γ_i.*

Proof. The statement can be easily derived from Propositions 3.4, 3.5 and Theorem 3.8 of Maurras et al. (2010). Specifically, the two propositions and the theorem describe some facets of the convex hull of Huffman trees in terms of the path-lengths from the root of the tree to the leaves. We recall that, given a phylogeny T of Γ and a taxon $i \in \Gamma$, if we disregard edge $(i,\hat{i})$ in T, the remaining tree can be seen as a Huffman tree rooted in $\hat{i}$ and coding the remaining $(n-1)$ taxa in Γ_i. Hence, by definition of variables x^k_{ij} and constraint (7b), it is easy to see that the length of the path from $\hat{i}$ to each taxon $j \in \Gamma_i$, expressed in terms of variables x^k_{ij}, is equal to $\sum_{k\in L}(k-1)x^k_{ij}$. Note that the factor $(k-1)$ is due to the fact that edge $(i,\hat{i})$ is disregarded. □

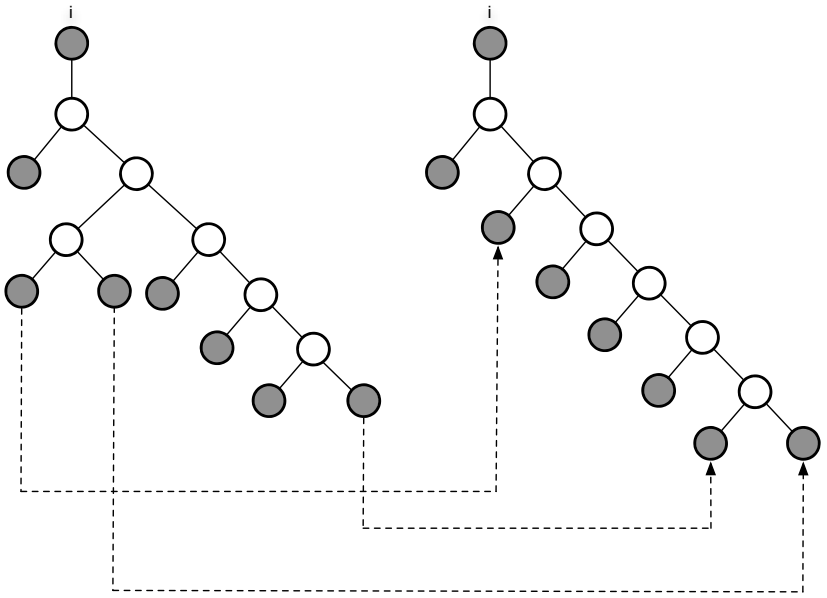

Figure 4: An example of imbalancing exchange between path-length sequences $\tau_i^0 = [2, \mathbf{4}, \mathbf{4}, 4, 5, 6, \mathbf{6}]$ (phylogeny on the left) and $\tau_i^1 = [2, \mathbf{3}, 4, 5, 6, \mathbf{7}, \mathbf{7}]$ (phylogeny on the right).

3.2 Advanced Strengthening Valid Inequalities

By exploiting the analogies between phylogenies and *Huffman trees* (see Parker and Ram, 1996), a further set of strengthening valid inequalities for PL4 can be derived from the propositions presented in this section by replacing τ_{ij} with $\sum_{k \in L} k x_{ij}^k$ and $2^{-\tau_{ij}}$ with $\sum_{k \in L} 2^{-k} x_{ij}^k$. Before proceeding, we introduce the following operators on sequences which will prove useful throughout this section:

1. ascending sort, in symbols $\overrightarrow{\mathbf{x}} = [\mathbf{x}$ put in ascending order$]$;
2. descending sort, in symbols $\overleftarrow{\mathbf{x}} = [\mathbf{x}$ put in descending order$]$.

Moreover, given a set Γ of n taxa and a taxon $i \in \Gamma$, we denote $\gamma_i = [2, 3, 4, \ldots, (n-3), (n-2), (n-1), (n-1)]$ as the most imbalanced path-length sequence. It is easy to see that if $\overline{T}$ is a most imbalanced tree with respect to i then $\gamma_i = \overrightarrow{\tau}_i$. For some $T \in \mathcal{T}$ and taxon $i \in \Gamma$, consider the path-length sequence τ_i. Assume that $\overrightarrow{\tau}_i$ is such that its k-th and $(k+1)$-th entries, $k < n-2$, are equal, i.e., $\overrightarrow{\tau}_i = [\ldots, p, p, \ldots, q]$, with $p \leq q < (n-1)$.

Definition 2. *An* imbalancing exchange *on τ_i is the operation that returns the sequence*

$$\tau_{ij}^1 = \begin{cases} \overrightarrow{\tau}_{ij} & j = 1, \ldots, k-1 \\ p-1 & j = k \\ \overrightarrow{\tau}_{i(j+1)} & j = k+1, \ldots, (n-3) \\ q+1 & j = (n-2), (n-1). \end{cases}$$

Figure 4 shows an example of imbalancing exchange.

Proposition 11. *(Parker and Ram, 1996) Consider a phylogeny $T \in \mathcal{T}$ and a taxon $i \in \Gamma$. Either $\overrightarrow{\tau}_i$ is equal to γ_i or $\overrightarrow{\tau}_i$ has at least two identical entries, say the k-th and the $(k+1)$-th ones, different from $(n-1)$. In the latter case, there exists τ_i^1 associated to a phylogeny $T^1 \in \mathcal{T}$ such that τ_i^1 is obtained from τ_i by means of an imbalancing exchange.*

The repeated application of previous proposition generates a finite sequence of imbalancing exchanges leading from τ_i to γ_i.

Consider a taxon $i \in \Gamma$, a path-length sequence τ_i, and an associated weight vector $\mathbf{v} = \{v_j : j \in \Gamma_i\}$. Let τ_i^1 be a path-length sequence obtained through an imbalancing exchange on τ_i. Then, we associate the

following weight vector $\mathbf{v}^1 = \{v_j^1 : j \in \Gamma_i\}$ to τ_i^1:

$$v_j^1 = \begin{cases} \overleftarrow{v}_j & j = 1, \ldots, k \\ \overleftarrow{v}_{j+1} & j = k+1, \ldots, (n-2) \\ \overleftarrow{v}_{k+1} & j = (n-1). \end{cases}$$

The following proposition holds:

Proposition 12. *Given a taxon $i \in \Gamma$ and a weight vector $\mathbf{v} = \{v_j : j \in \Gamma_i\}$, the following inequality holds for all path-length sequences $\tau_i \in \Theta_i$:*

$$\sum_{j \in \Gamma_i} v_j 2^{-\tau_{ij}} \leq \sum_{j \in \Gamma_i} \overleftarrow{v_j} 2^{-\gamma_{ij}} \tag{13}$$

where γ_{ij} is the j-th element of the most imbalanced path-length sequence $\gamma_i = [2, 3, 4, \ldots, (n-3), (n-2), (n-1), (n-1)]$.

Proof. Consider any path-length sequence $\tau_i^0 \in \Theta_i$ and note that the following condition holds:

$$\sum_{j \in \Gamma_i} v_j 2^{-\tau_{ij}} \leq \sum_{j \in \Gamma_i} \overleftarrow{v}_j^0 2^{-\overrightarrow{\tau}_{ij}^0}. \tag{14}$$

If $\overrightarrow{\tau}_i^0 = \gamma_i$ then the statement trivially holds. If $\overrightarrow{\tau}_i^0 \neq \gamma_i$, consider a path-length sequence τ_i^1 obtained from τ_i^0 through a imbalancing exchange. Then, it holds that

$$\sum_{j \in \Gamma_i} \overleftarrow{v}_j^0 2^{-\overrightarrow{\tau}_{ij}^0} \leq \sum_{j \in \Gamma_i} v_j^1 2^{-\overrightarrow{\tau}_{ij}^1} \leq \sum_{j \in \Gamma_i} \overleftarrow{v}_j^1 2^{-\overrightarrow{\tau}_{ij}^1} = \sum_{j \in \Gamma_i} \overleftarrow{v}_j^0 2^{-\overrightarrow{\tau}_{ij}^1}. \tag{15}$$

In fact, the first inequality in (15) holds as

$$\begin{aligned} &\sum_{j \in \Gamma_i} \overleftarrow{v}_j^0 2^{-\overrightarrow{\tau}_{ij}^0} - \sum_{j \in \Gamma_i} v_j^1 2^{-\overrightarrow{\tau}_{ij}^1} = \\ &= (2^{-p} \overleftarrow{v}_k^0 + 2^{-p} \overleftarrow{v}_{k+1}^0 + 2^{-q} \overleftarrow{v}_{n-1}^0) - (2^{-(p-1)} \overleftarrow{v}_k^0 + 2^{-(q+1)} (\overleftarrow{v}_{k+1}^0 + \overleftarrow{v}_{n-1}^0)) = \\ &= 2^{-p} \overleftarrow{v}_{k+1}^0 - (2^{-(p-1)} - 2^{-p}) \overleftarrow{v}_k^0 + (2^{-q} - 2^{-(q+1)}) \overleftarrow{v}_{k+1}^0 - 2^{-(q+1)} \overleftarrow{v}_{n-1}^0 = \\ &= 2^{-p} \underbrace{(\overleftarrow{v}_{k+1}^0 - \overleftarrow{v}_k^0)}_{\leq 0} + 2^{-(q+1)} \underbrace{(\overleftarrow{v}_{n-1}^0 - \overleftarrow{v}_{k+1}^0)}_{\leq 0} \leq 0. \end{aligned}$$

