# On the Practical Approximability of an Inapproximable Problem

Francesco Pisanu[a,*]

[a]CORE-Université Catholique de Louvain, Voie du Roman Pays 34, Ottignies-Louvain-la-Neuve, 1348, Belgium

ARTICLE INFO

*Keywords*:
unrooted binary trees
path-length matrix
approximation algorithms.

ABSTRACT

The Balanced Minimum Evolution Problem (BMEP) is a classical problem in phylogenetics, known to be NP-hard and inapproximable within a factor $c^n$ for some $c > 1$, unless $P = NP$. In this work, we show that when the input distance matrix satisfies a bounded ratio between its largest and smallest nonzero entries, the BMEP becomes polynomial-time approximable within a constant factor depending on this ratio. This observation suggests that the hardness of the BMEP is most prominent in worst-case instances with unbounded variation in input distances, and provides a more nuanced perspective for its use in practical applications involving molecular sequences.

## 1. Introduction

Let $\Gamma = \{1, 2, \dots, n\}$ be a set of target items, and let $D = (d_{ij})_{i,j\in\Gamma}$ be a symmetric matrix with non-negative entries and zero diagonal, representing pairwise dissimilarities. The emphBalanced Minimum Evolution Problem (BMEP) consists in finding an Unrooted Binary Tree (UBT) $T$ with $n$ leaves, bijectively labeled by the items in $\Gamma$, so as to minimize the total tree length

$$L(T) = \sum_{i\neq j} \frac{d_{ij}}{2^{\tau_{ij}}},$$

where $\tau_{ij}$ denotes the topological distance (i.e., the number of edges) between leaves $i$ and $j$ in $T$.

UBTs play a central role in computational biology, especially in
emphcomputational phylogenetics. Among the many approaches to phylogenetic inference,
emphdistance-based methods are particularly suited to the analysis of molecular data. Molecular phylogenetics studies the hierarchical evolutionary relationships among species by analyzing DNA, RNA, amino acid, or codon sequences. These relationships are typically represented by a weighted tree, called a
emphphylogeny, whose leaves correspond to observed taxa, internal vertices to unobserved ancestors, edges to evolutionary connections, and edge weights to dissimilarity estimates between taxa [3] (see Figure 1).

Phylogenetic inference has become increasingly important in diverse scientific domains, including medical research, drug discovery, epidemiology, and population genetics [16]. Applications include the prediction of influenza evolution [2], the study of HIV virulence [17, 15], the identification of emerging pathogens such as SARS [14], the reconstruction of ancestral proteins [7], the design of neuropeptides [1], and the study of macroevolutionary processes [12].

The phylogeny $T$ of a given set $\Gamma$ of $n \geq 3$ taxa is encoded as a UBT with $n$ terminal nodes (taxa), $n - 2$ internal nodes (speciation events), and $2n - 3$ edges (see Figure 1). In practice, the true tree is unknown and must be inferred by solving a combinatorial optimization problem over the $(2n - 5)!!$ possible UBTs on $\Gamma$. The formulation of this problem depends on biological assumptions and methodological choices, but all known variants are $\mathcal{NP}$-hard [4]. In particular, Fiorini and Joret [8] proved that the BMEP is $\mathcal{APX}$-hard and inapproximable within any factor $c^n$, for some $c > 1$.

Despite this worst-case inapproximability, heuristic methods such as the widely-used
emphNeighbor-Joining algorithm [10] remain the only tractable option for moderate to large datasets (e.g., 25–30 taxa or more), and consistently perform well in practice [10].

In this note, we address this apparent contradiction. We show that the BMEP becomes polynomial-time approximable with respect to the ratio function

$$\rho(D) := \frac{\max_{i\neq j} d_{ij}}{\min_{i\neq j} d_{ij}},$$

under the assumption that $\min_{i\neq j} d_{ij} > 0$. Our main result is the following.

