A NOTE ON THE APPROXIMABILITY OF THE BALANCED MINIMUM EVOLUTION PROBLEM

Daniele Catanzaro, Raffaele Pesenti, Francesco Pisanu

REPRINT | 3353

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

A Note on the Approximability of the Balanced Minimum Evolution Problem

Daniele Catanzaro* Raffaele Pesenti† Francesco Pisanu‡

Abstract

The *Balanced Minimum Evolution Problem (BMEP)* is a highly nonlinear $\mathcal{NP}$-hard optimization problem in molecular phylogenetics that has attracted significant attention from the bioinformatics and mathematical programming communities. We investigate conditions under which its practical instances become efficiently approximable. We show that when all pairwise distances are positive and bounded, the problem admits a polynomial-time approximation algorithm with a performance guarantee linked to the interval width. We also characterize polynomially solvable instances, and derive tight bounds on the optimal solution.

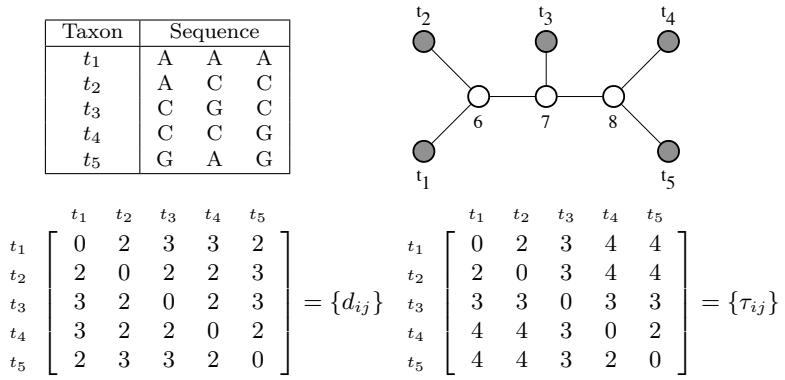

Figure 1: Above: a simple example of a set Γ of 5 taxa and the associated hypothetical DNA sequences (left), and a possible phylogeny of Γ encoded as a UBT (right). Below: an example of a *distance matrix* associated with Γ (left), obtained by assuming the Hamming distance as a measure of dissimilarity between pairs of sequences, along with the corresponding path-length matrix τ encoding the above UBT (right).

1 Introduction

Consider a set $\Gamma = \{1, 2, \ldots, n\}$ of $n \geq 3$ items and let $\mathbf{D} = \{d_{ij}\}$ be a *dissimilarity* matrix on Γ, i.e., a symmetric matrix with all diagonal entries equal to 0 and nonnegative off-diagonal entries. An *Unrooted Binary Tree* (UBT) T on Γ is a tree having Γ as its leaf set and internal vertices of degree 3 [10]. For example, Figure 1 shows a possible UBT for a set Γ of five items. Let Θ_n denote the set of all UBTs on Γ, and for a given UBT $T \in \Theta_n$, let τ_{ij} denote the number of edges on the unique path in T between items i and j. Then, the *Balanced Minimum*

*CORE, Université Catholique de Louvain, Belgium. Email: daniele.catanzaro@uclouvain.be

†Venice School of Management, Università Ca' Foscari di Venezia, Italy. Email: pesenti@unive.it

‡CORE, Université Catholique de Louvain, Belgium. Corresponding author. Email: francesco.pisanu@uclouvain.be

Evolution Problem (BMEP) seeks a UBT $T^* \in \Theta_n$ that solves the following nonlinear network design problem:

$$\min_{T \in \Theta_n} L_{\mathbf{D}}(T) = \sum_{i \in \Gamma} \sum_{\substack{j \in \Gamma \\ j \neq i}} \frac{d_{ij}}{2^{\tau_{ij}}}. \tag{1}$$

The BMEP is a combinatorial optimization problem arising from Pauplin's work [22] on the estimation of evolutionary trees in molecular phylogenetics [9]. In this setting, Γ represents a collection of distinct aligned molecular sequences extracted from target species (commonly referred to as *taxa*), such as DNA, RNA, codon sequences, or entire genomes. The matrix $\mathbf{D}$ encodes estimated evolutionary distances between pairs of molecular sequences from different taxa [3, 4, 5, 13], while a UBT on Γ encodes a candidate *phylogeny*, i.e., a tree graph describing the evolutionary relationships of taxa based on their molecular sequences. The BMEP seeks to identify, among an exponential number of candidate phylogenies, the one that minimizes the cross-entropy associated with the molecular sequences of taxa, thereby selecting the phylogeny with the highest likelihood of having occurred over the course of their evolution [7].

Beyond its biological motivation, recently reviewed in [9], the BMEP has also become an object of independent interest in algorithmic and combinatorial optimization. Research efforts have focused on its statistical foundations [12, 17, 18], its combinatorial structure [6, 8, 11, 15, 16, 19, 23], and algorithmic solution approaches, including both exact methods [1, 6, 21] and heuristics [14, 17, 21]. From a complexity-theoretic perspective, Fiorini and Joret [14] proved that the BMEP is $\mathcal{NP}$-hard, via a reduction from the 3-coloring problem, and inapproximable within any factor c^n, for some constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$. The authors also showed that when the dissimilarity matrix $\mathbf{D}$ is metric (i.e., when its entries satisfy the triangle inequality), the problem admits a polynomial-time approximation algorithm with an approximation factor of two. These hardness and inapproximability results rely on reductions that construct instances of the BMEP in which $\mathbf{D}$ contains zero-valued off-diagonal entries. Such cases, however, are rarely encountered in practical phylogenetic analyses, where taxa are typically assumed to be all distinct (leading so to dissimilarity matrices having positive off-diagonal entries). Motivated by this observation, we investigate here the broader conditions under which practical instances of the BMEP become efficiently approximable. We show that when all pairwise distances are positive and confined within a prescribed interval, the problem admits a polynomial-time approximation algorithm, whose performance guarantee is explicitly linked to the width of this interval. In addition, we identify specific subclasses of instances that are solvable in polynomial time and establish tight lower and upper bounds on the optimal solution to the problem. Altogether, these results provide new insight into the approximability of the BMEP in settings that are directly relevant to phylogenetic inference. Before proceeding, we introduce in the next section some notation and definitions that will prove useful throughout the article, and we briefly recall some prior results from the literature that will be instrumental in the subsequent sections.