Similarly, the last equality in (15) holds as v^0 and v^1 include the same entries, possibly only in a different order. Now, if $\overrightarrow{\tau}_i^1 = \gamma_i$ the statement is proved. Otherwise, redefine $\tau_i^0 = \overrightarrow{\tau}_i^1$ and $v^0 = \overleftarrow{v}^1$ and iterate the above argumentation until $\overrightarrow{\tau}_i^1 = \gamma_i$. Note that, due to Proposition 11, the number of such iterations is finite. □

For any $S \subset \Gamma_i$, let $\mathbf{v}^S$ be the incidence vector of S. Then, an immediate consequence of Proposition 12 is that the following inequalities hold:

$$\begin{aligned} \textstyle\sum_{j \in \Gamma_i} v_j^S 2^{-\tau_{ij}} &\leq \textstyle\sum_{k=1}^{|S|} 2^{-(k+1)} = 2^{-1} - 2^{-(|S|+1)} \\ \textstyle\sum_{j \in \Gamma_i} -v_j^S 2^{-\tau_{ij}} &\leq -(2^{-(n-1)} + \textstyle\sum_{k=1}^{|S|-1} 2^{(-n+k)}) = -2^{(-n+|S|)} \end{aligned}$$

i.e., $2^{(-n+|S|)} \leq \sum_{j \in \Gamma_i} v_j^S 2^{-\tau_{ij}} \leq 2^{-1} - 2^{-(|S|+1)}$. Note that, when $S = \Gamma_i$, Proposition (12) implies the Kraft equality.

Definition 3. *Given a phylogeny T of Γ and a subset $S \subseteq \Gamma$, a* cycle through *S, denoted as C_S, is a sequence of paths between pair of taxa in S that forms a closed walk in which each taxon in S is visited only once, each used edge is visited twice, and taxa in $\Gamma \setminus S$ are not visited.*

As an example, if $S = \Gamma$, a possible cycle for the phylogeny shown in Figure 1 is $\{e_{A1}, e_{1B}, e_{B1}, e_{13}, e_{3E}, e_{E3}, e_{32}, e_{2C}, e_{C2}, e_{2D}, e_{D2}, e_{23}, e_{31}, e_{1A}\}$. The topological length of a cycle C_S is equal to the number of edges defining the closed walk and is denoted by $length(C_S)$. We denote C_S^* as a cycle of minimal length and by $length(C_S^*)$ its length. We understand that C_S is identified by the ordered set of pairs of taxa delimiting the paths composing the the closed walk, e.g., $C_S = \{(i_1, i_2), (i_2, i_3), (i_3, i_1)\}$ defines the closed walk on T from taxon i_1 to taxon i_2 to taxon i_3 and, finally, back to taxon i_1.

For any given $S \subseteq \Gamma$, $length(C_S^*)$ is equal to twice the number of edges of the smallest subtree Y of T having as leaves taxa in S. In turn, Y has at least $2(|S| + 1) - 4 = 2|S| - 2$ edges, as it must describe phylogenies whose leaves are taxa in S. Y is minimal when the phylogeny T presents a bridge edge (j, k) such that, if we remove edge (j, k) from T, we obtain two rooted binary trees: one rooted in j and whose leaves are taxa in S, and another rooted in k and whose leaves are taxa in $\Gamma \setminus S$. Hence, $length(C_\Gamma^*) = 2(2n - 3)$, for all $T \in \mathcal{T}$ and, in general, $length(C_S^*) \geq 4(|S| - 1)$, for $S \subset \Gamma$ and for all $T \in \mathcal{T}$. This insight justifies the following proposition.

Proposition 13. *(Cycle inequalities) For all $\tau \in \Theta$, the following inequality holds*

$$\sum_{(i,j) \in C_S} \tau_{ij} \geq 4(|S| - 1), \quad \forall C_S, \ \forall S \subset \Gamma. \tag{16}$$

Given a sequence σ determining the existence of an inequality of type (16) that separates σ from Θ is generally not easy. However, many heuristics for the Traveling Salesman Problem (TSP) (Garey and Johnson, 2003) can be employed to determine a cycle C_S suboptimal for $\sum_{(i,j) \in C_S} \tau_{ij}$.

Hereafter, we say that two taxa of a phylogeny T are *twins* if they share their immediate common ancestor (e.g., taxa A and B in Figure 1).

Proposition 14. *(2-Tree inequality) Given a phylogeny T of Γ, the following inequality holds*

$$2^{-\tau_{ik}} - 2^{-\tau_{jk}} \leq 2^{-2}(1 - 4 \cdot 2^{-\tau_{ij}}) \tag{17}$$

for any three distinct taxa i, j, and k in Γ.

Proof. If it holds that $2^{-\tau_{ij}} = 2^{-2}$ then it follows that i and j are twins T, hence $\tau_{ik} = \tau_{jk}$ and (17) holds as an equality. Alternatively, if $2^{-\tau_{ij}} \leq 2^{-3}$ then $2^{-2}(1 - 4 \cdot 2^{-\tau_{ij}}) \geq 2^{-3}$. In this case two situations may occur: either $2^{-\tau_{ik}} = 2^{-2}$ or $2^{-\tau_{ik}} \leq 2^{-3}$. If $2^{-\tau_{ik}} = 2^{-2}$ then it follows that i and k are twins in T, hence $2^{-\tau_{jk}} = 2^{-\tau_{ij}}$ and (17) holds again as an equality. Differently, if $2^{-\tau_{ik}} \leq 2^{-3}$ then (17) trivially holds. □

A possible extension of the 2-Tree inequality can be obtained as follows.

Proposition 15. *(3-Tree inequality) Let S be a proper subset of Γ containing three distinct taxa i_1, i_2, and i_3, and let k be a taxon not in S. Then, given a phylogeny T of Γ, the following inequality holds:*

$$2^{-\tau_{i_1 k}} - 2^{-\tau_{i_2 k}} - 2^{-\tau_{i_3 k}} \leq 2^{-1} - 2^{-\tau_{i_1 i_2}} - 2^{-\tau_{i_2 i_3}} - 2^{-\tau_{i_1 i_3}}. \tag{18}$$

Proof. Note first that in a phylogeny T of at least four taxa the path-lengths relative to taxa in S cannot be all equal to 2. Moreover, note also that the sum $2^{-\tau_{i_1 i_2}} + 2^{-\tau_{i_2 i_3}} + 2^{-\tau_{i_1 i_3}}$ in T is either equal to 2^{-1} or assumes values less than or equal to $2^{-2} + 2^{-3}$. Specifically, the value 2^{-1} is obtained only when two taxa in S are twins and the path-lengths from the twins to the third taxon are equal to 3 (consider e.g., taxa A, B and E in Figure 1). Alternatively, a value less than or equal to $2^{-2} + 2^{-3}$ is obtained if: (i) two taxa in S are twins and the path-lengths from the twins to the third taxon are greater than or equal to 4, in which case we would have $2^{-\tau_{i_1 i_2}} + 2^{-\tau_{i_2 i_3}} + 2^{-\tau_{i_1 i_3}} \leq 2^{-2} + 2^{-4} + 2^{-4} = 2^{-2} + 2^{-3}$; (ii) no pair of taxa in S are twins, in which case we would have $2^{-\tau_{i_1 i_2}} + 2^{-\tau_{i_2 i_3}} + 2^{-\tau_{i_1 i_3}} \leq 2^{-3} + 2^{-3} + 2^{-3} = 2^{-2} + 2^{-3}$. Hence, the minimum value of the right-hand-side of (18) is 0. If this circumstance occurs as i_1 and i_2 (or i_1 and i_3) are twins, then $2^{-\tau_{i_1 k}} = 2^{-\tau_{i_2 k}}$ (or $2^{-\tau_{i_1 k}} = 2^{-\tau_{i_3 k}}$), hence (18) holds since its left-hand-side is negative. If the right-hand-side of (18) is null as i_2 and i_3 are twins and $2^{-\tau_{i_1 i_2}} = 2^{-\tau_{i_1 i_3}} = 2^{-3}$, then the path-length from each taxon $k \notin S$ to i_2 and i_3 contains an edge more than the path-length from k to i_1 (consider, e.g., Figure 1 in which $i_1 = E$, $i_2 = A$, $i_3 = B$, and $k = D$). Hence, we have $2^{-\tau_{i_2 k}} = 2^{-\tau_{i_3 k}} = 2^{-1} 2^{-\tau_{i_1 k}}$ and (18) holds as an equality. When the right-hand-side of (18) is greater than 0, its value is at least equal to 2^{-3}. Then, if $2^{-\tau_{i_1 k}} \leq 2^{-3}$, (18) trivially holds. When the right-hand-side is strictly greater than zero and $2^{-\tau_{i_1 k}} = 2^{-2}$ then i_1 and k are twins, hence $2^{-\tau_{i_2 k}} = 2^{-\tau_{i_1 i_2}}$ and $2^{-\tau_{i_3 k}} = 2^{-\tau_{i_1 i_3}}$. As $2^{-\tau_{i_2 i_3}} \leq 2^{-2}$, (18) holds again. Finally, when $2^{-\tau_{i_2 i_3}} = 2^{-2}$ (18) holds as an equality. □