**Theorem 1.1.** *Let $D = (d_{ij})$ be a symmetric dissimilarity matrix with zero diagonal and strictly positive off-diagonal entries. Then the BMEP admits a polynomial-time approximation algorithm with an approximation ratio at most $2 + \rho(D)$.*

In particular, the BMEP is $2 + \rho$-approximable whenever $\rho(D)$ is bounded by a constant $\rho > 0$. This is often the case in practical applications, such as those involving codon or nucleotide sequences, where dissimilarity values are derived from sequence alignment scores or evolutionary models and tend to fall within narrow ranges.

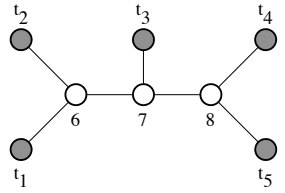
## 2. Preliminaries

An instance of the BMEP is said to be *metric* if the dissimilarity matrix $D = (d_{ij})$ satisfies the triangle inequality, that is,

$$d_{ij} + d_{jk} \geq d_{ik}, \quad \text{for all distinct } i, j, k \in \Gamma.$$

*Corresponding author
francesco.pisanu@uclouvain.be (F. Pisanu)
ORCID(s): 0000-0003-0799-5760 (F. Pisanu)

| Taxon | Sequence | | |
|---|---|---|---|
| $t_1$ | A | A | A |
| $t_2$ | A | C | C |
| $t_3$ | C | G | C |
| $t_4$ | C | C | G |
| $t_5$ | G | A | G |




$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{bmatrix} 0 & 2 & 3 & 3 & 2 \\ 2 & 0 & 2 & 2 & 3 \\ 3 & 2 & 0 & 2 & 3 \\ 3 & 2 & 2 & 0 & 2 \\ 2 & 3 & 3 & 2 & 0 \end{bmatrix} \end{array} = \{d_{ij}\} \qquad \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{bmatrix} 0 & 2 & 3 & 4 & 4 \\ 2 & 0 & 3 & 4 & 4 \\ 3 & 3 & 0 & 3 & 3 \\ 4 & 4 & 3 & 0 & 2 \\ 4 & 4 & 3 & 2 & 0 \end{bmatrix} \end{array} = \{\tau_{ij}\}$$

**Figure 1:** Above: an example of a set $\Gamma$ of 5 taxa and the associated hypothetical DNA sequences, with a possible phylogeny of $\Gamma$ encoded as an *Unrooted Binary Tree* (UBT) on the right. Below: an example of a *distance matrix* associated with $\Gamma$, obtained by assuming the Hamming distance as a measure of similarity between any pair of sequences, along with the corresponding path-length matrix.

In the same work where they established the general inapproximability of the BMEP, Fiorini and Joret [8] also showed that in the case of metric dissimilarities, the problem becomes approximable within a constant factor.

**Theorem 2.1 (Fiorini and Joret [8]).** *The BMEP admits a 2-approximation algorithm when restricted to metric instances.*

The algorithm of Fiorini and Joret proceeds as follows. Given a metric dissimilarity matrix $D$, one first computes a *minimum spanning tree* (MST) over the complete graph with vertex set $\Gamma$ and edge weights $d_{ij}$. This tree captures the minimum total dissimilarity needed to connect all taxa. Then, a phylogenetic tree is derived from the MST by recursively replacing each internal node of degree greater than three with a binary subtree, ensuring that the resulting structure is a valid UBT.

The core idea behind the approximation guarantee lies in bounding the BME cost using properties of the MST and the Kraft inequality.

**Theorem 2.2 (Catanzaro et al. [5]).** *Let $T$ be a UBT with leaf-set $\Gamma$. Then for every leaf $i$, we have*

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

This result, adapted from the well-known *Kraft inequality* for binary trees [13], plays a fundamental role in bounding the contribution of added distances during the reduction to a metric instance.

In the next section, we leverage these results to prove that the BMEP is polynomial-time approximable under a bounded-ratio condition on the entries of the dissimilarity matrix.

