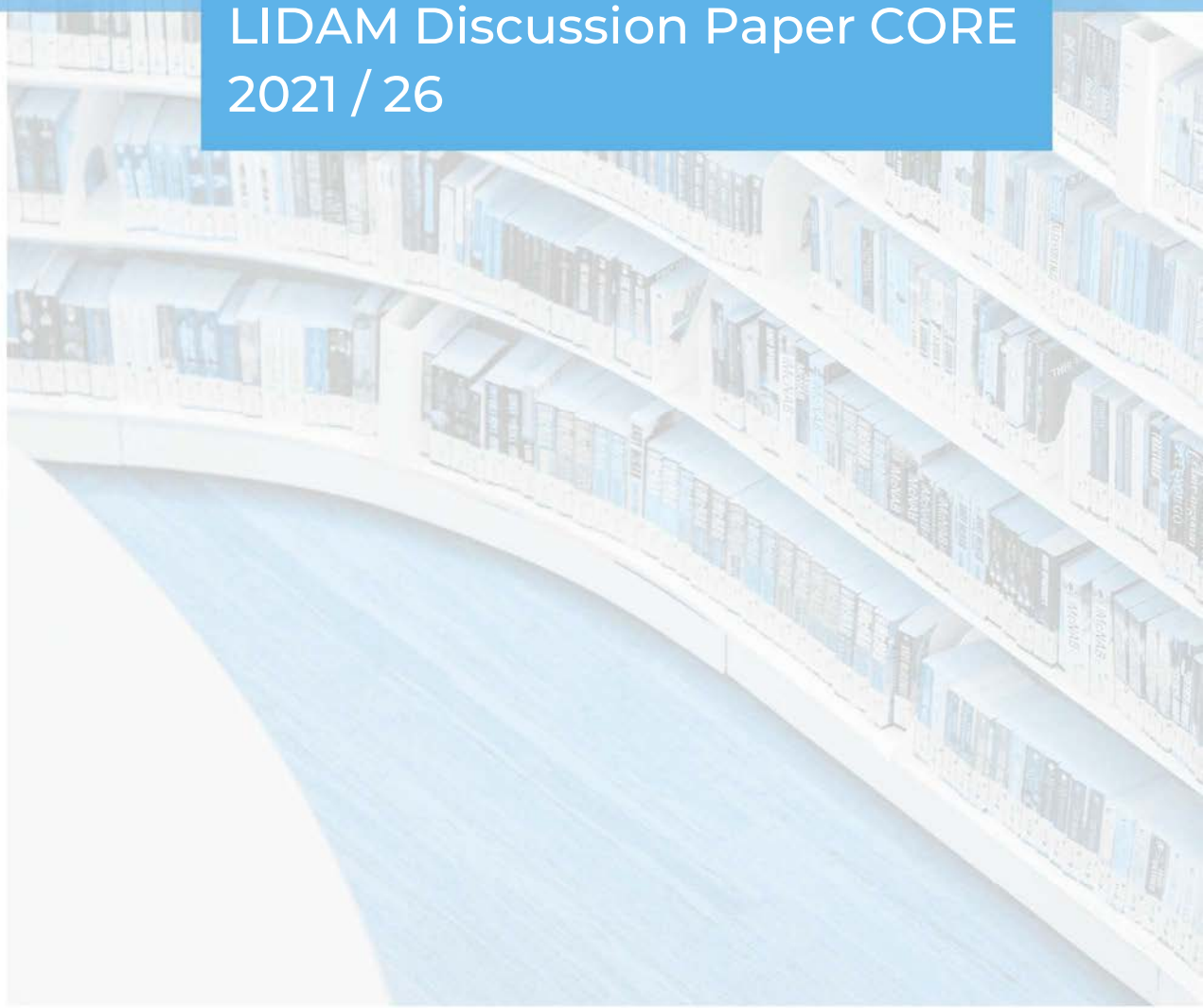
ON NUMERICAL STABILITY AND STATISTICAL CONSISTENCY OF THE BALANCED MINIMUM EVOLUTION PROBLEM

Daniele Catanzaro, Martin Frohn, Raffaele Pesenti

LIDAM Discussion Paper CORE
2021 / 26



UCLouvain

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

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/core-discussion-papers.html

On Numerical Stability and Statistical Consistency of the Balanced Minimum Evolution Problem

Daniele Catanzaro[a,*,1], Martin Frohn[a,1] and Raffaele Pesenti[c,3]

[a]Université Catholique de Louvain, Voie du Roman Pays 34, L1.03.01, B-1348 Louvain-la-Neuve, Belgium.
[c]Department of Management, Università Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venezia, Italy.

