On the approximability of the fixed-tree balanced minimum evolution problem

Martin Frohn¹

Received: 13 April 2020 / Accepted: 30 November 2020 / Published online: 2 January 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

The *Fixed-Tree BMEP (FT-BMEP)* is a special case of the *Balanced Minimum Evolution Problem (BMEP)* that consists of finding the assignment of a set of n taxa to the n leaves of a given unrooted binary tree so as to minimize the BMEP objective function. Deciding the computational complexity of the FT-BMEP has been an open problem for almost a decade. Here, we show that a few modifications to Fiorini and Joret's proof of the $\mathcal{NP}$ -hardness of the BMEP suffice to prove the general $\mathcal{NP}$ -hardness of the FT-BMEP as well as its strong inapproximability.

Keywords Fixed-tree balanced minimum evolution problem · Computational complexity · Phylogenetics

1 Introduction

A *phylogeny* is a weighted tree that describes the hierarchical evolutionary relationships of a given set of species (also referred to as *taxa*), based on their observed inherited traits (e.g., DNA, RNA, codon sequences, or whole genomes) [1–5]. The topology of a phylogeny and its corresponding biological meaning may depend on the specific application or use [6]. For example, in the context of tumor evolution, a phylogeny can be represented as an arborescence that connects the sampled tumor clones to the healthy one (see e.g., [7–9]). In systematics, instead, a phylogeny is encoded as an *Unrooted Binary Tree (UBT)* in which the terminal vertices (or *leaves*) represent the observed taxa; *internal vertices* represent speciation events occurred throughout evolution of taxa; *edges* represent estimated evolutionary relationships; and *edge weights* represent measures of the similarity between pairs of taxa [10]. As an example, Fig. 1 shows the phylogeny (in the systematic context) of a set of eight biosafety level 2, 3

✉ Martin Frohn
martin.frohn@uclouvain.be

¹ CORE, Université Catholique de Louvain, Voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium

Fig. 1 An example of a phylogeny of a set of eight taxa (red vertices), including the whole genomes of the Crimean-Congo Hemorrhagic Fever (CCHF) orthonairovirus, Ebolavirus, the Lassa mammarenavirus, Yersinia Pestis (Y-Pestis), the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the Human Immunodeficiency Viruses (HIVs) 1 and 2, and the Nipah virus. The internal vertices of the phylogeny are marked in blue. Edge weights have been removed for the sake of readability. The above complete genomes are available at GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) via the reference numbers GCF_000854165.1, NC_002549.1, GCF_000851705.1, NC_004777.1, NC_045512.2, NC_001802.1, NC_001722.1, and NC_002728.1, respectively

Fig. 2 An example of an instance of the BMEP, including a set $\Gamma = \{t_1, t_2, t_3, t_4, t_5\}$ of five taxa and the corresponding input distance matrix D

$$D = \begin{matrix} & \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 \end{matrix} \\ \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} & \begin{bmatrix} 0 & 2 & 3 & 3 & 2 \\ 2 & 0 & 2 & 2 & 3 \\ 3 & 2 & 0 & 2 & 3 \\ 3 & 2 & 2 & 0 & 2 \\ 2 & 3 & 3 & 2 & 0 \end{bmatrix} \end{matrix}$$

and 4 pathogens (taxa), based on the knowledge of their respective complete genomes (see caption for more details).

Consider a set $\Gamma = \{1, 2, \dots, n\}$ of $n \geq 3$ taxa and a $n \times n$ symmetric distance matrix D whose generic entry d_{ij} represents a measure of the dissimilarity between the pair of taxa $i, j \in \Gamma$. Each entry d_{ij} is nonnegative and equal to zero on the main diagonal of D . Then, the *Balanced Minimum Evolution Problem* (BMEP) consists of finding a phylogeny T of Γ — i.e., a pair constituted by (i) an *Unrooted Binary Tree* (UBT) having n leaves and (ii) a bijection between the set of leaves of this tree and the taxa in Γ — so as to minimize the following *length function*

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

where τ_{ij} represents the *topological distance* between taxa i and j , i.e., the number of edges belonging to the (unique) path in T connecting taxon i to taxon j [11,12]. Figures 2 and 3 provide an example of an instance of the BMEP and the corresponding optimal solution, respectively.