2 Notation, Definitions, and Prior Work

Consider a UBT $T \in \Theta_n$. We say that two distinct leaves i and j of T form a *cherry* if the length of the unique path in T connecting i to j is equal to 2, i.e., $\tau_{ij} = 2$. For example, the leaves t_1 and t_2 in the UBT in Figure 1 form a cherry. We say that $T \in \Theta_n$ is a *caterpillar* if, for some pair of distinct leaves $i, j \in \Gamma$, $\tau_{ij} = n - 1$ [10]. Stated otherwise, a caterpillar is a particular UBT characterized by the existence of precisely two cherries.

For a given tree $T = (V, E)$ and any vertex $v \in V$, we denote by $N(v)$ the set of *neighbors of* v, i.e., the set of vertices that are adjacent to v in T.

Consider a dissimilarity matrix $\mathbf{D}$ on Γ and a fixed $i \in \Gamma$. We denote $\mathbf{D}^i$ as the (dissimilarity) matrix obtained from $\mathbf{D}$ by removing its i-th row and column. Finally, whenever appropriate, we shall refer to a dissimilarity matrix as an *instance* of the BMEP.

We recall that a dissimilarity matrix $\mathbf{D}$ on Γ is said to be *metric* if its entries satisfy the following *triangle inequality*

$$d_{ij} + d_{jk} \geq d_{ik} \qquad \forall \text{ distinct } i, j, k \in \Gamma.$$

A metric dissimilarity matrix $\mathbf{D}$ on Γ is said to be *additive* if for every distinct $i, j, p, q \in \Gamma$, precisely one of the following *four point conditions* holds [2]:

$$\begin{aligned} d_{ij} + d_{pq} &\leq d_{ip} + d_{jq} = d_{iq} + d_{jp}, \\ d_{ip} + d_{jq} &\leq d_{ij} + d_{pq} = d_{iq} + d_{jp}, \\ d_{iq} + d_{jp} &\leq d_{ij} + d_{pq} = d_{ip} + d_{jq}. \end{aligned}$$

For metric dissimilarity matrices the following result holds:

Proposition 2.1 (Fiorini and Joret [14])**.** *Any metric instance of the BMEP admits a polynomial-time approximation algorithm having an approximation factor of* 2*.*

Fiorini and Joret's approximation algorithm [14] proceeds in two steps. First, it computes a *Minimum Spanning Tree* (MST) over the complete graph having a vertex for each item $i \in \Gamma$ and a weight d_{ij} for each edge (i, j). Subsequently, the algorithm transforms this MST into a valid UBT on Γ with a total cost no greater than that of the MST by recursively replacing each internal vertex of degree greater than 3 with a binary subtree. Fiorini and Joret [14] shows that the cost of the resulting UBT is at most twice the optimal value of the BMEP, thus yielding the 2-approximation guarantee.

Consider a dissimilarity matrix $\mathbf{D}$ on Γ and any tree T having Γ as a leaf set. Let w denote a function that associates a nonnegative weight $w(e)$ to each edge e of T. In addition, let P_{ij}^T denote the unique path in T between the distinct leaves $i, j \in \Gamma$ and define

$$w(P_{ij}^T) = \sum_{e \in P_{ij}^T} w(e).$$

Then, we say that T is *additive* for $\mathbf{D}$ (or *fits* $\mathbf{D}$, for short) if the following *additivity constraint* holds:

$$d_{ij} = w(P_{ij}^T) \qquad \forall \text{ distinct } i, j \in \Gamma. \tag{2}$$

A classical result from Buneman [2] shows that the set of additive dissimilarity matrices on Γ corresponds to the set of $W = \{w(P_{ij}^T)\}$ matrices associated with additive trees on Γ, whose entries satisfy (2). Waterman et al. [24] further show that an additive tree T, with no zero weight internal edge, always exists for a given additive dissimilarity matrix $\mathbf{D}$. This tree is unique if all four-point conditions hold with strict inequalities, and can be constructed in polynomial time by means of the iterative algorithm described in [24]. A natural question is whether this algorithm could also serve as a certificate of optimality for additive instances of the BMEP. In this respect, there is, in principle, no clear evidence supporting such a possibility. Indeed, Waterman et al.'s algorithm is designed to solve a feasibility problem (i.e., finding an additive tree for $\mathbf{D}$), whereas the BMEP is an optimization problem that neither enforces nor assumes additivity, whether explicitly or implicitly. Thus, the tree returned by Waterman et al.'s algorithm may, in general, differ from the optimal solution to an additive instance $\mathbf{D}$ of the BMEP. Nonetheless, we will see in Section 3 that both Buneman's and Waterman et al.'s foundational results still provide valuable insights for understanding the relationship between additive trees and the optimal

solutions of additive BMEP instances. Before moving to the next section, we briefly recall here that any UBT $T \in \Theta_n$ can be encoded by means of a *path-length matrix* $\tau = \{\tau_{ij}\}$ whose combinatorial properties have been extensively discussed in [10]. Some properties that will be frequently invoked throughout the article are the following:

Proposition 2.2 (Catanzaro et al. [6])**.** *Let T be a UBT in Θ_n. Then, the path-length matrix $\tau = \{\tau_{ij}\}$ encoding T satisfies the following* Kraft equality*:*