## 3. Proof of Theorem 1.1

We prove that any instance of the BMEP with a dissimilarity matrix $D$ having strictly positive entries admits a polynomial-time approximation algorithm with approximation ratio at most $2 + \rho(D)$, where

$$\rho(D) := \frac{\max_{i\neq j} d_{ij}}{\min_{i\neq j} d_{ij}}.$$

Let $D = (d_{ij})$ be a BMEP instance of size $n$, with optimal tree $T^*$ and cost $OPT := c(T^*) = \sum_{i\neq j} \frac{d_{ij}}{2^{\tau_{ij}}}$. Define a new instance $D' = (d'_{ij})$ where:

$$d'_{ij} := d_{ij} + d^+, \quad \text{for all } i \neq j, \quad \text{and } d'_{ii} := 0,$$

and where $d^+ := \max_{i\neq j} d_{ij}$.

We now show that $D'$ is a metric matrix. Indeed, for all distinct $i, j, k \in \Gamma$,

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

since by definition $d^+ \geq d_{ik}$. Hence, $D'$ satisfies the triangle inequality and is metric.

We next analyze how the cost changes when applying the BME objective to $D'$. For any UBT $T$ on $\Gamma$, we have:

$$\begin{aligned} \sum_{i,j\in\Gamma} d'_{ij} 2^{-\tau_{ij}} &= \sum_{i,j\in\Gamma} (d_{ij} + d^+) 2^{-\tau_{ij}} \\ &= \sum_{i,j\in\Gamma} d_{ij} 2^{-\tau_{ij}} + d^+ \sum_{i,j\in\Gamma} 2^{-\tau_{ij}} \\ &= \sum_{i,j\in\Gamma} d_{ij} 2^{-\tau_{ij}} + d^+ \sum_{i\in\Gamma} \left( \sum_{j\in\Gamma} 2^{-\tau_{ij}} \right) \\ &= \sum_{i,j\in\Gamma} d_{ij} 2^{-\tau_{ij}} + d^+ \cdot \frac{n}{2}, \end{aligned} \tag{1}$$

where the last equality follows from Theorem 2.2.

In particular, the optimal tree $T^*$ for the original instance $D$ is also optimal for $D'$, with cost

$$OPT' := c'(T^*) = OPT + d^+ \cdot \frac{n}{2}.$$

By Theorem 2.1, we can find in polynomial time a feasible solution $T'$ for $D'$ such that:

$$c'(T') \leq 2 \cdot OPT'.$$

Then, by Equation (1), we get:

$$c(T') = \sum_{i,j} d_{ij} 2^{-\tau_{ij}(T')} = c'(T') - d^+ \cdot \frac{n}{2} \leq 2 \cdot OPT + d^+ \cdot \frac{n}{2}. \tag{2}$$

Now, let $d^- := \min_{i \neq j} d_{ij} > 0$. Using Theorem 2.2 again, we observe that:

$$OPT = c(T^*) \geq \sum_{i,j} d^- 2^{-\tau_{ij}} = d^- \cdot \frac{n}{2}.$$

Substituting this into Equation (2), we obtain:

$$\frac{c(T')}{OPT} \leq 2 + \frac{d^+ \cdot \frac{n}{2}}{OPT} \leq 2 + \frac{d^+ \cdot \frac{n}{2}}{d^- \cdot \frac{n}{2}} = 2 + \frac{d^+}{d^-} = 2 + \rho(D).$$

Finally, the metric instance $D'$ can be computed in $\mathcal{O}(n^2)$, and the 2-approximate solution for $D'$ in $\mathcal{O}(n^3)$ [8]. Thus, the entire algorithm runs in polynomial time and achieves approximation ratio at most $2 + \rho(D)$. □

## 4. Practical consequences of Theorem 1.1

Theorem 1.1 shows that the BMEP is efficiently approximable whenever the dissimilarity matrix $D$ has a bounded ratio $\rho(D)$. This result is particularly relevant in practice, where dissimilarity measures are derived from biological data and are rarely pathological.

In real-world applications, it is reasonable to assume that all dissimilarities are strictly positive, i.e., $d_{ij} > 0$ for all $i \neq j$, since distinct taxa typically exhibit at least some divergence. Moreover, numerical values for $d_{ij}$ are subject to finite precision due to data acquisition methods or floating-point representation, which naturally bounds $\rho(D)$ by a polynomial function of the input size.