An optimal solution to the BMEP (i.e., an *optimal phylogeny*) provides an estimation of the hierarchical evolutionary relationships of a given set of biological entities (i.e., taxa) based on a measure of the dissimilarity between pairs of taxa (i.e., the distances) [1,3,5,13]. These relationships can be reinterpreted as the cross-entropy

Fig. 3 A possible phylogeny T for the set Γ provided in Fig. 2. Note that each input taxon in Γ is assigned precisely to a leaf of T and that each internal vertex of T has degree three (i.e., T is unrooted and binary). The topological distance between the pair of taxa t_1 and t_3 is $\tau_{13} = 4$. Similarly, the topological distance between the pair of taxa t_2 and t_3 is $\tau_{23} = 3$. In the context of the BMEP, T is provably optimal for the input distance matrix D shown in Fig. 2

minimization of the information related to the molecular sequences extracted from taxa (see [14] for details). The BMEP can be solved in polynomial-time if the input distance matrix D is *additive*, i.e., if its entries satisfy the following condition [4]:

$$d_{ij} + d_{kr} \leq \max\{d_{ik} + d_{jr}, d_{ir} + d_{jk}\} \quad \forall i, j, k, r \in \Gamma. \quad (2)$$

Unfortunately, if D is a generic matrix, then the BMEP becomes $\mathcal{NP}$ -hard and inapproximable within c^n , for some positive constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$ [15]. If the input distance matrix D is just *metric*, i.e., if its entries satisfy the triangle inequality, then the optimal solution to the BMEP can be approximated within a factor of two [15].

The BMEP was introduced in the literature on molecular phylogenetics by Desper and Gascuel [16], based on a phylogenetic estimation model proposed by Pauplin [17] about 20 years ago. It was subsequently the object of thorough studies carried out in both the computational biology and the operations research communities [18–27]. In particular, the biological interpretation as well as the statistical consistency properties of the BMEP have been investigated in Gascuel [4], whereas the computational and combinatorial aspects have been deepened in Aringhieri et al. [25], Catanzaro et al. [11, 14, 28] and Forcey et al. [29, 30], Catanzaro et al. [31], Catanzaro and Pesenti [32], respectively. A recent survey on the state-of-the-art on the BMEP can be found in [33].

This letter addresses an open theoretical question related to the work of Aringhieri et al. [25] and concerning the computational complexity of a particular version of the BMEP that consists of finding an optimal assignment of taxa to the leaves of a fixed UBT so as to minimize the length function (1). We refer to this problem as the *Fixed-Tree Balanced Minimum Evolution Problem* (FT-BMEP). We show here that a few modifications to Fiorini and Joret's proof of the $\mathcal{NP}$ -hardness of the BMEP suffice to prove the general $\mathcal{NP}$ -hardness of the FT-BMEP as well as its strong inapproximability. For the sake of clarity and completeness, we will present the whole proof in the remainder of this letter.

2 On the complexity of the taxa assignment problem on a fixed unlabeled phylogeny

We denote $\Pi(n)$ as the set of permutations of a given set of n elements and $\mathbb{R}_{0+}$ as the set of the non-negative real numbers. Given a set Γ of $n \geq 3$ taxa and a taxon $i \in \Gamma$,

Fig. 4 An example of an unlabeled phylogeny of five taxa

we denote Γ_i as the set $\Gamma \setminus \{i\}$. We say that a phylogeny T of Γ is *unlabeled* when the input set of taxa has not yet been assigned to its leaves. In other words, an unlabeled phylogeny of Γ is just an UBT with n leaves (see e.g., Fig. 4). We say that two unlabeled phylogenies T_1 and T_2 are *isomorphic* if there exists a graph isomorphism between T_1 and T_2 , i.e., a bijection ρ from the vertex set of T_1 to the vertex set of T_2 such that two vertices, say u and v , are adjacent in T_1 if and only if $\rho(u)$ and $\rho(v)$ are adjacent in T_2 . We denote $\mathcal{T}$ as the set of the $(2n - 5)!!$ possible phylogenies of Γ and $\mathcal{T}_U$ as the set of the possible unlabeled non-isomorphic phylogenies of Γ [1]. Finally we recall that the phylogenies of Γ satisfy the following *Kraft equality* [11]:

$$\sum_{j \in \Gamma_i} \frac{1}{2^{\tau_{ij}}} = \frac{1}{2} \quad \forall i \in \Gamma. \quad (3)$$

In the light of the above notation and definitions, in this section we investigate the computational complexity of the following problem:

Problem 1 (The Fixed-Tree Balanced Minimum Evolution Problem [FTBMEP]) Given a positive integer $n \geq 3$, a set $\Gamma = \{1, 2, \dots, n\}$ of n taxa, a symmetric distance matrix $D \in \mathbb{R}_{0+}^{n \times n}$, and a fixed unlabeled phylogeny $T \in \mathcal{T}_U$, find a permutation $\pi^* \in \Pi(n)$ such that

$$\pi^* = \arg \min_{\pi \in \Pi(n)} \left\{ L_{\pi}(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} d_{ij} 2^{-\tau_{\pi(i)\pi(j)}} \right\}.$$

To determine the computational complexity of the FT-BMEP we consider the following decision problem, hereafter referred to as the *Fixed-Tree Balanced Assignment Problem* (FT-BAP):

Problem 2 (The Fixed-Tree Balanced Assignment Problem [FTBAP]) Given a positive constant B , a positive integer $n \geq 3$, a set $\Gamma = \{1, 2, \dots, n\}$ of n taxa, a symmetric distance matrix $D \in \mathbb{R}_{0+}^{n \times n}$, and a fixed unlabeled phylogeny $T \in \mathcal{T}_U$, is there a permutation $\hat{\pi} \in \Pi(n)$ such that $L_{\hat{\pi}}(T) \leq B$?

We will show that the FT-BAP is $\mathcal{NP}$ -complete and inapproximable within a constant factor unless $\mathcal{P} = \mathcal{NP}$. The $\mathcal{NP}$ -hardness of the FT-BMEP will then follow as a direct consequence of this result. Similarly to Fiorini and Joret [15], in our proof we will use a reduction from the following $\mathcal{NP}$ -Complete decision problem [34]:

Problem 3 (The 3-Colorability Problem [3CP]) Given an undirected graph $G = (V, E)$, can V be partitioned into three stable sets, i.e. sets in which no two vertices are adjacent in G ?

Denoting $(\Gamma, D, T, B)_{FT-BAP}$ and $G = (V, E)_{3CP}$ as an instance of FT-BAP and an instance of 3CP, respectively, the following proposition holds:

Proposition 1 *The FT-BAP is NP-Complete.*

Proof Given any instance $G = (V, E)_{3CP}$ of the 3CP we show that we can construct in polynomial-time an instance $(\Gamma, D, T, B)_{FT-BAP}$ of the FT-BAP such that the following claim holds true:

The graph $G = (V, E)_{3CP}$ is 3-colorable $\iff$

$$\exists \hat{\pi} \in \Pi(n) : L_{\hat{\pi}}(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\tau_{\hat{\pi}}(i)\hat{\pi}(j)}} \leq B. \quad (4)$$

where $n = |\Gamma|$. To this end, we set $p := |V|$ and $m := |E|$. Let λ be an arbitrary constant such that $\frac{2}{3} - \frac{\log_2(p(p-1)/2)}{p} < \lambda < \frac{2}{3}$. We remark that, for $p \geq 218$, we have $\frac{3}{5} < \lambda < \frac{2}{3}$ and $m \leq p(p-1)/2 \leq 2^{(2/3-\lambda)}$. This assumption is without loss of generality as we are interested in proving condition (4) asymptotically. We also denote k as the smallest positive integer satisfying $k \geq 3p/(2\lambda - 1)$ and $k \equiv 0 \pmod{3}$. Finally, let $V = \{v_1, v_2, \dots, v_p\}$. Now, we define an instance of the FT-BAP as follows. We set $n := 3p + k$, $\Gamma := \{1, 2, \dots, n\}$, and we associate the first p taxa in Γ with the corresponding vertices of $G = (V, E)_{3CP}$ so that whenever we consider taxon $i \leq p$ in Γ we also refer implicitly to the corresponding vertex v_i in $G = (V, E)_{3CP}$ and vice versa. This means that any permutation that assigns the taxa in Γ to the leaves of T is defined by the following two properties: (i) assign taxa $\{1, \dots, p\}$ to p leaves of T ; (ii) assign taxa in Γ not corresponding to vertices in G to the remaining $n - p$ leaves of T . We also define the generic entry d_{ij} of the distance matrix D as

$$d_{ij} := \begin{cases} 1 & \text{if } \max\{i, j\} \leq p \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \quad \forall i, j \in \Gamma.$$

Finally, we set the constant $B := 2^{1-\lambda k - (2/3-\lambda)(7-8\lambda)n/3}$ and we construct an unlabeled phylogeny $T \in \mathcal{T}_U$ as in Fig. 5, i.e., we join an internal vertex v , hereafter referred to as *centroid* [35], with three rooted subtrees, referred to as *subcaterpillars* T_1, T_2 and T_3 , each containing $p + k/3$ leaves. It is easy to see that the construction process of $(\Gamma, D, T, B)_{FT-BAP}$ can be carried out in polynomial-time and that it is valid because $k \equiv 0 \pmod{3}$.