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

Proposition 2.3 (Catanzaro et al. [6])**.** *Let T be a UBT in Θ_n. Then, the path-length matrix $\tau = \{\tau_{ij}\}$ encoding T satisfies the following* manifold condition*:*

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

3 Efficiently solvable instances of the BMEP

In this section, we present some instances of the BMEP that can be solved efficiently. We start by considering the simplest case in which $\mathbf{D}$ is additive and has all equal non-diagonal entries. In such a case, the following result holds:

Proposition 3.1. *Let $\mathbf{D}$ be a dissimilarity matrix such that $d_{ij} = k$, for some positive real k. Then, for every UBT $T \in \Theta_n$, $L_{\mathbf{D}}(T) = k \cdot n/2$.*

Proof. Let $T \in \Theta_n$. Then, we have that

$$\begin{aligned} L_{\mathbf{D}}(T) &= \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} k 2^{-\tau_{ij}} = \\ &= k \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} 2^{-\tau_{ij}} \overset{\text{Kraft}}{=} k \sum_{i \in \Gamma} \frac{1}{2} = k\,\frac{n}{2}. \end{aligned}$$

□ □

Proposition 3.1 highlights a first conceptual difference between Waterman et al.' s framework and the BMEP. Specifically, in Waterman et al.'s setting, an additive distance matrix $\mathbf{D}$ with constant entries $d_{ij} = k$ admits the star-tree with n leaves and edge weights equal to $k/2$ as its unique tree representation (additive tree). In contrast, within the BMEP framework, the same distance matrix $\mathbf{D}$ does not induce a unique optimal solution: every UBT $T \in \Theta_n$ attains the same objective value and is therefore optimal. This loss of uniqueness is explained by the topological restriction of the BMEP to UBTs, as opposed to Waterman et al.'s framework, where the additive tree may be non-binary. This distinction is not merely technical. The two frameworks address fundamentally different problems: Waterman et al. study the problem of exactly fitting an additive distance matrix by a tree metric, while the BMEP seeks a tree that optimizes a global functional of the distances. As a consequence, even when the input matrix is additive, the BMEP objective may not single out a preferred topology, revealing a different notion of optimality. In particular, for constant distance matrices, the BMEP objective becomes invariant over the space of UBTs, reflecting the absence of topological information in the data rather than a limitation of the estimation model. An interesting question, which lies beyond the scope of the present article, is whether allowing non-binary trees in the BMEP could restore uniqueness, at least for constant distance matrices. Addressing this question clearly warrants further investigation.

Remark 3.2. Proposition 3.1 can be straightforwardly extended to the case in which $d_{ij} = (k_i + k_j)/2$, for some reals $k_i > 0$, $i \in \Gamma$. In this setting, for every UBT $T \in \Theta_n$, it holds that $L_{\mathbf{D}}(T) = \frac{1}{2}\sum_{i\in\Gamma} k_i$. An analogous extension can be formulated for all subsequent propositions that assume the dissimilarities d_{ij} contain a constant term k.

As a second case, consider a dissimilarity matrix $\mathbf{D}$ such that $d_{ij} \gg d_{pq}$ for some item $i \in \Gamma$ and for every distinct $j, p, q \in \Gamma \setminus \{i\}$. Without loss of generality, assume that: first, $i = 1$; second, the entries d_{1j} are sorted in increasing order; and third, for some positive real $k < d_{12}$, $d_{pq} = k$, for all distinct $p, q \in \{2, \ldots, n\}$. This last assumption holds whenever the entries d_{1j} are large enough to imply

$$\max_{p,q,k,\ell\in\Gamma\setminus\{1\}} \{(d_{pq} - d_{k\ell})/d_{12} \approx 0\}$$

i.e., whenever the difference $d_{pq} - d_{k\ell}$ in negligible. An example of such a situation occurs when $d_{1j}/d_{pq} > 2^{n-2}$. Indeed, in this case, the penalty incurred by reducing the path length τ_{1j} by one unit outweighs the maximum possible benefit obtained by increasing up to $n-1$ other path lengths τ_{pq}.

In the light of what we discussed above, we say that a BMEP instance matrix is *nearly rooted* if, for some positive real k, we have that $d_{ij} \geq k$ for all $j \in \Gamma \setminus \{i\}$ and $d_{pq} = k$, for all distinct $p, q \in \Gamma \setminus \{i\}$. Then, the following result holds:

Proposition 3.3. *Let* $\mathbf{D}$ *be any nearly rooted BMEP instance such that* $d_{12} \leq d_{13} \leq \ldots \leq d_{1n}$, *and* $d_{pq} = k$, *for all distinct* $p, q \in \{2, \ldots, n\}$ *and for some* $0 < k < d_{12}$. *Then, the value of the optimal solution* $T^*_{\mathbf{D}}$ *to the instance of the BMEP* $\mathbf{D}$ *is*

$$L_{\mathbf{D}}(T^*_{\mathbf{D}}) = \frac{n}{2}k + \frac{d_{12}}{4} + \ldots + \frac{d_{1j}}{2^j} + \ldots + \frac{d_{1\,n-1}}{2^{n-1}} + \frac{d_{1n}}{2^{n-1}}.$$

Proof. Consider the matrix $\mathbf{D}' = \{d'_{ij}\}$, such that $d'_{ij} = d_{ij} - k$, for every distinct $i, j \in \Gamma$, and 0 otherwise. Then, by (1) and Proposition 3.1, for every UBT $T \in \Theta_n$, we have that:

$$\begin{aligned} L_{\mathbf{D}}(T) &= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d_{ij}2^{-\tau_{ij}} = \\ &= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d'_{ij}2^{-\tau_{ij}} + \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} k\cdot 2^{-\tau_{ij}} = \\ &= L_{\mathbf{D}'}(T) + k\cdot n/2. \end{aligned}$$