We formalize this observation as follows:

**Corollary 4.1.** *If $\rho(D)$ is bounded by a polynomial in $n$, then the BMEP admits a polynomial-time approximation algorithm with polynomial approximation guarantee.*

*Example: Hamming distance.* Many classical approaches to molecular phylogenetics rely on Hamming distances between aligned sequences. See, e.g., [6, 11, 9]. Assuming that no two taxa share identical sequences, the minimum Hamming distance is at least 1, and the maximum distance $d^+$ is bounded by the sequence length $k$. Therefore:

$$\rho(D) \leq d^+ \leq k.$$

This implies:

**Corollary 4.2.** *If $D$ is based on Hamming distances over sequences of length $k$, the BMEP is $2 + k$-approximable.*

*Example: Gamma-distributed dissimilarities.* In [18], the authors propose to model pairwise dissimilarities using a Gamma distribution with shape parameter $\alpha$ and rate $\beta$. In such models,

$$\rho(D) = \left(\frac{x}{y}\right)^{\alpha - 1} e^{-\beta(x-y)},$$

where $x$, and $y$ are the argmax and argmin of the samples' measure. For many parameter settings used in practice (e.g., $\alpha = 2$, $\beta = 1$, and moderate dispersion), $\rho(D)$ is typically bounded by a small constant.

**Corollary 4.3.** *Under Gamma-distributed dissimilarities with biologically plausible parameters, the BMEP is constant-factor approximable.*

These examples illustrate that Theorem 1.1 bridges the gap between worst-case inapproximability and practical tractability.

## Conclusions

Theorem 1.1 provides a theoretical explanation for the consistent empirical performance of heuristics such as Neighbor-Joining, despite the worst-case inapproximability of the BMEP. In particular, while Fiorini and Joret [8] show that the BMEP is $\mathcal{APX}$-hard and inapproximable within $c^n$ for some $c > 1$, their reduction relies on sparse dissimilarity matrices in which the minimum entry is zero. In such cases, $\rho(D)$ becomes unbounded. However, the assumption $d_{ij} = 0$ is biologically implausible, as it implies that taxa $i$ and $j$ are genetically identical. Our result complements this negative result by showing that BMEP instances with strictly positive and reasonably bounded dissimilarities—typical in molecular phylogenetics—are indeed efficiently approximable. This encourages further algorithmic investigations into robust and practical approximation algorithms for the BMEP, which remains central to computational phylogenetics.

## References


[1] D. A. Bader, B. M. E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In *SPIE ITCom: Commercial application for high-performance computing*, pages 159–168. SPIE, Bellingham, WA, July 2001.

[2] 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, December 1999.

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

[4] 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.

[5] D. Catanzaro, M. Labbé, R. Pesenti, and J.J. Salazar-González. The balanced minimum evolution problem. *INFORMS J. Comput.*, 24:276–294, 2012.

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

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

[8] S. Fiorini and G. Joret. Approximating the balanced minimum evolution problem. *Oper. Res. Lett.*, 40:31–35, 2011.

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

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

[11] Gaston H Gonnet, Chantal Korostensky, and Steve Benner. Evaluation measures of multiple sequence alignments. *Journal of Computational Biology*, 7(1-2):261–276, 2000.

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

[13] Leon Gordon Kraft. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949.

[14] 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, May 2003.

[15] Chin-Yih Ou, Carol A Ciesielski, Gerald Myers, Claudiu I Bandea, Chi-Cheng Luo, Bette TM Korber, James I Mullins, Gerald Schochetman, Ruth L Berkelman, A Nikki Economou, et al. Molecular epidemiology of hiv transmission in a dental practice. *Science*, 256(5060):1165–1171, 1992.

[16] Lior Pachter and Bernd Sturmfels. The mathematics of phylogenomics. *SIAM review*, 49(1):3–31, 2007.

[17] Howard A Ross and Allen G Rodrigo. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *Journal of virology*, 76(22):11715–11720, 2002.

[18] Ana Helena Tavares, Jakob Raymaekers, Peter J Rousseeuw, Paula Brito, and Vera Afreixo. Clustering genomic words in human dna using peaks and trends of distributions. *Advances in Data Analysis and Classification*, 14(1):57–76, 2020.