With analogous argumentations we can prove that the 4-Tree inequality, involving a taxon k not in $S = \{i_1, i_2, i_3, i_4\}$, is:

$$2^{-\tau_{i_1 k}} - 2^{-\tau_{i_2 k}} - 2^{-\tau_{i_3 k}} - 2^{-\tau_{i_4 k}} \leq 2^{-1} + 2^{-3} - 2^{-\tau_{i_1 i_2}} - 2^{-\tau_{i_2 i_3}} - 2^{-\tau_{i_3 i_4}} - 2^{-\tau_{i_1 i_4}}. \quad (19)$$

4 Testing the Performances of PL4

In order to evaluate the efficiency of our exact approach to solution of the BMEP we tested the performances of PL4 on a number of real aligned DNA datasets, namely: "Primates12/898", a dataset of 12 sequences, 898 characters each from primates mitochondrial DNA; "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 62 sequences of 3768 characters each from hyracoidae; finally, "SeedPlant25/19784", a dataset of 25 sequences of 19784 characters each from pinoles. From each dataset we have extracted the first 20 taxa (or all taxa if $n < 20$) and built the associated $n \times n$ distance matrices by using the General Time Reversible (GTR) model of DNA sequence evolution in which all the gaps were treated as 'N'. The estimation method used to obtained GTR distances is described in Catanzaro et al. (2006). 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 BMEP. Datasets and corresponding distance matrices can be found in the online supplement for codes and data.

We implemented PL4 in ANSI C++ by using Xpress Optimizer libraries v18.10.00. The experiments run on a Pentium 4, 3.2 GHz, equipped with 2 GByte RAM and operating system Gentoo release 7 (kernel linux 2.6.17). During the runtime of PL4 we activated the Xpress automatic cuts, the Xpress pre-solving strategy, and used the Xpress primal heuristic to generate the first upper bound for the problem. Moreover, we used a branch-and-cut approach to add dynamically the four-point, the cycle, the triangular, and the r-Tree inequalities. Actually, already for $n = 12$ the number of inequalities introduced in the formulation just by the four-point condition approaches about a million, slowing down sensibly the simplex solver. We assumed one hour as maximum runtime per instance and rescaled the objective function by a factor 2^n in order to reduce possible numerical stability problems.

In order to obtain a measure of the performances of PL4 we considered, as reference, the performances of a simplified version of Pardi (2009) exact approach to solution of the BMEP running on the same instances. Specifically, Pardi's approach is based on a *stepwise addition strategy* (see Felsenstein, 2004), a peculiar implicit enumeration procedure that can be resumed as follows. For any subset $S \subseteq \Gamma$, define a *subphylogeny* $Y(S)$ as any phylogeny that involves only taxa in S. Let $\mathcal{E}(Y(S))$ be the edgeset of $Y(S)$. Moreover, for a given subphylogeny $Y(S)$, taxon $i \in \Gamma \setminus S$, and edge $(r, s) \in \mathcal{E}(Y(S))$, define a *branching* as the operation

$$Y(S \cup \{i\}) = Y(S) \oplus_{(r,s)} i = (S \cup \{i\}, (\mathcal{E}(Y(S)) \setminus \{(r, s)\}) \cup \{(r, \hat{i}), (\hat{i}, s), (\hat{i}, i)\})$$

i.e., as the process that returns the subphylogeny $Y(S \cup \{i\})$ obtained inserting a new edge $(\hat{i}, i)$ on the edge (r, s) of $Y(S)$. Figure 5 shows an example of branching.

We say that a phylogeny T is *generated* from $Y(S)$ if T is obtained by recursive branching of $Y(S)$. Finally, consider the following subroutines:

Head(t, Γ) returning the t-th element of set Γ.

Bound$(S, Y(S), T)$ computing a lower bound on the length of the shortest phylogeny $\hat{T}$ that can be generated from $Y(S)$. If the lower bound is less than the length of the currently optimal phylogeny T the subroutine returns TRUE, FALSE otherwise.

Search($S, Y(S), T$**)** recursively branching the subphylogeny $Y(S)$ in search of the shortest phylogeny $\hat{T}$ that can be generated from $Y(S)$ (see Algorithm 1). SEARCH() interrupts its recursion whenever BOUND$(S, Y(S), T)$ return FALSE, in which case we say that the phylogenies that can be generated from $Y(S)$ are *pruned*. Alternatively, SEARCH() continues the branching process until all the phylogenies generable from $Y(S)$ are computed.

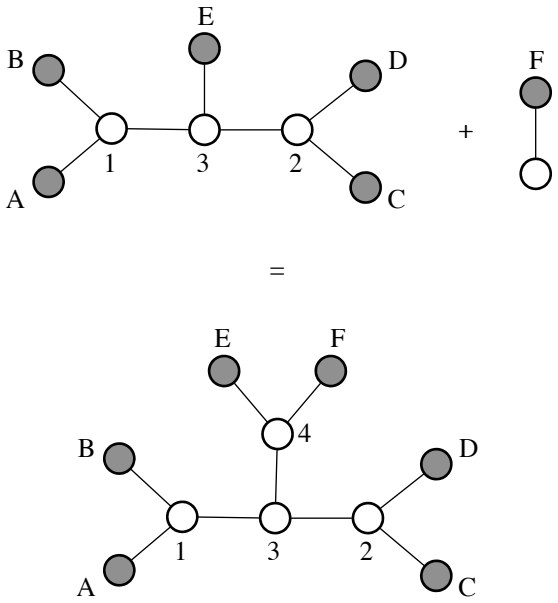

Figure 5: An example of branching: $Y(\{A, B, C, D, E, F\}) = Y(\{A, B, C, D, E\}) \oplus_{(E,3)} F$

Then, the stepwise addition strategy can be outlined as in Algorithm 2. Specifically, the algorithm initially sets the currently optimal phylogeny T to an empty tree ($NULL$) and fixes its length to $+\infty$. Subsequently, it generates the only possible subphylogeny consisting of the first three taxa in Γ and finally calls the subroutine SEARCH() to find the optimal phylogeny to BMEP. We show in Figure 6 some of the first subtrees explored by the subroutine SEARCH().

Pardi (2009) investigated a number of possible combinatorial lower bounds for the BMEP and developed several computational techniques, inspired by Desper and Gascuel (2002), that may significantly speed-up computations. The implementation of those techniques is out the scope of the article, thus in our experiments we just considered a simplified version of Pardi's procedure in which no speed-up techniques was implemented. As regards to the lower bound for the problem, we used the one proposed in (7.3.7) from Pardi (2009), which can be stated as follows:

$$\mathcal{L}(T^*) \geq \mathcal{L}(Y(S)) + \sum_{f \notin S} \min_{\substack{i,j \in S: \\ i<j<f}} \frac{1}{2}(d_{if} + d_{fj} - d_{ij}) \qquad \forall\ S \subseteq \Gamma.$$

In fact, it is possible to prove that the change induced in the length of a subphylogeny $Y(S)$ by means of a

```
1 SEARCH(S, Y(S), T);
  Input  : S: a subset of taxa
           Y(S): a subphylogeny
           T: current optimal phylogeny
  Output : A phylogeny T solution of the BMEP
2 if BOUND(S, Y(S), T) then
3     if S = Γ then
4         T = Y(S);
5     else
6         set i = HEAD(|S| + 1, Γ);
7         for e ∈ E(T(S)) do
8             SEARCH(S, Y(S) ⊕_e i, T);
9 return T;
```

Algorithm 1: Subroutine SEARCH().