ARTICLE INFO

Keywords:
Combinatorial optimization; network design; balanced minimum evolution problem.

ABSTRACT

The *Balanced Minimum Evolution Problem* (BMEP) is affected by numerical instabilities that may preclude its use on practical instances. In this article, we investigate the impact of rescaling the objective function of the problem to overcome numerical instabilities and we show how this strategy may affect the statistical consistency of the optimal solution. In particular, we show that the numerical instabilities at the core of the BMEP cannot be overcome by any kind of rescaling of the objective function. As an extension of this negative result, we also characterize the class of the input distance matrices that may give rise to numerical instabilities for large scale instances.

1. Introduction

Consider a set $\Gamma = \{1, 2, \dots, n\}$ of $n \geq 3$ distinct aligned molecular sequences (e.g., DNA, RNA, or codon sequences), hereafter referred to as *taxa*, and a $n \times n$ symmetric distance matrix $\mathbf{D}$ whose generic entry d_{ij} – equal to zero on the main diagonal and strictly positive otherwise – encodes a measure of the dissimilarity between the pair of taxa $i, j \in \Gamma$ (see, e.g., Figure 1). 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 having n leaves and (ii) a bijection between these leaves and the taxa in Γ, so as to minimize the following *length function*

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

where τ_{ij} represents the *path-length* between taxa i and j in T, i.e., the number of edges belonging to the (unique) path in T connecting taxon i to taxon j [11, 43] (see Figure 2).

The BMEP was introduced in the literature on molecular phylogenetics by Desper and Gascuel [19], based on a particular phylogenetic edge weight estimation model proposed by Pauplin [47] about 21 years ago. The optimal phylogeny to the BMEP is proven to be statistically consistent [38] and determining it is a computationally less demanding task then solving the optimization problems related to alternative phylogeny estimation criteria such as the maximum parsimony or the maximum likelihood [5]. Moreover, since the BMEP belongs to the class of the phylogenetic estimation methods based on distance methods, it inherits from them the benefit of being able to handle virtually any kind of data for which a dissimilarity measure is available [50]. This makes the optimal phylogeny to the BMEP particularly suitable for general-purpose phylogenetic analyses.

Taxa	Sequence		
Taxon 1	A	A	A
Taxon 2	A	C	C
Taxon 3	C	G	C
Taxon 4	C	C	G
Taxon 5	G	A	G

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 3 & 3 & 2 \\ 2 & 0 & 2 & 2 & 3 \\ 3 & 2 & 0 & 2 & 3 \\ 3 & 2 & 2 & 0 & 2 \\ 2 & 3 & 3 & 2 & 0 \end{bmatrix} \end{array} = \mathbf{D}$$

Figure 1: A dummy example of a set Γ of five molecular sequences (*taxa*) and the corresponding distance matrix $\mathbf{D}$.

The BMEP can be seen as: (i) a restriction of the *Minimum Evolution Problem*, discussed in [6]; (ii) a particular version of the *network design problem* [35, 48] defined over the class of the UBTs; and (iii) a particular version of the *Steiner tree problem* in which the length function depends upon the topology of the tree, the tree must be unrooted an binary and the number of Steiner nodes is a priori known [15, 16, 22, 23, 31, 34, 39, 40, 49, 54]. The BMEP is in general $\mathcal{NP}$-hard and inapproximable within c^n, for some positive constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$ [25]. The problem is instead polynomially solvable if the input distance matrix $\mathbf{D}$ is *additive*, i.e., if its entries satisfy the following condi-

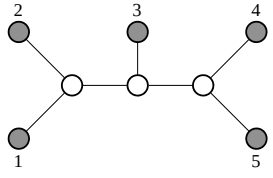


Figure 2: An example of a phylogeny T for the set Γ of five taxa of Figure 1. Each internal vertex of T has degree three (i.e., T is unrooted and binary) and each taxon in Γ is assigned precisely to a leaf of T. The topological distance between the pair of taxa 1 and 3, denoted as τ_{13}, is equal to four. Similarly, the topological distance between the pair of taxa 2 and 3, denoted as τ_{23}, is equal to three. T is provably optimal for the input distance matrix D shown in Figure 1.