This equality also holds for the optimal solution $T^*_{\mathbf{D}}$, hence the optimal solution does not depend on k. Thus, suppose, without loss of generality, that $k = 0$. We proceed by induction. For $n = 4$, a direct computation shows that as $d_{12} \leq d_{13} \leq d_{14}$, the optimal solution is T^*_4, satisfying $\tau_{12} = 2$, $\tau_{13} = \tau_{14} = 3$. This configuration coincides with the claim that the UBT must be a caterpillar. Now, assume that the statement holds true for $n = r > 4$; then, we show that it also holds true for $n = r + 1$. To this end, for each $j \in \Gamma \setminus \{1\}$, denote by

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

i.e., $L^j(T)$ represents the BMEP objective function $L_{\mathbf{D}}(T)$ with all the terms involving j removed. Minimizing the BMEP on the instance $\mathbf{D}^j$ and replacing exactly one arbitrary edge with a path of length 2 is therefore equivalent to minimizing $L^j(T)$ over all UBTs on $r + 1$ items. Since each pair $(1, \ell)$ appears in every $L^j(T)$, except when ℓ is equal to 1 or j, then

$$\sum_{j\in\Gamma\setminus\{1\}} L^j(T) = (r-1)L_{\mathbf{D}}(T).$$

Additive BMEP Solver 1: Solver for additive instances of the BMEP

Input: An additive dissimilarity matrix $\mathbf{D} = \{d_{ij}\}$ on a set Γ of $n \geq 3$ items;
Output: A UBT $T = (V, E)$ on Γ with edge weights $w(e)$, $e \in E$;

1 Run Waterman et al.'s algorithm on $\mathbf{D}$ to compute the additive tree $T = (V, E)$ with edge-weights $w(e)$, $e \in E$;
2 **while** *there exists an internal vertex* $v \in V \setminus \Gamma$ *with* $|N(v)| > 3$ **do**
3 Select two arbitrary vertices $u, u' \in N(v)$;
4 Set $N = N(v) \setminus \{u, u'\}$;
5 Remove all the edges of the form (v, u'') with $u'' \in N$;
6 Create a new internal vertex $v' \notin V$ and add it to V;
7 Add the edge (v, v') to E and set $w(v, v') = 0$;
8 Add an edge of the form (v', u'') to E for all $u'' \in N$;
9 Set $w(v', u'') = w(v, u'')$ for all $u'' \in N$;
10 **end**
11 **return** (T, w)

Since removing any item $j \in \Gamma \setminus \{1\}$ does not alter the order in the sequence $d_{12} \leq \ldots \leq d_{1n}$, we can use induction over $\mathbf{D}^j$. If we consider the caterpillar $T' \in \Theta_{r+1}$ satisfying the claim we can note that, after removing j and contracting the two edges adjacent to the unique (internal) vertex adjacent to j in T', the new caterpillar T^j so obtained coincides with the caterpillar respecting the claim for the instance $\mathbf{D}^j$. Moreover, T^j is the optimal solution to the instance $\mathbf{D}^j$, by the induction hypothesis, and hence, $L^j(T')$ is optimal by construction. As $(r-1)L_{\mathbf{D}}(T')$ decomposes into $\sum_{j \in \Gamma \setminus \{i\}} L^j(T')$ and the minimum of each term is achieved by T', we can conclude that $(r-1)L_{\mathbf{D}}(T')$ is minimal. Thus, $L_{\mathbf{D}}(T')$ itself is minimal by linearity. □ □

Remark 3.4. Consider a caterpillar T'' obtained by reversing the label order with respect to Proposition 3.3, i.e.,

$$L_{\mathbf{D}}(T'') = \frac{(n-2)}{2}k + \frac{d_{12}}{2^{n-1}} + \frac{d_{13}}{2^{n-1}} + \frac{d_{14}}{2^{n-2}} + \ldots + \frac{d_{1j}}{2^{n-j+2}} + \ldots + \frac{d_{1\,n-1}}{8} + \frac{d_{1n}}{4}.$$

Then, $L_{\mathbf{D}}(T'') \geq L_{\mathbf{D}}(T''')$ for every UBT $T''' \in \Theta_n$, i.e., T'' is the UBT with the greatest cost.

Now consider the case in which $\mathbf{D}$ is an additive dissimilarity matrix for a set Γ of $n \geq 3$ items, and let T denote the additive tree for $\mathbf{D}$ computed by Waterman et al.'s algorithm. We show that whenever T is not a UBT (i.e., whenever T is not a feasible solution for the BMEP), the degree of any internal vertex exceeding 3 can always be reduced by replacing the edges incident to that vertex with a suitable zero-weight subtree, so that the resulting graph is an *additive UBT for* $\mathbf{D}$, hence a feasible solution for the BMEP instance encoded by $\mathbf{D}$. In particular, we present an algorithm, called the *Additive BMEP Solver*, that computes such an additive UBT in $\mathcal{O}(n^2)$ time and we will prove, in Proposition 3.5, that such a UBT is optimal for $\mathbf{D}$.

The algorithm takes as input the matrix $\mathbf{D}$ and proceeds in two phases. First, it constructs an additive tree $T = (V, E)$ for $\mathbf{D}$ by using Waterman et al.'s algorithm (line 1 of Additive BMEP Solver). If T is already a UBT on Γ, then the algorithm terminates and returns T. Otherwise, T must contain at least one internal vertex of degree strictly greater than three. In this case, the algorithm enters a second phase, illustrated in Figure 2, in which it iteratively applies a sequence of topological transformations to T until turning it into a UBT on Γ, thereby producing a feasible solution for this additive instance of the BMEP. The core of this second phase is the while loop in lines 2–10, which reduces the degree of each internal vertex v in

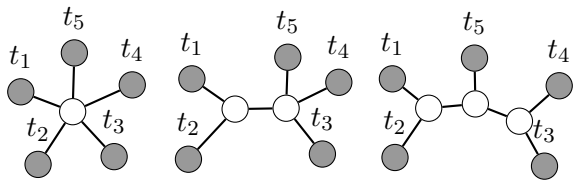

Figure 2: A simple example that shows how Additive BMEP Solver works. Left: An initial additive tree (in this case, a star-tree) is first computed by running Waterman et al.'s algorithm (line 1). Center: the insertion of a first internal edge of weight zero ensures that one endpoint has degree 3 (lines 3-9). Right: a reiterated insertion of a further internal edge yields a UBT, i.e., a feasible solution to the BMEP (lines 3-9).