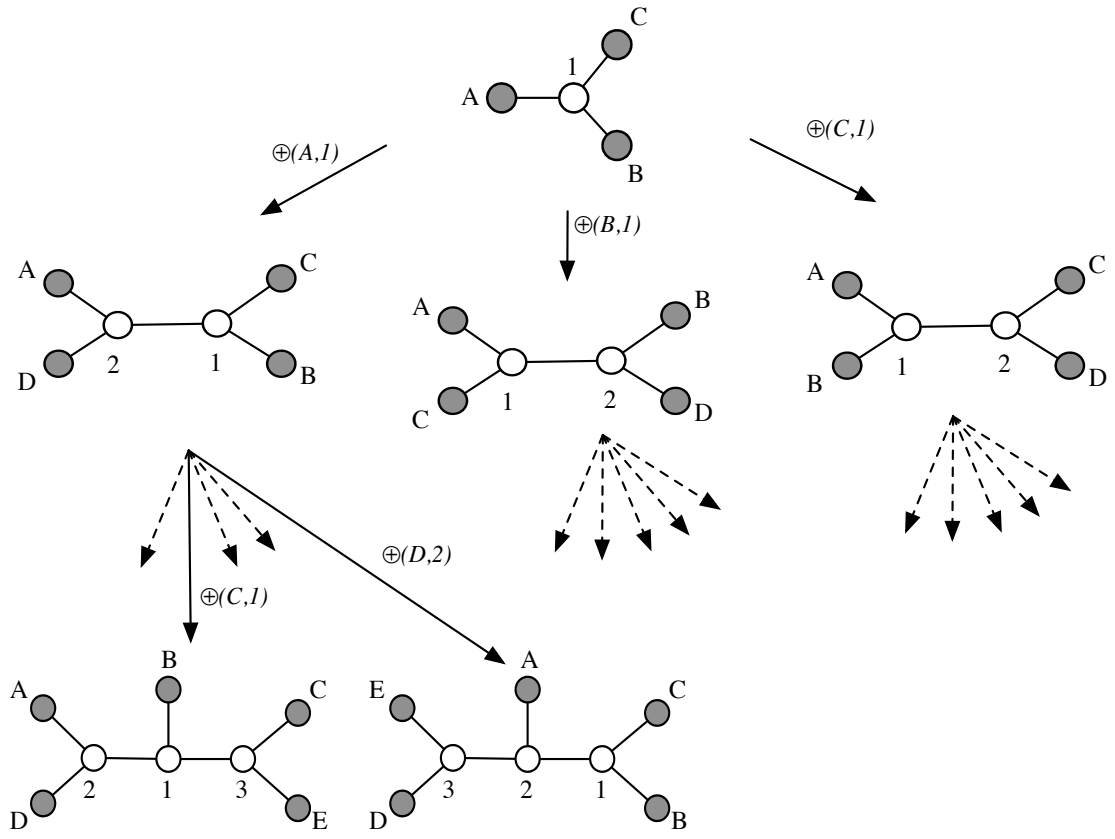

Figure 6: An example of some of the first subtrees explored by the implicit enumeration procedure.

branching is the average weight of many terms $\frac{1}{2}(d_{if} + d_{fj} - d_{ij})$. Hence, a very simple lower bound for the BMEP can be obtained by taking the sum of the minima of these terms. The reader interested in the issue is referred to Pardi (2009) for more details.

The results obtained from the analysis of the considered instances are summarized in Table 1, in which the instances are sorted and listed in function of their number of taxa. Specifically, Table 1 shows the numerical results obtained by PL4 and Algorithm 2 using Pardi's lower bound with respect to the running time (expressed in seconds) taken to solve a generic instance of the BMEP, the number of branches needed, and the gap (expressed in percentage) i.e., the difference between the optimal value found and the value of linear relaxation (or Pardi's lower bound) at the root node of the search tree, divided by the optimal value. The symbol > 3600 is used in the columns "Time" to highlight that the run relative to a specific instance took longer than 1 hour. In this circumstance the values relative to the columns "Branches" and "Gap" refer to the number of branches performed within 1 hour and the best upper bound found within 1 hour, respectively.

As general trend, Table 1 shows that PL4 is a tight formulation for the problem, being characterized

```
1 Implicit Enumeration Procedure;
  Input  : Γ: the set of taxa
  Output : A phylogeny T solution of the BMEP
2 set T = NULL;
3 set S = {Head(1, Γ), Head(2, Γ), Head(3, Γ)};
4 let Y(S) be the only subphylogeny that can be made with the three taxa in S;
5 T = Search(S, Y(S), T);
6 return T;
```

Algorithm 2: Implicit enumeration procedure.

everywhere by very small number of branches and gap values. However, the running time performances of PL4 result very poor causing, in many cases, the inability of the formulation to tackle instances containing more than a dozen of taxa within the limit time. This result may appear in contrast with the trend showed by the number of branches and gap values. Numerical experiments have shown that the cause of the slowness of PL4 is due to the simplex execution. Specifically, the simplex execution becomes extremely onerous in terms of computing time when valid inequalities and other constraints different from the Kraft, the unicity, and the third equalities are considered. Actually, if from one hand their presence increases the quality of the root relaxation, from the other hand such an increment is not sufficiently to compensate the overhead imposed to the simplex algorithm. In order to improve the runtime performances of PL4, in the next section we shall merge Algorithm 2 with PL4, developing a set of possible branching rules and lower bounds for the problem.

5 Improving the Performances of PL4

It is worth noting that the runtime taken by Algorithm 2 depends on how many subphylogenies are generated by the subroutine SEARCH() and on how efficiently this task is performed. In turn, the number of subphylogenies generated by the subroutine SEARCH() and the efficiency of the generation process depend on: (i) the quality of the bound provided within the subroutine BOUND(); (ii) the runtime of subroutine BOUND(); (iii) the order in which taxa are extracted from Γ by subroutine HEAD(); and (iv) the order in which the edges of each $Y(S)$ are branched. Aspects (i) and (ii) have a major impact on the performances of Algorithm 2, for this reason in the rest of the section we shall focus mainly on them.

Given a subphylogeny $Y(S)$, a possible strategy for designing a subroutine BOUND() having a good trade-off between the quality of the bound provided and time taken to compute it, consists of determining which values the topological distances τ_{ij} may assume in the phylogenies generated from $Y(S)$. To this end, consider a subset $S \subseteq \Gamma$, a subphylogeny $Y(S)$, and two taxa q and $t \in Y(S)$. Let σ_{qt} be the topological distance between taxa q and t in $Y(S)$. Then, the following three situations may occur:

1. Taxa i and $j \in S$. In this case
$$\sigma_{ij} \leq \tau_{ij} \leq \sigma_{ij} + |\Gamma \setminus S|. \tag{20}$$
These inequalities hold as, on each of the remaining $|\Gamma \setminus S|$ branchings needed to obtain a complete phylogeny for Γ, the distance between i and j increases by one only if the branched edge is on the paths between i and j.

2. Taxa $i \in S$ and $j \in \Gamma \setminus S$. In this case
$$2 \leq \tau_{ij} \leq \max_{q \in S}\{\sigma_{iq}\} + |\Gamma \setminus S|. \tag{21}$$
Specifically, $\tau_{ij} = 2$ is achieved when edge $(\hat{j}, j)$ is inserted on the edge $(\hat{i}, i)$ and the two edges are not branched any more. Differently, $\tau_{ij} = \max_{q \in S}\{\sigma_{iq}\} + |\Gamma \setminus S|$ is achieved when $(\hat{j}, j)$ is inserted on the edge $(\hat{q}^*, q^*)$, being $q^* = arg\max_{q \in S}\{\sigma_{iq}\}$, and the subsequent branchings are always performed on an edge belonging to the path between i and j.

3. Taxa i and $j \in \Gamma \setminus S$. In this case
$$2 \leq \tau_{ij} \leq \max_{t,q \in S}\{\sigma_{tq}\} + |\Gamma \setminus S|. \tag{22}$$
Specifically, $\tau_{ij} = 2$ is achieved when edge $(\hat{j}, j)$ is inserted on the edge $(\hat{i}, i)$ and the two edges are not branched any more. Differently, $\tau_{ij} = \max_{t,q \in S}\{\sigma_{tq}\} + |\Gamma \setminus S|$ is achieved when: $(\hat{i}, i)$ is inserted on the edge $(\hat{t}^*, t^*)$; $(\hat{j}, j)$ is inserted on the edge $(\hat{q}^*, q^*)$, being $(t^*, q^*) = arg\max_{t,q \in S}\{\sigma_{tq}\}$; and the subsequent branchings are always performed on an edge belonging to the path between i and j.

Note that, when $S = \Gamma$ the above bounds reduce trivially to the equality $\tau_{ij} = \sigma_{ij}$ for all $i \in \Gamma$ and $j \in \Gamma_i$. Hence, given a subphylogeny $Y(S)$, $S \subseteq \Gamma$, a lower bound on the length $\mathcal{L}(\hat{T})$ of the shortest phylogeny $\hat{T}$ generated from $Y(S)$ can be obtained by solving the linear relaxation of the following mixed integer programming problem:

Dataset	Number of taxa	Optimum	PL4+All Strengthening Valid Inequalities			Algorithm 2 with Pardi's lower bound		
			Time (sec.)	Branches	Gap (%)	Time (sec.)	Branches	Gap (%)
Primates12	10	124.9682159	9.64	**9**	**0.84**	**0.04**	1787	12.80
	11	332.8341675	127.65	**257**	**1.19**	**0.33**	13915	13.34
	12	802.5893555	155.32	**37**	**1.23**	**1.27**	47596	13.47
M17	10	105.1336746	258.5	**851**	**0.66**	**0.18**	9538	8.91
	11	261.2330627	2679.3	**3986**	**0.75**	**3.42**	133890	10.02
	12	541.632019	> 3600	n.a.	**0.71**	**8.75**	321885	10.63
	13	1181.597656	> 3600	n.a.	**0.99**	**60.18**	1604601	11.41
	14	2408.065674	> 3600	n.a.	**1.00**	**71.56**	2326884	11.59
	15	4998.294922	> 3600	n.a.	**0.99**	**140.14**	4462772	11.86
	16	10225.56055	> 3600	n.a.	**1.00**	**473.07**	14093125	12.84
	17	20788.11133	> 3600	n.a.	**1.02**	**710.4**	20537205	13.03
M18	10	190.3310699	130.76	**765**	**0.98**	**0.94**	45897	21.70
	11	396.9421692	580.27	**823**	**1.36**	**5.35**	229968	25.94
	12	805.9367065	> 3600	n.a.	**1.46**	**7.51**	310081	26.41
	13	1758.11145	> 3600	n.a.	**1.97**	**243.5**	7449682	31.12
	14	3599.677734	> 3600	n.a.	**2.43**	**1306.67**	36498904	33.03
	15	7746.217773	> 3600	n.a.	**2.32**	> 3600	91389391	35.90
	16	16006.95703	> 3600	n.a.	**2.64**	> 3600	80815699	37.64
	17	32589.09766	> 3600	n.a.	**3.22**	> 3600	75306196	41.28
	18	66423.09375	> 3600	n.a.	**3.12**	> 3600	64292767	42.93
SeedPlant25	10	91.45757294	86.34	**457**	**3.49**	**0.05**	2538	27.84
	11	206.2500763	591.76	**1327**	**3.39**	**0.16**	7000	27.92
	12	427.815918	457.68	**155**	**3.26**	**0.3**	12064	28.75
	13	943.774292	> 3600	n.a.	**3.25**	**2.74**	88757	30.54
	14	1929.218628	> 3600	n.a.	**3.07**	**3.68**	116417	30.49
	15	4085.763428	> 3600	n.a.	**2.79**	**8.05**	237472	30.19
	16	8353.457031	> 3600	n.a.	**3.81**	**117.21**	2693382	35.12
	17	17314.42969	> 3600	n.a.	**4.23**	**2113.09**	39231198	39.02
	18	36156.52734	> 3600	n.a.	**4.93**	> 3600	61292490	41.91
	19	75261.32031	> 3600	n.a.	**4.85**	> 3600	60305931	42.72
	20	167026.4375	> 3600	n.a.	**8.53**	> 3600	43270865	43.98
M43	10	105.0199661	71.65	**207**	**0.73**	**0.04**	2167	8.67
	11	209.282196	333.02	**631**	**1.25**	**0.06**	3292	9.14
	12	434.3260803	970.82	**577**	**1.01**	**0.78**	31398	10.72
	13	895.4237061	> 3600	n.a.	**1.24**	**1.7**	64081	11.04
	14	1808.969604	> 3600	n.a.	**1.27**	**5.58**	189992	11.94
	15	3965.234131	> 3600	n.a.	**0.99**	**92.35**	2151347	13.37
	16	8219.834961	> 3600	n.a.	**0.97**	**213.99**	5938295	13.90
	17	16798.75977	> 3600	n.a.	**1.20**	**411.07**	10708997	14.33
	18	35383.16016	> 3600	n.a.	**0.88**	**1580.28**	36626921	14.95
	19	71769.90625	> 3600	n.a.	**0.91**	**2116.88**	47647147	15.02
	20	157472.4219	> 3600	n.a.	**0.93**	> 3600	75184251	15.95
RbcL55	10	152.4825439	436.26	**1657**	**1.21**	**0.53**	26276	13.05
	11	328.6446838	771.97	**1578**	**1.11**	**1.54**	67269	13.20
	12	685.9721069	> 3600	n.a.	**1.44**	**5.8**	226756	14.02
	13	1502.870361	> 3600	n.a.	**1.71**	**292.55**	8961512	16.74
	14	3094.887939	> 3600	n.a.	**1.76**	**2751.94**	73837168	19.21
	15	6448.258789	> 3600	n.a.	**1.91**	> 3600	96810253	19.95
	16	13455.29297	> 3600	n.a.	**1.88**	> 3600	87850516	22.05
	17	27804.24609	> 3600	n.a.	**2.28**	> 3600	76687705	25.35
	18	56237.69141	> 3600	n.a.	**2.32**	> 3600	67207122	28.54
	19	115898.8203	> 3600	n.a.	**3.25**	> 3600	62230289	31.11
	20	235713.0938	> 3600	n.a.	**3.39**	> 3600	54387017	34.98
M62	10	163.8762207	21.86	**67**	**0.51**	**0.03**	1476	5.26
	11	350.4248047	138.67	**263**	**0.89**	**0.09**	3836	5.56
	12	753.8209839	618.25	**365**	**1.37**	**0.28**	8902	5.81
	13	1580.764282	2559.54	**589**	**1.51**	**1.1**	35144	6.35
	14	3345.563965	> 3600	n.a.	**1.61**	**6.09**	168507	6.76
	15	7161.077637	> 3600	n.a.	**1.76**	**14.46**	374111	6.67
	16	14980.69238	> 3600	n.a.	**1.70**	**29**	713416	6.67
	17	31293.08203	> 3600	n.a.	**1.67**	**163.28**	3466812	7.04
	18	66187.97656	> 3600	n.a.	**1.46**	> 3600	58204356	8.54
	19	146516.7031	> 3600	n.a.	**2.02**	> 3600	56001704	8.98
	20	298416.5938	> 3600	n.a.	**2.00**	> 3600	56996243	9.16
Rana64	10	41.22337723	87.45	**399**	**1.36**	**0.04**	2230	7.91
	11	87.16202545	818.27	**1675**	**4.19**	**0.15**	6877	9.18
	12	183.0419006	1306.58	801	**3.76**	**1.43**	54820	12.45
	13	382.6997375	> 3600	n.a.	**4.39**	**3.56**	131909	14.09
	14	773.8104248	> 3600	n.a.	**4.68**	**8.22**	285096	15.75
	15	1603.119629	> 3600	n.a.	**5.58**	**27.36**	652079	17.04
	16	3290.744873	> 3600	n.a.	**6.98**	**87.25**	2206517	19.29
	17	6745.447266	> 3600	n.a.	**7.22**	**109.29**	2664998	19.16
	18	15435.26758	> 3600	n.a.	**4.32**	**1032.09**	19093712	19.3
	19	36052.45312	> 3600	n.a.	**3.95**	> 3600	56570558	17.53
	20	81194.32031	> 3600	n.a.	**3.76**	> 3600	58504444	17.78
M82	10	53.17028427	1052.97	**3571**	**2.97**	**1.36**	52596	22.51
	11	106.4434509	> 3600	n.a.	**3.01**	**13.23**	462980	29.32
	12	225.8356628	> 3600	n.a.	**3.14**	**21.1**	717164	28.98
	13	543.1273804	> 3600	n.a.	**2.52**	**91.14**	2810302	27.39
	14	1238.321899	> 3600	n.a.	**3.00**	**1132.56**	30609154	29.07
	15	2515.808105	> 3600	n.a.	**4.15**	**2555.66**	65180385	30.16
	16	5098.458984	> 3600	n.a.	**3.61**	> 3600	82047430	31.94
	17	10483.7168	> 3600	n.a.	**3.47**	> 3600	84456435	36.15
	18	21625.17188	> 3600	n.a.	**4.72**	> 3600	52577113	39.49
	19	44545.28125	> 3600	n.a.	**6.19**	> 3600	68686094	41.64
	20	89202.41406	> 3600	n.a.	**6.48**	> 3600	55217421	43.48

Table 1: Numerical results obtained by PL4 and Pardi's implicit enumeration procedure (Pardi, 2009) on the analyzed datasets.

Formulation 2. *Reduced PL4 (RPL4)*

$$\min z_{lin}(Y(S)) = \sum_{i\in\Gamma}\sum_{j\in\Gamma_i} d_{ij}\left(\sum_{k\in L(i,j,Y(S))} 2^{-k}x_{ij}^k\right) \tag{23a}$$

$$s.t. \quad \sum_{k\in L(i,j,Y(S))} x_{ij}^k = 1 \qquad \forall\ i\neq j\in\Gamma \tag{23b}$$

$$x_{ij}^k = x_{ji}^k \quad \forall\ i\neq j\in\Gamma,\ k\in L(i,j,Y(S)) \tag{23c}$$

$$\sum_{j\in\Gamma_i}\sum_{k\in L(i,j,Y(S))} 2^{-k}x_{ij}^k = \frac{1}{2} \qquad \forall\ i\in\Gamma \tag{23d}$$

$$\sum_{k\in L(i,j,Y(S))} k2^{-k}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i} x_{ij}^k = (2n-3) \tag{23e}$$

$$x_{ij}^k\in\{0,1\} \qquad \forall\ i,j\in\Gamma, k\in L(i,j,Y(S)), \tag{23f}$$

where $L(i,j,Y(S))$ are subsets of L such that

$$L(i,j,Y(S)) = \begin{cases} \{k\in L : \sigma_{ij}\leq k\leq \sigma_{ij}+|\Gamma\setminus S|\} & \text{if } i \text{ and } j\in S,\\ \{k\in L : 2\leq k\leq \max_{q\in S}\{\sigma_{iq}\}+|\Gamma\setminus S|\} & \text{if } i\in S \text{ and } j\in\Gamma\setminus S,\\ \{k\in L : 2\leq k\leq \max_{t,q\in S}\{\sigma_{tq}\}+|\Gamma\setminus S|\} & \text{if both } i \text{ and } j\in\Gamma\setminus S. \end{cases}$$

RPL4 derives from PL4 by elimination of all constraints but the first, the symmetry equalities, the Kraft equalities, and third equality. RPL4 does not include any strengthening valid inequality. Actually, in preliminary numerical experiments we have observed that the inclusion of the strengthening valid inequalities imposes a computational overload which is not compensated by the increment of the quality of the lower bound so obtained. However, we stress the fact that the above argumentation may be not valid for large instances of the BMEP. Actually, in these cases the introduction of strengthening valid inequalities may turn necessary to obtain bounds that prune a number of phylogenies sufficiently high to maintain computationally acceptable the runtime of the implicit enumeration procedure.