*Corresponding author
ORCID(s): 0000-0001-9427-1562 (D. Catanzaro); 0000-0002-5002-4049 (M. Frohn); 0000-0001-5890-4238 (R. Pesenti)
[1]Email: daniele.catanzaroŽuclouvain.be (Daniele Catanzaro)
[2]Email: martin.frohnŽuclouvain.be (Martin Frohn)
[3]Email: pesentiŽunive.it (Raffaele Pesenti)

tion [28]:

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

In the case the input distance matrix $\mathbf{D}$ 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 [25]. The last 20 years registered extensive research efforts focused on the characterization of the statistical consistency of the BMEP [20, 28, 29], the study of the connections between the BMEP and information entropy [8], the study of its combinatorial aspects [11, 13, 18, 26, 27, 32, 51], and the development of implicit enumeration algorithms [1, 11, 43] and heuristics [25, 28, 43] to tackle and solve its instances. The reader interested in an introduction to the problem and on a review of the most important advances currently described in the literature may refers to [7].

Apart from the $\mathcal{NP}$-hardness, a second aspect that limits the practical use of the BMEP is the numerical instability of the length function, which materializes whenever the generic addend $d_{ij}2^{-\tau_{ij}}$ approaches (or get smaller than) the machine epsilon. This aspect has never been systematically investigated in the literature and, as we will see in this article, has deep ties with the statistical consistency of the optimal solution T^* to the problem. In particular, we show here that there exists a class of input distance matrices for the BMEP for which the difference between any two distinct phylogenies becomes infinitesimal for increasing values of n. This fact poses both numerical stability and statistical consistency issues for the BMEP and justifies the search for strategies that may overcome them. A possible strategy consists of rescaling the objective function in such a way to enable the analysis of large datasets that are of interest in practice [2, 3, 4, 17, 21, 29, 36, 38, 43, 44, 45, 46, 55]. Unfortunately, we will see that no rescaling technique may guarantee at the same time the numerical stability of the problem and the statistical consistency of the optimal solution to the rescaled problem.

2. On the numerical stability of the BMEP

As anticipated in the introduction, the numerical instability of the length function of the BMEP materializes whenever the generic addend $d_{ij}2^{-\tau_{ij}}$ approaches (or get smaller than) the machine epsilon ϵ, i.e., if there exist at least a pair of distinct taxa $i, j \in \Gamma$, such that

$$d_{ij} \cdot 2^{-\tau_{ij}} \leq \epsilon. \tag{2}$$

This situation limits the analysis of practical instances involving hundreds or thousands taxa (see, e.g., the phylogenetic analysis of the datasets related to SARS-Cov2 [7]) and may occur because of specific values of either the path-length τ_{ij} or the distance d_{ij} (or both). In order to characterize the impact of τ_{ij} on the numerical stability of the BMEP, we start by observing that the path-lengths must encode an unrooted binary tree; this fact implies that $2 \leq \tau_{ij} \leq n - 1$,

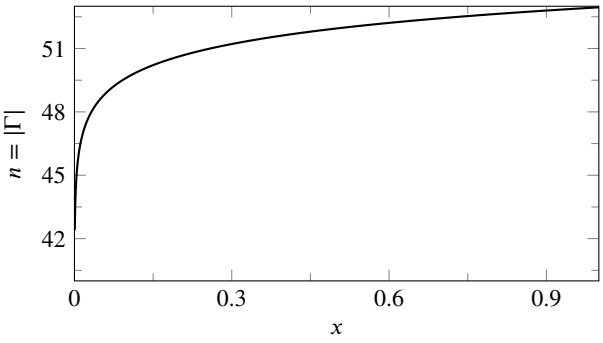


Figure 3: Behavior of the function $\log_2\left(\frac{x}{\epsilon}\right) + 1$ for $0 < x \leq 1$.