T with degree greater than three by: (i) selecting two arbitrary vertices u and u' adjacent to v (line 3); (ii) storing in $N = N(v) \setminus \{u, u'\}$ the remaining neighbors of v (line 4); (iii) removing all edges incident to v with an endpoint in N (line 5); (iv) adding a new internal vertex v' to V (line 6); (v) adding the edge (v, v') in E, with zero weight, so as to ensure that in the current forest the degree of v is exactly three (line 7); (vi) adding new edges to E of the form (v', u'') for all $u'' \in N$ (line 8), thereby restoring connectivity; and finally (vii) setting $w(v', u'') = w(v, u'')$, for all $u'' \in N$. Note that the order in which lines 5, 7, and 8 are executed guarantees that the graph produced at the end of each iteration remains connected and acyclic, hence a tree. Moreover, since the while loop continues until every internal vertex of T has degree exactly three, the final tree is a UBT on Γ. The UBT returned by the Additive BMEP Solver is not uniquely determined. Indeed, any arbitrary choice at line 3 may lead to alternative UBTs on Γ. Any of such a alternatives, however, preserves additivity; consequently, up to suitable topological transformations, the tree produced by Waterman et al.'s algorithm and any arbitrary tree returned by the Additive BMEP Solver are equivalent with respect to their additive structure.

As regards the computational complexity of the Additive BMEP Solver algorithm, observe that the number of the internal vertices in the original Waterman et al.'s additive tree T is at most $n-2$. For each internal vertex v in T with degree strictly greater than three the algorithm carried out exactly $|N(v)| - 3$ iterations. Since any tree with n leaves has at most $n-1$ edges and, by Euler's formula, $2|E| = \sum_{u \in V} |N(u)|$, we can easily derive that the overall complexity of Additive BMEP Solver is $\mathcal{O}\big(n^2 + (n-2)(n-1)\big) = \mathcal{O}(n^2)$. Note, in particular, this bound is tight when the additive tree provided by Waterman et al.'s algorithm is a star-tree.

We now show that any additive UBT is optimal for the BMEP instance $\mathbf{D}$.

Proposition 3.5. *Let* $\mathbf{D}$ *be an additive instance of the BMEP. Then any additive UBT for* $\mathbf{D}$ *is optimal for this instance.*

Proof. The proof is similar to that of Proposition 3.3, as both instances preserve their properties upon removing a leaf. We therefore proceed by induction, indicating only the specific differences. We start by considering the simple case in which $n = 3$. In such a case, there exists a unique UBT on Γ which necessarily coincides with Waterman et al.'s additive tree. Such a tree is trivially optimal for $\mathbf{D}$.

Assume now that the statement holds true for $n = r > 3$; then, we show that it also holds true for $n = r + 1$. To this end, for each $i \in \Gamma$, define

$$L^i(T) = \sum_{\ell \in \Gamma} \sum_{\ell' \in \Gamma \setminus \{i\}} d_{\ell\ell'} \, 2^{-\tau_{\ell\ell'}}.$$

Now, observe that minimizing the BMEP on the instance $\mathbf{D}^i$ and replacing exactly one arbitrary edge with a path of length 2 is equivalent to minimizing $L^i(T)$ over all UBTs on $r + 1$ items.

In addition, in the additive case, we also have

$$\sum_{i\in\Gamma} L^i(T) = (r-1)L_{\mathbf{D}}(T).$$

Then, the proof of the statement for $n = r + 1$ trivially follows by using arguments similar to those used in the proof of Proposition 3.3. □ □

It remains to prove that, for any additive UBT $T^*_{\mathbf{D}}$ fitting $\mathbf{D}$, the edge weights coincide with Pauplin's weights. Consequently, the Additive BMEP Solver returns precisely Pauplin's weights. The following proposition addresses this issue.

Proposition 3.6. *Let $\mathbf{D}$ denote an additive instance of the BMEP and let $T^*_{\mathbf{D}}$ denote any additive UBT for $\mathbf{D}$. Then, the weights on the edges of $T^*_{\mathbf{D}}$ coincide with those provided by Pauplin's edge weight model [22].*

Proof. We need to show that the edge weights of Pauplin's edge weight model coincide with those of the additive UBT $T^*_{\mathbf{D}}$ fitting $\mathbf{D}$. To this end, consider an item $i \in \Gamma$ and let u denote the unique vertex in $T^*_{\mathbf{D}}$ that is adjacent to i. In addition, let $w_A(i,u)$ and $w_B(i,u)$ denote the weights over the edges of $T^*_{\mathbf{D}}$ used for fitting $\mathbf{D}$ and Pauplin's edge weight model would assign to the edge (i,u), respectively, and denote by Γ' and Γ'' the sets of items of the two connected components of $T^*_{\mathbf{D}} \setminus \{i,u\}$. By Pauplin's edge weight model for edges incident to leaves, we have that