It is worth noting that solving RLP4 at each node of the branch-and-bound tree may result very time consuming due to the need of setting appropriately constraints (23f), a task that requires alone a computational complexity $O(n^3)$. A possible strategy to speed-up computations consists of considering the lagrangian relaxation of RPL4, i.e.,:

Formulation 3. *Lagrangian RPL4 (LRPL4)*

$$\min z_{lag}(Y(S),\mu,\lambda) = \sum_{i<j\in\Gamma}\sum_{k\in L(i,j,Y(S))} (2d_{ij}-\mu_i-\mu_j-2k\lambda)2^{-k}x_{ij}^k + (2n-3)\lambda + \sum_{i\in\Gamma}\frac{\mu_i}{2} \tag{24a}$$

$$s.t. \quad \sum_{k\in L(i,j,Y(S))} x_{ij}^k = 1 \qquad \forall\ i<j\in\Gamma \tag{24b}$$

$$x_{ij}^k\in\{0,1\} \qquad \forall\ i<j\in\Gamma,\ k\in L(i,j,Y(S)) \tag{24c}$$

where $\mu=\{\mu_i : i\in\Gamma\}$ and λ are the lagrangian multipliers of constraints (23d) and (23e). Formulation 3 is obtained from (23) by relaxing constraints (23d) and (23e) and substituting x_{ij}^k with x_{ji}^k when $i>j$ as required by constraints (23c). Note that, if we disregard the constant value $\sum_{i\in\Gamma}\frac{\mu_i}{2}+(2n-3)\lambda$, problem (24) can be decomposed in a set of smaller problems, such as

$$\min z_{ij,lag}(Y(S),\mu,\lambda) = \sum_{k\in L(i,j,Y(S))} (2d_{ij}-\mu_i-\mu_j-2k\lambda)2^{-k}x_{ij}^k \tag{25a}$$

$$s.t. \quad \sum_{k\in L(i,j,Y(S))} x_{ij}^k = 1 \tag{25b}$$

$$x_{ij}^k\in\{0,1\} \qquad \forall\ k\in L(i,j,Y(S)), \tag{25c}$$

for all $i,j\in\Gamma$ such that $i<j$. Since the solution to each problem (25) can be obtained analytically, computing the value $z_{lag}^*(Y(S),\mu,\lambda)$ results much faster than determining the value $z_{lin}^*(Y(S))$. This insight

```
BOUND(S, Y(S), T);
  Input  : S: a subset of taxa
           Y(S): a subphylogeny
           T: current optimal phylogeny
  Output : A boolean value
if |S| = 3 then
    if z*_lin(Y(S)) >= L(T) then
        return FALSE;
    else
        return TRUE;
else
    if z*_lag(Y(S), μ, λ) >= L(T) then
        return FALSE;
    else
        if z*_lin(Y(S)) >= L(T) then
            return FALSE;
        else
            return TRUE;
```

Algorithm 3: Subroutine BOUND().

suggests an alternative way to implement the subroutine BOUND(), which can be outlined as follows. When $|S| = 3$ BOUND() computes the value $z^*_{lin}(Y(S))$, optimal solution of RPL4. Subsequently, for all S such that $|S| > 3$ BOUND() computes the value $z^*_{lag}(Y(S))$, solution of LRPL4. If $z^*_{lag}(Y(S), \mu, \lambda) > \mathcal{L}(\hat{T})$, BOUND() returns FALSE, else, BOUND() computes $z^*_{lin}(Y(S))$ and if $z^*_{lin}(Y(S)) > \mathcal{L}(\hat{T})$ FALSE is returned, otherwise TRUE is returned. The whole procedure is formally described in Algorithm 3.

Subroutine BOUND() computes the value $z^*_{lag}(Y(S), \mu, \lambda)$ only if the value $z^*_{lin}(Y(S \setminus \{i\}))$ has been previously computed for some $i \in S$. We stress this point as, to save time, in computing the value $z^*_{lag}(Y(S))$ subroutine BOUND() does not determine, and hence does not use, the optimal dual values for $\mu = \{\mu_i : i \in \Gamma\}$ and λ. Subroutine BOUND() simply sets the elements of μ (respectively λ) equal to the shadow prices of constraints (23d) (respectively (23e)) obtained from the last time that the value $z^*_{lag}(Y(S \setminus \{i\}))$ has been computed for some $i \in S$. In preliminary experiments we have observed that the values $z^*_{lag}(Y(S), \mu, \lambda)$ and $z^*_{lin}(Y(S))$ differ very little, usually less than 1%. For this reason, when $z^*_{lag}(Y(S), \mu, \lambda) < \mathcal{L}(\hat{T})$ for more than 1%, we allow subroutine BOUND() to skip once the computation of the value $z^*_{lin}(Y(S))$ if the value $z^*_{lin}(Y(S \setminus \{i\}))$ has been previously computed for some $i \in S$. The rationale at the core of this choice is given by the fact that there is little hope that the value $z^*_{lin}(Y(S))$ be greater than $\mathcal{L}(\hat{T})$. In this case, as we need the shadow prices of constraints (23d) and (23e), subroutine BOUND() assumes that such values are equal to the corresponding values obtained when computing the value $z^*_{lin}(Y(S \setminus \{i\}))$.

In the next section we shall present the results obtained by embodying the new subroutine BOUND() inside Algorithm 2.

6 Numerical Results

Tables 2, 3, and 4 summarize the results obtained by all algorithms described in the article when solving the previously described instances. The algorithms are implemented in ANSI C++ and together with the analyzed instances can be found in the online supplement for codes and data.

Table 2 summarizes the numerical results with respect to the running time (expressed in seconds) taken to solve a generic instance of the BMEP. Specifically, Table 2 shows in the third column the running time of PL4 with all its strengthening valid inequalities; in the fourth column the running time of Algorithm 2 when using Pardi's bound; and finally in the fifth and sixth columns the running times of Algorithm 2 when using Pardi's bound and Algorithm 3, respectively, under a specific *taxa extraction order*, i.e., the order in which taxa are extracted from Γ by subroutine HEAD(). In fact, as observed in Pardi (2009), the taxa extraction order can affect the performances of Algorithm 2 in a way that is still not completely clear. In preliminary

Time (sec.)

Dataset	Number of taxa	PL4+ All Strengthening Valid Inequalities	Alg. 2 + Pardi's lower bound (No Leaf Order)	Alg. 2 + Pardi's lower bound (Hamiltonian Leaf Order)	Alg. 2 + Alg. 3 (Hamiltonian Leaf Order)
Primates12	10	9.64	0.04	**0.04**	0.12
	11	127.65	0.33	**0.11**	0.25
	12	155.32	1.27	**0.24**	0.38
M17	10	258.50	0.18	**0.10**	0.12
	11	2679.30	3.42	1.88	**0.91**
	12	> 3600	8.75	27.28	**1.70**
	13	> 3600	60.18	64.63	**9.57**
	14	> 3600	71.56	156.98	**12.66**
	15	> 3600	140.14	633.82	**26.00**
	16	> 3600	473.07	245.24	**4.65**
	17	> 3600	710.40	461.33	**8.14**
M18	10	130.76	0.94	0.59	**0.16**
	11	580.27	5.35	9.30	**1.48**
	12	> 3600	7.51	44.36	**4.76**
	13	> 3600	243.50	137.70	**19.71**
	14	> 3600	1306.67	594.57	**75.23**
	15	> 3600	> 3600	> 3600	**196.89**
	16	> 3600	> 3600	> 3600	**215.41**
	17	> 3600	> 3600	> 3600	**589.32**
	18	> 3600	> 3600	> 3600	**816.29**
SeedPlant25	10	86.34	0.05	**0.03**	0.12
	11	591.76	**0.16**	1.06	1.32
	12	457.68	0.30	6.16	**1.22**
	13	> 3600	2.74	**0.94**	2.05
	14	> 3600	**3.68**	298.75	11.76
	15	> 3600	**8.05**	2241.30	33.09
	16	> 3600	117.21	> 3600	**560.51**
	17	> 3600	2113.09	> 3600	> 3600
	18	> 3600	> 3600	3483.18	**779.35**
	19	> 3600	> 3600	> 3600	**2472.23**
	20	> 3600	> 3600	> 3600	> 3600
M43	10	71.65	**0.04**	0.31	0.22
	11	333.02	**0.06**	0.89	0.28
	12	970.82	0.78	2.59	**0.38**
	13	> 3600	1.70	7.92	**0.86**
	14	> 3600	5.58	25.92	**1.31**
	15	> 3600	92.35	210.49	**3.42**
	16	> 3600	213.99	742.86	**186.74**
	17	> 3600	411.07	1833.17	**373.16**
	18	> 3600	1580.28	> 3600	**222.17**
	19	> 3600	2116.88	> 3600	**306.50**
	20	> 3600	> 3600	> 3600	**110.50**
RbcL55	10	436.26	0.53	1.73	**0.45**
	11	771.97	1.54	2.37	**0.44**
	12	> 3600	**5.80**	11.27	7.77
	13	> 3600	292.55	42.65	**4.93**
	14	> 3600	2751.94	166.07	**13.22**
	15	> 3600	> 3600	514.02	**28.22**
	16	> 3600	> 3600	> 3600	**401.18**
	17	> 3600	> 3600	> 3600	**1119.52**
	18	> 3600	> 3600	> 3600	**3072.78**
	19	> 3600	> 3600	> 3600	> 3600
	20	> 3600	> 3600	> 3600	> 3600
M62	10	21.86	**0.03**	0.07	0.10
	11	138.67	**0.09**	0.20	0.14
	12	618.25	**0.28**	2.79	4.79
	13	2559.54	1.10	2.22	**0.38**
	14	> 3600	**6.09**	65.49	84.93
	15	> 3600	14.46	280.00	**8.95**
	16	> 3600	29.00	473.60	**10.26**
	17	> 3600	163.28	1745.37	**30.36**
	18	> 3600	> 3600	> 3600	> 3600
	19	> 3600	> 3600	> 3600	**132.07**
	20	> 3600	> 3600	> 3600	> 3600
Rana64	10	87.45	0.04	**0.02**	0.11
	11	818.27	**0.15**	0.17	2.15
	12	1306.58	1.43	**0.14**	0.18
	13	> 3600	3.56	**3.46**	3.58
	14	> 3600	8.22	7.37	**4.47**
	15	> 3600	27.36	29.55	**5.41**
	16	> 3600	87.25	326.87	**13.64**
	17	> 3600	**109.29**	3184.38	2760.74
	18	> 3600	**1032.09**	> 3600	1292.51
	19	> 3600	> 3600	> 3600	**45.09**
	20	> 3600	> 3600	> 3600	**2512.04**
M82	10	1052.97	1.36	2.69	**1.21**
	11	> 3600	13.23	21.56	**2.28**
	12	> 3600	21.10	126.09	**6.29**
	13	> 3600	91.14	546.63	**12.68**
	14	> 3600	1132.56	> 3600	**145.49**
	15	> 3600	2555.66	> 3600	**536.83**
	16	> 3600	> 3600	> 3600	**956.62**
	17	> 3600	> 3600	> 3600	**3367.62**
	18	> 3600	> 3600	> 3600	> 3600
	19	> 3600	> 3600	> 3600	> 3600
	20	> 3600	> 3600	> 3600	> 3600