for all distinct pairs of taxa $i, j \in \Gamma$ [13]. Then, a question that naturally arises is what is the minimum value of $n = |\Gamma|$ for which numerical instabilities start to occur when solving an instance of the BMEP. A trivial algebraic manipulation of (2) provide the following answer: numerical instabilities occur whenever

$$n \geq \left\lfloor \log_2\left(\frac{d_{ij}}{\epsilon}\right) + 1 \right\rfloor. \tag{3}$$

Now, in 64-bit floating point arithmetic, $\epsilon \approx 2.3 \cdot 10^{-16}$; then, if the smallest non-diagonal entry of $\mathbf{D}$ is at most 1, the BMEP becomes numerically instable for

$$n \geq \left\lfloor \log_2\left(\frac{1}{2.3 \cdot 10^{-16}}\right) + 1 \right\rfloor = 53. \tag{4}$$

This bound may even decrease in function of the smallest non-diagonal entry of $\mathbf{D}$ (see, e.g., Figure 3).

To overcome numerical instabilities it would be desirable to pull as away as possible the bound on n for which numerical stability issues arise. One way to achieve this goal consists of finding a scalar $k > 1$ such that

$$n \geq \left\lfloor k \cdot \log_2\left(\frac{d_{ij}}{\epsilon}\right) + 1 \right\rfloor \tag{5}$$

may result as large as possible. Imposing (5) means considering the following new length function for the BMEP:

$$L_k(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma \setminus \{i\}} \frac{d_{ij}}{2^{\tau_{ij}/k}}. \tag{6}$$

We observe that any rescaling of the length function (1) can be written in the form of (6) and we refer to the problem of minimizing (6) as the *Rescaled Balanced Minimum Evolution Problem* (RBMEP). An interesting property of (6) is that in some circumstances it allows to characterize the optimal solution to the rescaled problem, as shown by the following result:

Proposition 1. *For a sufficiently large scalar $k > 1$, there exists an input distance matrix $\mathbf{D}$ for which the caterpillar phylogeny is an optimal solution to the RBMEP.*

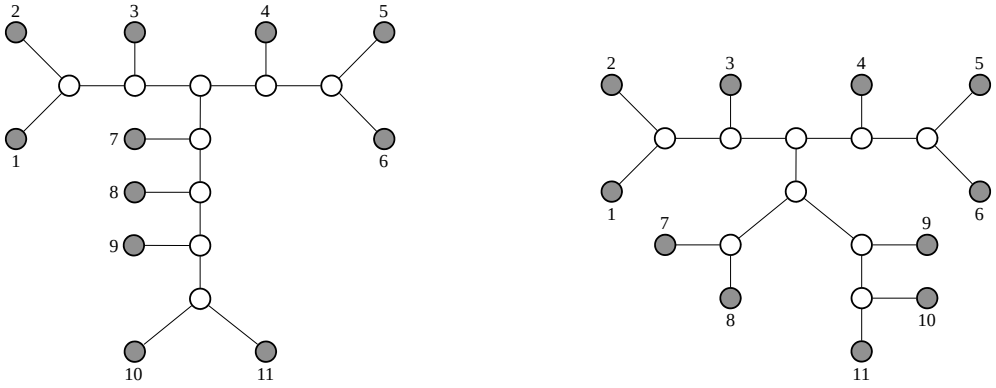


Figure 4: Two different phylogenies T (left) and T' (right) of 11 taxa.