$$\begin{aligned}
w_B(i,u) &= \sum_{p\in\Gamma'} d_{ip}\, 2^{1-\tau_{ip}} + \sum_{q\in\Gamma''} d_{iq}\, 2^{1-\tau_{iq}} \\
&\quad - \sum_{p\in\Gamma'}\sum_{q\in\Gamma''} d_{pq}\, 2^{1-\tau_{pq}} = \\
&= \sum_{p\in\Gamma'} w_A(P^{T^*_{\mathbf{D}}}_{ip})\, 2^{1-\tau_{ip}} + \sum_{q\in\Gamma''} w_A(P^{T^*_{\mathbf{D}}}_{iq})\, 2^{1-\tau_{iq}} \\
&\quad - \sum_{p\in\Gamma'}\sum_{q\in\Gamma''} w_A(P^{T^*_{\mathbf{D}}}_{pq})\, 2^{1-\tau_{pq}} = \\
&\quad \sum_{p\in\Gamma'}\sum_{e\in E(P^{T^*_{\mathbf{D}}}_{ip})} w_A(e)\, 2^{1-\tau_{ip}} \\
&\quad + \sum_{q\in\Gamma''}\sum_{f\in E(P^{T^*_{\mathbf{D}}}_{iq})} w_A(f)\, 2^{1-\tau_{iq}} \\
&\quad - \sum_{p\in\Gamma'}\sum_{q\in\Gamma''}\sum_{g\in E(P^{T^*_{\mathbf{D}}}_{pq})} w_A(g)\, 2^{1-\tau_{pq}}.
\end{aligned} \tag{3}$$

Now, it is worth noting that $P^{T^*_{\mathbf{D}}}_{pq}$ decomposes into $P^{T^*_{\mathbf{D}}}_{pu}$ and $P^{T^*_{\mathbf{D}}}_{uq}$. Similarly, $P^{T^*_{\mathbf{D}}}_{ip}$ and $P^{T^*_{\mathbf{D}}}_{iq}$ decompose into (i,u) and $P^{T^*_{\mathbf{D}}}_{up}$ and (i,u) and $P^{T^*_{\mathbf{D}}}_{uq}$, respectively. Then, we can rewrite the right-most side of (3) as

$$\begin{aligned}
&w_A(i,u)\left(\sum_{j\in\Gamma\setminus\{i\}} 2^{1-\tau_{ij}}\right) \\
&+ \sum_{p\in\Gamma'}\sum_{e\in E(P^{T^*_{\mathbf{D}}}_{up})} w_A(e)\left(2^{1-\tau_{ip}} - \sum_{q\in\Gamma''} 2^{1-\tau_{pq}}\right) \\
&+ \sum_{q\in\Gamma''}\sum_{f\in E(P^{T^*_{\mathbf{D}}}_{uq})} w_A(f)\left(2^{1-\tau_{iq}} - \sum_{p\in\Gamma'} 2^{1-\tau_{pq}}\right).
\end{aligned} \tag{4}$$

By applying Kraft's equality to the subtree of $T^*_{\mathbf{D}}$ (which incidentally is a UBT) with leaf-set $\{u\} \cup \Gamma''$, we have that

$$\begin{aligned}
&\sum_{q\in\Gamma''} 2^{-\tau_{pq}} = \sum_{q\in\Gamma''} 2^{-(\tau_{ip}+\tau_{iq}-2)} \\
&= 2^{-(\tau_{ip}-2)} \sum_{q\in\Gamma''} 2^{-\tau_{iq}} = 2^{-(\tau_{ip}-2)} \sum_{q\in\Gamma''} 2^{-(\tau_{uq}+1)} \\
&= 2^{-(\tau_{ip}-1)} \sum_{q\in\Gamma''} 2^{-\tau_{uq}} \overset{\text{Kraft}}{=} 2^{-(\tau_{ip}-1)} \cdot \frac{1}{2} = 2^{-\tau_{ip}}
\end{aligned} \tag{5}$$

hence, the terms in the second and third parentheses of (4) equal zero. Then, by (3), (4), and (5), we get

$$w_B(i,u) = w_A(i,u) \left(\sum_{j\in\Gamma\setminus\{i\}} 2^{1-\tau_{ij}} \right) \overset{\text{Kraft}}{=} w_A(i,u).$$

By using similar arguments, it is easy to see that the same result also holds for the weights assigned to any internal edge, thus the statement follows. □ □

In the light of Propositions 2.3, 3.1, 3.5, and 3.6 the following result holds:

Corollary 3.7. *Let* $\mathbf{D}$ *be an instance of the BMEP such that* $d_{ij} = \alpha\tau_{ij}+k$*, where* $\{\tau_{ij}\}$ *encodes any UBT* $T \in \Theta_n$*, and* α *and* k *are positive constants. Then,* $T \in \Theta_n$ *is optimum for* $\mathbf{D}$*, its optimal value is* $L_{\mathbf{D}}(T) = \alpha(2n-3) + k \cdot n/2$*, and the weight assigned to any edge is* $\alpha + k/2$ *if the edge is incident to a leaf of* T*, and* α *otherwise.*

This result implies that if the dissimilarity matrix $\mathbf{D}$ is derived from an affine transformation of the path-length matrix τ of a UBT in Θ_n, then this new instance of the BMEP can be solved efficiently and its optimal solution is τ itself.

4 Upper and lower bounds on the optimal value of the BMEP

Given a generic dissimilarity matrix $\mathbf{D}$, not necessarily additive, we define $d^+ = \max_{i\neq j} d_{ij}$ and $d^- = \min_{i\neq j} d_{ij}$. In addition, for a given $i \in \Gamma$, denote by σ_i a permutation over $\Gamma \setminus \{i\}$ such that $d_{i\sigma_i(j)} \leq d_{i\sigma_i(j+1)}$, for each $j \in \Gamma \setminus \{i\}$, and define:

$$L_i^- = \left(\sum_{j\in\Gamma\setminus\{i,\,\sigma_i(n)\}} d_{i\sigma_i(j)} 2^{-\sigma_i(j)} \right) + d_{i\sigma_i(n)} 2^{-(n-1)}$$

and

$$L_i^+ = d_{i\sigma_i(2)} 2^{n-1} + \left(\sum_{j\in\Gamma\setminus\{i,\,\sigma_i(2)\}} d_{i\sigma_i(j)} 2^{n-\sigma_i(j)+2} \right).$$