In order to prove claim (4) we first show the following intermediate result:

$$\begin{aligned} &\text{If } \exists \hat{\pi} \in \Pi(n) : \tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k \text{ for every edge } (v_i, v_j) \in E, \\ &\text{then } G = (V, E)_{3CP} \text{ is 3-colorable.} \end{aligned} \quad (5)$$

To prove (5), consider any edge $(v_i, v_j) \in E$ such that $\tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k$. Then, the considered permutation $\hat{\pi}$ must assign i and j to leaves located in distinct subcaterpillars $T_l, l \in \{1, 2, 3\}$, otherwise one has

Fig. 5 An example of an unlabeled phylogeny with a centroid v and three subcaterpillars

$$\tau_{\hat{\pi}(i)\hat{\pi}(j)} \leq p + \frac{k}{3} \leq \frac{2}{3}\lambda k < \lambda k,$$

where the first inequality is derived from the maximal distance between two leaves in a subcaterpillar T_l , $l \in \{1, 2, 3\}$, and the second inequality is a reformulation of our assumption $k \geq 3p/(2\lambda - 1)$. This fact, however, contradicts the hypothesis of having $\tau_{\hat{\pi}(i)\hat{\pi}(j)} > \lambda k$. Therefore, the sets S_l of taxa assigned to the leaves of the subcaterpillars T_l , $l \in \{1, 2, 3\}$ induce a partition of the vertices of $G = (V, E)_{3CP}$ into three stable sets. Thus, $G = (V, E)_{3CP}$ is 3-colorable and the statement of claim (5) follows.

It is worth noting that if $G = (V, E)_{3CP}$ is not 3-colorable then, by contraposition, for any permutation $\pi \in \Pi(n)$, there exists at least one pair of adjacent vertices in $G = (V, E)_{3CP}$, say v_s and v_l , such that $\tau_{\pi(s)\pi(l)} \leq \lambda k$. As a consequence, for any permutation $\pi \in \Pi(n)$,

$$L_\pi(T) = 2 \cdot \sum_{(v_i, v_j) \in E} \frac{1}{2^{\tau_{\pi(i)\pi(j)}}} \geq 2 \cdot 2^{-\tau_{sl}} \geq 2^{1-\lambda k}.$$

Since $B < 2^{1-\lambda k}$ by our choice of B , we can deduce that if there exists a permutation $\pi \in \Pi(n)$ with $L_\pi(T) \leq B$ then $G = (V, E)_{3CP}$ is certainly 3-colorable.

Conversely, assume that $G = (V, E)$ is 3-colorable. Then, let S_1 , S_2 and S_3 denote the sets constituting the tripartition of V induced by the 3-coloration. Moreover, consider the following permutation $\pi \in \Pi(n)$ that assigns the taxa in Γ to the leaves of T :

- for each $l \in \{1, 2, 3\}$, assign arbitrarily the taxa corresponding to the vertices in S_l to the leaves of the l th subcaterpillar of T that are farthest from the centroid (see Fig. 5);
- assign the remaining $2p + k$ taxa arbitrarily to the remaining leaves.

Fig. 6 An example of an unlabeled phylogeny with a centroid v and two subcaterpillars (up and left) and a generic subtree (right). The gray triangles represent two generic rooted binary trees

By construction, for any $(v_i, v_j) \in E$, it holds that $\tau_{\pi(i)\pi(j)} > 2k/3$ (see, again, Fig. 5). Moreover, by recalling that $m = |E|$, it also holds that

$$L_{\pi}(T) = 2 \cdot \sum_{(v_i, v_j) \in E} \frac{1}{2^{\tau_{\pi(i)\pi(j)}}} < \sum_{(v_i, v_j) \in E} 2^{1-2k/3} = 2^{1-2k/3} m.$$