Table 2: Overview, with respect to the running time, of the numerical results obtained from the analysis of the considered datasets.

Branches

Dataset	Number of taxa	PL4+ All Strengthening Valid Inequalities	Alg. 2 + Pardi's lower bound (No Leaf Order)	Alg. 2 + Pardi's lower bound (Hamiltonian Leaf Order)	Alg. 2 + Alg. 3 (Hamiltonian Leaf Order)
Primates12	10	**9**	1787	2233	522
	11	**257**	13915	4501	781
	12	**37**	47596	8647	1146
M17	10	**851**	9538	3062	853
	11	**3986**	133890	53549	5161
	12	n.a.	321885	656750	**9188**
	13	n.a.	1604601	1469337	**47351**
	14	n.a.	2326884	3182059	**54730**
	15	n.a.	4462772	11653303	**98682**
	16	n.a.	14093125	4786400	**14781**
	17	n.a.	20537205	7943220	**23420**
M18	10	**765**	45897	20234	1020
	11	**823**	229968	260789	9565
	12	n.a.	310081	1092911	**26986**
	13	n.a.	7449682	3244055	**97650**
	14	n.a.	36498904	12209798	**319106**
	15	n.a.	91389391	67354519	**738592**
	16	n.a.	80815699	57035287	**806433**
	17	n.a.	75306196	54504361	**2025672**
	18	n.a.	64292767	53568165	**2214344**
SeedPlant25	10	**457**	2538	1420	792
	11	**1327**	7000	46582	8438
	12	**155**	12064	243425	7035
	13	n.a.	88757	33376	**10505**
	14	n.a.	116417	9223929	**53272**
	15	n.a.	237472	60183365	**130140**
	16	n.a.	2693382	86194772	**1999243**
	17	n.a.	39231198	78183397	**11536834**
	18	n.a.	61292490	68820012	**2270736**
	19	n.a.	60305931	62839170	**6060338**
	20	n.a.	43270865	58993984	**7791673**
M43	10	**207**	2167	9719	1299
	11	**631**	3292	24775	1639
	12	**577**	31398	66107	1923
	13	n.a.	64081	180883	**3797**
	14	n.a.	189992	527059	**5222**
	15	n.a.	2151347	3831367	**12448**
	16	n.a.	5938295	12865319	**553470**
	17	n.a.	10708997	29277538	**978552**
	18	n.a.	36626921	50944050	**477234**
	19	n.a.	47647147	48029558	**559416**
	20	n.a.	75184251	42814507	**232440**
RbcL55	10	**1657**	26276	73885	3118
	11	**1578**	67269	87726	2739
	12	n.a.	226756	419058	**42253**
	13	n.a.	8961512	1185697	**23335**
	14	n.a.	73837168	4358506	**51360**
	15	n.a.	96810253	12422524	**101269**
	16	n.a.	87850516	65735578	**1402870**
	17	n.a.	76687705	59550075	**2971315**
	18	n.a.	67207122	57206496	**6828465**
	19	n.a.	62230289	44885393	**10922539**
	20	n.a.	54387017	45258284	**5944082**
M62	10	**67**	1476	2377	530
	11	**263**	3836	5677	577
	12	**365**	8902	76052	22778
	13	**589**	35144	51677	1446
	14	n.a.	168507	1393416	**346476**
	15	n.a.	374111	4787355	**32029**
	16	n.a.	713416	7350931	**33092**
	17	n.a.	3466812	24309519	**88322**
	18	n.a.	58204356	47005536	**10089505**
	19	n.a.	56001704	44077944	**313505**
	20	n.a.	56996243	41807194	**7693580**
Rana64	10	399	2230	567	**383**
	11	**1675**	6877	7399	7477
	12	801	54820	4855	**578**
	13	n.a.	131909	116150	**7858**
	14	n.a.	285096	217075	**8484**
	15	n.a.	652079	702512	**8752**
	16	n.a.	2206517	6116978	**24179**
	17	n.a.	2664998	77091081	**3848622**
	18	n.a.	19093712	77476337	**1496640**
	19	n.a.	56570558	52425796	**60328**
	20	n.a.	58504444	71502713	**3032704**
M82	10	**3571**	52596	83722	9877
	11	n.a.	462980	588638	**15679**
	12	n.a.	717164	3036651	**38083**
	13	n.a.	2810302	11424629	**66662**
	14	n.a.	30609154	66271501	**720757**
	15	n.a.	65180385	58637723	**2290648**
	16	n.a.	82047430	56646829	**3698560**
	17	n.a.	84456435	53120128	**11953699**
	18	n.a.	52577113	46559698	**10092896**
	19	n.a.	68686094	43368382	**9759937**
	20	n.a.	55217421	40689604	**8106293**

Table 3: Overview, with respect to the number of branches performed, of the numerical results obtained from the analysis of the considered datasets.