Then, the following result holds:

Proposition 4.1. *The objective function of the BMEP satisfies*

$$d^- \cdot n/2 \leq \sum_{i\in\Gamma} L_i^- \leq L_{\mathbf{D}}(T) \leq \sum_{i\in\Gamma} L_i^+ \leq d^+ \cdot n/2$$

for any $T \in \Theta_n$*.*

Proof. We first observe that, by definition of L_i^- and Kraft's equality, $d^-/2 = d^- \cdot (1/4 + 1/8 + \ldots + 2 \cdot 1/2^{n-1} \leq L_i^-$ holds true for every $i \in \Gamma$. Moreover, by Proposition 3.3, we have that $L_i^- \leq \sum_j d_{ij} 2^{-\tau_{ij}}$, hence $\sum_i L_i^- \leq \sum_i \sum_j d_{ij} 2^{-\tau_{ij}} = L_{\mathbf{D}}(T)$ holds true for every UBT T in Γ. Similarly, we have that $d^+/2 \geq L_i^+ \geq \sum_j d_{ij} 2^{-\tau_{ij}}$. By summing with respect to i, the statement follows. □ □

Note that the inequalities in Proposition 4.1 are tight, as both the equality $d^- \cdot n/2 = \sum_i L_i^-$ and $d^+ \cdot n/2 = \sum_i L_i^+$ hold if and only if $d^- = d^+$ (and the solution of this case is claimed in Proposition 3.1). In addition, Proposition 4.1 implies that the error between any feasible solution and the optimal one grows linearly with respect to the size of the instance. In particular, if the difference d^- and d^+ is negligible, the quality of any feasible solution can be regarded as consistently good.

5 On the approximation of practical instances of the BMEP

Given any dissimilarity matrix $\mathbf{D} = \{d_{ij}\}$, with $d_{ij} > 0$, for all distinct $i, j \in \Gamma$, define the following two differential quantities:

$$\delta = \max_{i,j,k \in \Gamma} |d_{ik} - d_{ij} - d_{jk}|$$

and

$$\Delta = \frac{2}{n} \sum_{s \in \Gamma} (L_s^+ - 2L_s^-).$$

In addition, define

$$\varphi = \frac{\delta}{2 \sum_{i \in \Gamma} L_i^-}$$

and

$$\rho = \frac{\sum_{i \in \Gamma} L_i^+}{\sum_{i \in \Gamma} L_i^-}.$$

Then, the following proposition holds:

Proposition 5.1. *The BMEP admits a polynomial-time approximation algorithm with an approximation ratio of at most*

$$\begin{cases} 2 + \varphi\, n & \text{if } \delta < \Delta, \\ \rho & \text{otherwise.} \end{cases}$$

Proof. If $\delta \geq \Delta$, the statement trivially follows from Proposition 4.1. We then consider the case in which $\delta < \Delta$. Observe that in this case $2 + \varphi\, n < \rho$ by definition of δ and Δ, i.e., the approximation ratio $2 + \varphi\, n$ is not trivial to obtain. Define a new instance of the BMEP $\mathbf{D}' = \{d'_{ij}\}$ as:

$$d'_{ij} := d_{ij} + \delta, \quad \text{for all distinct } i, j \in \Gamma$$

and $d'_{ij} := 0$ otherwise. By definition of δ, $\mathbf{D}'$ is metric. Indeed, the following relationship holds

$$\begin{aligned} d'_{ij} + d'_{jk} &= (d_{ij} + \delta) + (d_{jk} + \delta) = d_{ij} + d_{jk} + 2\delta \\ &\geq d_{ik} + \delta = d'_{ik} \end{aligned}$$

for all distinct $i, j, k \in \Gamma$. Now, we prove that

$$T^*_{\mathbf{D}} = T^*_{\mathbf{D}'}.$$

Indeed, with respect to $\mathbf{D}'$ the following chain of equalities holds for any UBT $T \in \Theta_n$:

$$\begin{aligned}
L(T) &= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d'_{ij}2^{-\tau_{ij}} = \\
&= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} (d_{ij}+\delta)2^{-\tau_{ij}} \\
&= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d_{ij}2^{-\tau_{ij}} + \delta\sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} 2^{-\tau_{ij}} \\
&= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d_{ij}2^{-\tau_{ij}} + \delta\sum_{i\in\Gamma}\left(\sum_{j\in\Gamma} 2^{-\tau_{ij}}\right) \\
&= \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} d_{ij}2^{-\tau_{ij}} + \frac{n}{2}\delta
\end{aligned}$$

where the last equality follows from Proposition 3.1. Hence, $T^*_{\mathbf{D}}$ is also optimal for $\mathbf{D}'$ and has value

$$L_{\mathbf{D}'}(T^*_{\mathbf{D}}) = L_{\mathbf{D}}(T^*_{\mathbf{D}}) + \frac{n}{2}\delta.$$

Now, because $\mathbf{D}'$ is metric, by Proposition 2.1, it is possible to compute in polynomial time a UBT $T' \in \Theta_n$ for $\mathbf{D}'$ such that $L_{\mathbf{D}'}(T') \leq 2 \cdot L_{\mathbf{D}'}(T^*_{\mathbf{D}})$, i.e.,

$$L_{\mathbf{D}'}(T') \leq 2\left(L_{\mathbf{D}}(T^*_{\mathbf{D}}) + \frac{n}{2}\delta\right). \tag{6}$$

By Proposition 4.1, we get

$$L_{\mathbf{D}}(T^*_{\mathbf{D}}) \geq \sum_{i\in\Gamma} L_i^- \tag{7}$$

and by combining (6) and (7), we get

$$\frac{L_{\mathbf{D}}(T')}{L_{\mathbf{D}}(T^*_{\mathbf{D}})} \leq 2 + \frac{\frac{n}{2}\delta}{L_{\mathbf{D}}(T^*_{\mathbf{D}})} \leq 2 + \frac{\frac{n}{2}\delta}{\sum_{i\in\Gamma} L_i^-} = 2 + \varphi\, n.$$