Now, observe that because $n = k + 3p$, $m \leq 2^{(2/3-\lambda)p}$ and $|V| = p \leq (2\lambda - 1)k/3 \leq (2\lambda - 1)n/3$, it holds that

$$\begin{aligned} \frac{2^{1-\lambda k}}{m \cdot 2^{1-2k/3}} &= \frac{2^{(2/3-\lambda)k}}{m} \geq 2^{(2/3-\lambda)(k-p)} \\ &= 2^{(2/3-\lambda)(n-4p)} \geq 2^{(2/3-\lambda)(7-8\lambda)n/3} \end{aligned} \quad (6)$$

which is equivalent to $2^{1-2k/3} m \leq 2^{1-\lambda k - (2/3-\lambda)(7-8\lambda)n/3} = B$. In other words, the length function satisfies $L_{\pi}(T) \leq B$. This completes the proof of claim (4), and the statement of the proposition follows. $\square$

Proposition 1 shows that there exists at least one specific unlabeled phylogeny $T \in \mathcal{T}_U$ for which the FT-BAP is $\mathcal{NP}$ -complete. Indeed, there exist exponentially many other unlabeled phylogenies in $\mathcal{T}_U$ for which a similar result holds. For example, all of the unlabeled phylogenies that can be obtained from the tree shown in Fig. 5 by arbitrarily rearranging the topology connecting the p leaves in each target subcaterpillar T_l (see, e.g., Fig. 6). Moreover, it is easy to realize that by adjusting appropriately the values of k and λ in the above proof, the $\mathcal{NP}$ -completeness still persists also for unlabeled phylogenies characterized by a *bicentroid* instead of a centroid, i.e., a unique edge whose removal decomposes T into two subtrees containing roughly $n/2$ taxa each. Note that, according to the following result, a tree always contains either a centroid or a bicentroid:

Proposition 2 [35] *Let T be a tree with n vertices.*

1. *If $n = 2k + 1$ for some $k \in \mathbb{N}$, then there exists a unique vertex c in T , called the centroid, such that all (two or more) subtrees obtained by removing c contain at most k vertices.*
2. *If $n = 2k$ for some $k \in \mathbb{N}$, then there exists in T either*
  - (a) *a unique vertex c , called the centroid, such that all (three or more) subtrees obtained by removing c contain less than k vertices, or*
  - (b) *a unique edge b , called the bicentroid, such that the two subtrees obtained by removing b contain exactly k vertices.*

The following result completes our study:

Proposition 3 *There exists a constant $c > 1$ such that FT-BAP has no c^n -approximation algorithm unless $\mathcal{P} = \mathcal{NP}$.*

Proof Consider again a 3CP instance $(V, E)_{3CP}$ and reduce this instance to $(\Gamma, D, T, B)_{FT-BAP}$ as described in the previous proof. In addition, set

$$c := 2^{(2/3-\lambda)(7-8\lambda)/3} > 1$$

Then, it follows from inequality (6) that a c^n -approximation algorithm for the FT-BAP could be used to decide whether G is 3-colorable or not. $\square$

An interesting open question is whether there exist instances of the FT-BMEP that can be ε -approximated, for some $\varepsilon > 0$, when dealing with particular types of input distances matrices D . This question further adds to the ones discussed in [15] and definitely warrants additional research effort.

Acknowledgements The author thanks Prof. Daniele Catanzaro, Prof. Raffaele Pesenti and Prof. Maurice Quayranne for fundamental feedbacks and helpful discussions. Part of this work has been developed when M. Frohn was Visiting Scholar at the Sauder School of Business of the University of British Columbia, Canada.

References

1. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Sunderland (2004)
2. Semple, C., Steel, M.: Phylogenetics. Oxford University Press, New York (2003)
3. Page, R.D.M., Holmes, E.C.: Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford (1998)
4. Gascuel, O.: Mathematics of Evolution and Phylogeny. Oxford University Press, New York (2005)
5. Catanzaro, D.: Estimating phylogenies from molecular data. In: Bruni, R. (ed.) Mathematical Approaches to Polymer Sequence Analysis and Related Problems, pp. 149–176. Springer, Berlin (2011)
6. Grinstaff, G., Owen, M.: Representations of partial leaf sets in phylogenetic tree space. SIAM J. Appl. Algebra Geom. **3**(4), 691–720 (2019)
7. Catanzaro, D., Schackney, S.E., Schäffer, A.A., Schwartz, R.: Classifying the progression of Ductal Carcinoma from single-cell sampled data via integer linear programming: a case study. IEEE/ACM Trans. Comput. Biol. Bioinf. **13**(4), 643–655 (2016)
8. Beerenwinkel, N., Schwarz, R.F., Gerstung, M., Markowetz, F.: Cancer evolution: mathematical models and computational inference. Syst. Biol. **64**(1), e1–e25 (2015)