		Gap (%)			
Dataset	Number of taxa	PL4+All Strengthening Valid Inequalities	Alg. 2 + Pardi's lower bound (No Leaf Order)	Alg. 2 + Pardi's lower bound (Hamiltonian Leaf Order)	Alg. 2 + Alg. 3 (Hamiltonian Leaf Order)
Primates12	10	**0.84**	12.80	13.43	2.57
	11	**1.19**	13.34	14.30	2.59
	12	**1.23**	13.47	14.27	2.47
M17	10	**0.66**	8.91	7.52	1.05
	11	**0.75**	10.02	7.88	1.18
	12	**0.71**	10.63	10.64	1.21
	13	**0.99**	11.41	10.41	1.65
	14	**1.00**	11.59	10.39	1.65
	15	**0.99**	11.86	11.19	1.65
	16	**1.00**	12.84	19.38	1.70
	17	**1.02**	13.03	19.16	1.64
M18	10	**0.98**	21.70	22.74	2.11
	11	**1.36**	25.94	17.20	2.57
	12	**1.46**	26.41	19.43	2.71
	13	**1.97**	31.12	25.79	3.27
	14	**2.43**	33.03	27.82	3.62
	15	**2.32**	35.90	31.33	3.47
	16	**2.64**	37.64	31.28	3.51
	17	**3.22**	41.28	32.38	3.53
	18	**3.12**	42.93	39.98	3.55
SeedPlant25	10	**3.49**	27.84	42.92	5.51
	11	**3.39**	27.92	33.54	5.10
	12	**3.26**	28.75	34.63	4.87
	13	**3.25**	30.54	41.20	5.30
	14	**3.07**	30.49	39.08	5.06
	15	**2.79**	30.19	38.47	4.66
	16	**3.81**	35.12	42.53	5.76
	17	**4.23**	39.02	44.45	6.21
	18	**4.93**	41.91	44.51	6.92
	19	**4.85**	42.72	45.63	6.97
	20	**8.53**	43.98	49.83	9.11
M43	10	**0.73**	8.67	9.75	1.29
	11	**1.25**	9.14	9.99	2.39
	12	**1.01**	10.72	12.04	1.79
	13	**1.24**	11.04	12.14	1.74
	14	**1.27**	11.94	13.03	1.75
	15	**0.99**	13.37	14.34	1.53
	16	**0.97**	13.90	15.60	1.51
	17	**1.20**	14.33	16.24	1.82
	18	**0.88**	14.95	16.90	1.44
	19	**0.91**	15.02	17.38	1.43
	20	**0.93**	15.95	20.96	1.48
RbcL55	10	**1.21**	13.05	11.74	1.79
	11	**1.11**	13.20	11.26	1.61
	12	**1.44**	14.02	14.53	2.20
	13	**1.71**	16.74	13.89	2.33
	14	**1.76**	19.21	15.54	2.27
	15	**1.91**	19.95	15.80	2.46
	16	**1.88**	22.05	23.67	2.50
	17	**2.28**	25.35	21.93	2.94
	18	**2.32**	28.54	22.75	2.90
	19	**3.25**	31.11	29.01	3.67
	20	**3.39**	34.98	26.85	4.63
M62	10	**0.51**	5.26	4.32	1.07
	11	**0.89**	5.56	5.79	1.50
	12	**1.37**	5.81	4.85	2.06
	13	**1.51**	6.35	6.60	2.21
	14	**1.61**	6.76	6.40	2.35
	15	**1.76**	6.67	7.30	2.64
	16	**1.70**	6.67	7.19	2.54
	17	**1.67**	7.04	7.44	2.50
	18	**1.46**	8.54	8.44	2.52
	19	**2.02**	8.98	12.54	2.19
	20	**2.00**	9.16	9.64	2.81
Rana64	10	**1.36**	7.91	9.44	5.74
	11	**4.19**	9.18	9.17	4.94
	12	**3.76**	12.45	15.10	4.64
	13	**4.39**	14.09	13.77	5.39
	14	**4.68**	15.75	14.71	5.60
	15	**5.58**	17.04	16.03	6.47
	16	**6.98**	19.29	20.03	7.78
	17	**7.22**	19.16	20.03	8.01
	18	**4.32**	19.30	20.53	5.07
	19	**3.95**	17.53	19.48	4.48
	20	**3.76**	17.78	19.64	4.74
M82	10	**2.97**	22.51	20.97	4.39
	11	**3.01**	29.32	27.78	4.52
	12	**3.14**	28.98	33.62	4.43
	13	**2.52**	27.39	37.71	3.92
	14	**3.00**	29.07	28.75	4.51
	15	**4.15**	30.16	31.05	6.26
	16	**3.61**	31.94	32.03	5.13
	17	**3.47**	36.15	42.67	4.76
	18	**4.72**	39.49	35.50	5.95
	19	**6.19**	41.64	42.44	6.36
	20	**6.48**	43.48	47.01	8.21

Table 4: Overview, with respect to the number of gap, of the numerical results obtained from the analysis of the considered datasets.

experiments we tested different taxa extraction orders, namely: the random; *the ascending greedy* consisting of computing the number $c_i = \sum_{j \in \Gamma_i} d_{ij}$, for all $i \in \Gamma$, and by sorting the vector $\mathbf{c} = \{c_i\}$ in ascending order; *the descending greedy*, consisting of computing the number $c_i = \sum_{j \in \Gamma_i} d_{ij}$, for all $i \in \Gamma$, and by sorting the vector $\mathbf{c} = \{c_i\}$ in descending order; and the *Hamiltonian Leaf Order* provided by the solution of the shortest hamiltonian circuit on the instance represented by the input distance matrix. In the tables we just present the Hamiltonian Leaf Order which was the one characterized by better average performances. It is worth noting that the problem of finding the shortest hamiltonian circuit can be efficiently tackled by using Concorde (Applegate et al., 2001), a solver for the Traveling Salesman Problem (TSP) (Garey and Johnson, 2003) and other related network optimization problems. Concorde is written in ANSI C and is able to solve instances of the TSP having thousands cities. In our experiments Concorde took a negligible time (typically milliseconds) for solving the considered instances for this reason we omitted its running time in the table.

Table 3 summarizes the numerical results with respect to the number of branches needed to solve a specific instance. As in Section 4, if the corresponding running time was longer than 1 hour, the value denotes the number of branches performed within 3600 seconds. Finally, Table 4 summarizes the numerical results with respect to gap shown in percentage terms. We recall that if for a specific instance the corresponding running time was longer than 1 hour, the best upper bound found within 3600 seconds is used to compute the gap.

As general trend, Tables 2, 3, and 4 show that the combination of Algorithm 2 and Algorithm 3 provides in average good performances. Specifically, the algorithm results the fastest when analyzing datasets M18 and M82 and predominantly the fastest when analyzing datasets M17, M43, and RbcL55. Moreover, the algorithm is able to tackle instances that are unsolved by the remaining solution approaches (see e.g., RbcL55 for 18 taxa and M82 for 17 taxa). The performances of the Algorithm 2 with Algorithm 3 decrease when dealing with instances characterized by small number of taxa (usually, less then a dozen). This phenomenon is mainly due to the overhead introduced by the runtime generation of RPL4 and tends to disappear when tackling bigger instances.

The major impact of the leaf order on the solution time becomes evident when considering datasets such as SeedPlant25, M62, and Rana64, in which the solution time sensibly changes. In our experiments we observed that the leaf order influences in general all the exact solution approaches based on Algorithm 2, independently of the type of bound used. However, we observed major influence of the leaf order on Pardi's lower bound with respect to Algorithm 3. For example, in Tables 3 and 4 it is possible to see that the number of branches and the gap values for Pardi's lower bound may change drastically when tackling the same instance under different leaf order (see e.g., SeedPlant25 for $n = 10$, RbcL55 for $n = 20$, M82 for $n = 13$).

Finally, it is worth noting that the bound provided by Algorithm 3, in average about 3.61%, is slightly worse than the one provided by PL4, in average about 2.54%. This fact confirms the major impact that the properties of the topological distances have on the problem. Our believe is that a further deeper investigation of those properties could suggest new directions on the development of efficient exact approaches to solution of this problem.

7 Conclusion

The Balanced Minimum Evolution Problem (BMEP) is a recent version of the Phylogenetic Estimation Problem (PEP) firstly introduced by Pauplin (2000). Given a set Γ of n taxa and the corresponding matrix $\mathbf{D}$ of evolutionary distances, the BMEP consists of finding a phylogeny for Γ having minimum length (Catanzaro, 2009). The BMEP is based on the minimum evolution criterion of phylogenetic estimation which states that if the evolutionary distances were unbiased estimates of the true evolutionary distances (i.e., the distances that one would obtain if all the molecular data from the analyzed taxa were available), then the true phylogeny would have an expected length shorter than any other possible phylogeny compatible with $\mathbf{D}$. Interestingly, the minimum evolution criterion does not asses that molecular evolution follows minimum paths, but states, according to classical evolutionary theory, that a minimum length phylogeny may properly approximate the real phylogeny of well-conserved molecular data i.e., data whose basic biochemical functions undergone small change throughout the evolution of the observed taxa (Beyer et al., 1974). Since the selective forces acting on taxa may not be constant over time, evolution proceeds by small rather than smallest change (Beyer et al., 1974; Waterman et al., 1977). Thus, a minimum length phylogeny provides a lower bound on the overall number of mutation events that could have occurred along evolution of the observed taxa.

In this article we presented a possible exact approach to solution of the BMEP based on mathematical programming. Specifically, we investigated the properties of the topological distances in order to provide a

valid polynomial size formulation for the problem. Moreover, we developed families of strengthening valid inequalities, branching rules, and lower bounds aiming at improving the performances of the formulation. Our results give perspective on the mathematics of the BMEP and suggest new directions on the development of future efficient exact approaches to solution of this problem.

Acknowledgement

The first author acknowledges support from the Belgian National Fund for Scientific Research (F.N.R.S.), of which he is "Chargé de Recherches". Both the first and the second authors also acknowledge support from Communauté Française de Belgique — Actions de Recherche Concertées (ARC). The authors thank Dr. Fabio Pardi for helpful discussions and Dr. Rosa Maria Lo Presti for the datasets provided. Finally, the authors thank the area editor, the associate editor, and the anonymous reviewers for their valuable comments on the previous version of the manuscript.

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.

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.

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

P. Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory*, 17:48–50, 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*. Springer-Verlag, New York (In press), 2010.

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

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. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. *Networks*, 53(2):126–140, 2009.

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

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

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

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

S. Fiorini and G. Joret. The balanced minimum evolution problem is hard. Technical report, Département de Mathématique - Université Libre de Bruxelles (U.L.B.), 2010.

M. Fischetti, G. Lancia, and P. Serafini. Exact algorithms for minimum routing cost trees. *Networks*, 39(3): 161–173, 2002.

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

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

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. 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.

J. F. Maurras, T. H. Nguyen, and V. H. Nguyen. On the Convex Hull of Huffman Trees. *Electronic Notes in Discrete Mathematics*, 36(36):1009–1016, 2010.

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.

F. Pardi. *Algorithms on Phylogenetic Trees*. PhD thesis, University of Cambridge, UK, 2009.

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.

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.

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

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