As $\mathbf{D}'$ can be constructed in $\mathcal{O}(n^2)$ and T' can be computed in $\mathcal{O}(n^3)$ [14], the statement follows. □ □

Proposition 5.1 establishes that the BMEP admits an efficient approximation whenever the dissimilarity matrix $\mathbf{D}$ has positive off-diagonal entries and a bounded ratio ρ. This assumption is particularly meaningful in practice, since dissimilarity matrices of this kind naturally arise from biological data. Indeed, standard models of molecular evolution [5, 13, 20] typically yield positive dissimilarities between distinct molecular sequences. Moreover, the entries d_{ij} are inherently subject to finite precision (stemming from both data acquisition and floating-point representation), which effectively bounds ρ by a polynomial function of the input size. Furthermore, when ρ grows exponentially and $\mathbf{D}$ has a diagonal block structure, a good-quality solution can still be obtained by applying Proposition 3.3 to each block, under the assumption that every diagonal block is itself a nearly rooted instance of the BMEP.

To conclude this section, we investigate the problem of providing a meaningful measure of algorithmic quality for the BMEP in contexts where no constant-factor approximation is possible. To this end, we recall that for a minimization problem, the *differential approximation ratio* is defined as

$$\frac{c_{\text{apx}} - OPT}{c_{\text{worst}} - OPT},$$

where c_{apx}, c_{worst}, and OPT denote, respectively, the value of the solution found by an approximation algorithm, the maximum value achievable in the set of feasible solutions to the problem, and the value of the optimal solution. Interestingly, although Proposition 5.1 does not yield tight bounds for Fiorini and Joret's approximation algorithm, the following general elementary result still holds:

Proposition 5.2. *The differential approximation ratio of the BMEP is the same of the metric BMEP.*

Proof. Let $\mathbf{D}$ denote any instance of the BMEP. Consider the matrix $\mathbf{D'} = \{d'_{ij}\}$, defined by $d'_{ij} = d_{ij} + d^+$ for every $i \neq j$. It is easy to see that $\mathbf{D'}$ is a metric instance of the BMEP. By Proposition 3.1, $L_{\mathbf{D'}}(T) = L_{\mathbf{D}}(T) + d^+ \cdot n/2$, for every $T \in \Theta_n$. Thus, if T' is the UBT obtained from Fiorini and Joret's approximation algorithm and $T'' \in \text{argmax}_{T \in \Theta_n}\{L_{\mathbf{D}}(T)\}$, then we have

$$\frac{L_{\mathbf{D'}}(T') - L_{\mathbf{D'}}(T^*_{\mathbf{D}})}{L_{\mathbf{D'}}(T'') - L_{\mathbf{D'}}(T^*_{\mathbf{D}})} = \frac{L_{\mathbf{D}}(T') - L_{\mathbf{D}}(T^*_{\mathbf{D}})}{L_{\mathbf{D}}(T'') - L_{\mathbf{D}}(T^*_{\mathbf{D}})}.$$

□ □

The last proposition leaves open the possibility to provide better upper bounds for Fiorini and Joret's approximation algorithm whenever the ratio ρ is bounded.

A simple example. Consider the matrix

$$\mathbf{D} = \begin{bmatrix} 0.000 & 5.691 & 3.872 & 4.268 & 0.163 & 5.610 \\ 5.691 & 0.000 & 6.401 & 0.051 & 2.812 & 2.721 \\ 3.872 & 6.401 & 0.000 & 4.200 & 3.098 & 2.244 \\ 4.268 & 0.051 & 4.200 & 0.000 & 2.796 & 3.396 \\ 0.163 & 2.812 & 3.098 & 2.796 & 0.000 & 4.502 \\ 5.610 & 2.721 & 2.244 & 3.396 & 4.502 & 0.000 \end{bmatrix}.$$

A direct computation shows that $2 + \varphi n = 3.226$, with $\delta = 2.717$, while $\rho = 4.08$. In particular, the optimum (computed with a branch and cut algorithm) is 7.39, while the cost of the solution obtained by the Fiorini and Joret's approximation algorithm is 8.343, and hence, the actual ratio is 1.129. Note that the differential approximation ratio is 0.1906, where we used $\max_{T \in \Theta_n} L_{\mathbf{D}}(T) = 12.39$ to compute it.

Acknowledgments

The first and the third authors acknowledge support from the Belgian National Fund for Scientific Research (FNRS) via the grant FNRS PDR 40007831. This first author also acknowledges support from the Fondation Louvain, via the grant "COALESCENS". The authors thank the anonymous referees for their helpful comments, which significantly improved the paper, in particular the algorithm for finding optimal solutions to the BMEP on additive trees.

References

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

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

[3] D. Catanzaro. The minimum evolution problem: Overview and classification. *Networks*, 53, 2009.

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

[5] 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.

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

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

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

[9] D. Catanzaro, M. Frohn, O Gascuel, and R. Pesenti. A tutorial on the balanced minimum evolution problem. *European Journal of Operational Research*, 300(1):1–19, 2022.

[10] D. Catanzaro, R. Pesenti, A. Sapucaia, and L. Wolsey. Optimizing over path-length matrices of unrooted binary trees. *Mathematical Programming*, pages 1–53, 2025.

[11] M. A. Cueto and F. A. Matsen. Polyhedral geometry of phylogenetic rogue taxa. *Bulletin of Mathematical Biology*, 73(6):1202–1226, 2011.

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

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

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

[15] S. Forcey, L. Keefe, and W. Sands. Facets of the balanced minimal evolution polytope. *Journal of Mathematical Biology*, 73(2):447–468, December 2015.

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

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

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

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

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

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

[22] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51(1):41–47, July 2000.

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

[24] M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64(2):199–213, 1977.