9. Pennington, G., Smith, C.A., Shackney, S., Schwartz, R.: Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol.* **5**(2a), 407–427 (2006)
10. Catanzaro, D.: The minimum evolution problem: overview and classification. *Networks* **53**(2), 112–125 (2009)
11. Catanzaro, D., Labbé, M., Pesenti, R., Salazar-González, J.J.: The balanced minimum evolution problem. *INFORMS J. Comput.* **24**(2), 276–294 (2012)
12. Pardi, F.: Algorithms on Phylogenetic Trees. University of Cambridge, Cambridge (2009). Ph.D. thesis
13. Warnow, T.: Bioinformatics and Phylogenetics. Springer, New York (2019)
14. Catanzaro, D., Frohn, M., Pesenti, R.: An information theory perspective on the balanced minimum evolution problem. *Oper. Res. Lett.* **48**(3), 362–367 (2020a)
15. Fiorini, S., Joret, G.: Approximating the balanced minimum evolution problem. *Oper. Res. Lett.* **40**(1), 31–35 (2012)
16. Desper, R., Gascuel, O.: Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *J. Comput. Biol.* **9**(5), 687–705 (2002)
17. Pauplin, Y.: Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.* **51**, 41–47 (2000)
18. Cueto, M.A., Matsen, F.A.: Polyhedral geometry of phylogenetic rogue taxa. *Bull. Math. Biol.* **73**(6), 1202–1226 (2011)
19. Desper, R., Gascuel, O.: Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting. *Mol. Biol. Evol.* **21**(3), 587–598 (2004)
20. Gascuel, O., Steel, M.: Neighbor-joining revealed. *Mol. Biol. Evol.* **23**(11), 1997–2000 (2006)
21. Haws, D.C., Hodge, T.L., Yoshida, R.: Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. *Bull. Math. Biol.* **73**(11), 2627–2648 (2011)
22. Semple, C., Steel, M.: Cyclic permutations and evolutionary trees. *Adv. Appl. Math.* **32**(4), 669–680 (2004)
23. Eickmeyer, K., Huggins, P., Pachter, L., Yoshida, R.: On the optimality of the neighbor-joining algorithm. *Algorithms Mol. Biol.* **3**(5), 1–11 (2008)
24. Bordewich, M., Gascuel, O., Huber, K., Moulton, V.: Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE Trans. Comput. Biol. Bioinform.* **6**(1), 110–117 (2009)
25. Aringhieri, R., Catanzaro, D., Di Summa, M.: Optimal solutions for the balanced minimum evolution problem. *Comput. Oper. Res.* **38**, 1845–1854 (2011)
26. Yasui, N., Vogiatzis, C., Yoshida, R., Fukumizu, K.: imPhy: Imputing phylogenetic trees with missing information using mathematical programming. *IEEE Trans. Comput. Biol. Bioinform.* **17**(4), 1222–1230 (2018)
27. Durell, C., Forcey, S.: Level-1 phylogenetic networks and their balanced minimum evolution polytopes, Technical report, Department of Mathematics, University of Akron, Akron, OH (2019)
28. Catanzaro, D., Labbé, M., Pesenti, R.: The balanced minimum evolution problem under uncertain data. *Discrete Appl. Math.* **161**(13–14), 1789–1804 (2013)
29. Forcey, S., Keefe, L., Sands, W.: Facets of the balanced minimal evolution polytope. *Math. Biol.* **73**, 447–468 (2016)
30. Forcey, S., Keefe, L., Sands, W.: Split-facets for balanced minimal evolution polytopes and the permutoassociahedron. *Bull. Math. Biol.* **79**, 975–994 (2017)
31. Catanzaro, D., Pesenti, R., Wolsey, L.A.: On the balanced minimum evolution polytope. *Discrete Optim.* **36**, 1–33 (2020b)
32. Catanzaro, D., Pesenti, R.: Enumerating vertices of the balanced minimum evolution polytope. *Comput. Oper. Res.* **109**, 209–217 (2019)
33. Catanzaro, D., Frohn, M., Gascuel, O., Pesenti, R.: Twenty years of balanced minimum evolution, Technical report 06-2020, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain (2020c)
34. Garey, M.R., Johnson, D.S.: Computers and Intractability: a guide to the theory of NP-completeness. Freeman, New York (2003)
35. Jordan, C.: Sur les assemblages des lignes. *Journal für die reine und angewandte Mathematik* **70**, 185–190 (1869)