Proof. Assume $n = 11$ and first observe that

$$\lim_{k\to\infty} L_k(T) = \frac{n(n-1)}{2} - \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} \tau_{ij} d_{ij}.$$

Then, for sufficiently large k, the phylogeny T that maximizes

$$\sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} \tau_{ij} d_{ij}$$

is optimal for the RBMEP. Now, observe that, by the Cauchy-Schwarz inequality ([14, 41]), the following inequality holds

$$\sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} \tau_{ij} d_{ij} \leq ||\tau||_F \cdot ||D||_F, \tag{7}$$

where τ is the matrix of path-lengths and $||\cdot||_F$ is the Frobenius norm of a matrix ([30]). Hence, the phylogeny T with maximal $||\tau||_F$ is optimal to the RBMEP. Since we assumed $n = 11$, the maximum $||\tau||_F^2 = 4,064$ is achieved by the caterpillar phylogeny. □

Proposition 1 shows us that there exist scalars $k > 1$ for which the optimal solution to the RBMEP can be identified independently of a given input distance matrix $\mathbf{D}$. This is not the case for the BMEP. Hence, a licit question is whether the new length function (6) allows to preserve the statistical consistency of the optimal solution to the RBMEP for sufficiently small $k > 1$. Unfortunately, the results presented in the next section provide a negative answer.

3. On the statistical consistency and numerical stability of the RBMEP

To investigate the statistical consistency of the RBMEP, we start by observing that the following propositions hold:

Proposition 2. *For any scalar $k > 1$, $L_k(T)$ overestimates the length of the true phylogeny of the set Γ of taxa encoded by the input distance matrix $\mathbf{D}$.*

Proof. As shown in Desper and Gascuel [20], $L(T)$ is a consistent estimation of the length of the true phylogeny associated to $\mathbf{D}$. Then, the statement follows by observing that $2^{-\tau_{ij}/k} > 2^{-\tau_{ij}}$ holds true for any scalar $k > 1$, hence $L_k(T) > L(T)$, for any phylogeny T of Γ. □

Proposition 3. *There exists an input distance matrix $\mathbf{D}$ for which the RBMEP is not statistically consistent, for any $k \geq 1.2447$.*

Proof. Assume $n = 11$ and consider the phylogenies T and T', shown in Figure 4, and the associated path-length matrices τ and τ', respectively, shown in Figure 5. Set $D = \tau'$. Then, it is possible to see numerically that $L_k(T) < L_k(T')$ for any $k \geq 1.2447$. Now, D corresponds precisely to T', hence T' is the true phylogeny for this specific instance of the RBMEP. Thus, for any $k \geq 1.2447$, the problem of minimizing $L_k(T)$ over the set of the phylogenies for Γ does not yield the desired true phylogeny. □

Proposition 3 still leaves some hope to find a scalar $1 < k < 1.2447$ for which the RBMEP could be statistically consistent. However, checking whether this hope is well-grounded is of little interest in practice mainly because, in the best case, this k (if it existed) would pull away the arising numerical instability up to

$$n \geq \left\lfloor 1.2446 \cdot \log_2\left(\frac{1}{2.3 \cdot 10^{-16}}\right) + 1 \right\rfloor = 65, \tag{8}$$

a value that is not suited for large scale phylogenetic analyses involving hundreds or thousands taxa.

As a striking example of the implications of the above propositions, consider the input distance matrix "Primates12" used in [1, 6, 9, 10, 11]. This matrix is characterized by a very strong *phylogenetic signal* (see [9]) and encodes the *General Time Reversible* (GTR) distances (see [24, 37, 42, 52, 53, 55]) for Hayasaka et al. [33]'s real aligned mitochondrial DNA sequences relative to the cytochrome b extracted from 12 primates. In particular, the matrix has been obtained (i) by treating all gaps as 'N' and (ii) by running the estimation model described in [12]. The optimal solutions to the BMEP and the RBMEP for $k = 2$ are shown as T and T' in Figure 6, respectively, and are unfortunately different.

$$\tau = \begin{bmatrix} 0 & 2 & 3 & 5 & 6 & 6 & 5 & 6 & 7 & 8 & 8 \\ 2 & 0 & 3 & 5 & 6 & 6 & 5 & 6 & 7 & 8 & 8 \\ 3 & 3 & 0 & 4 & 5 & 5 & 4 & 5 & 6 & 7 & 7 \\ 5 & 5 & 4 & 0 & 3 & 3 & 4 & 5 & 6 & 7 & 7 \\ 6 & 6 & 5 & 3 & 0 & 2 & 5 & 6 & 7 & 8 & 8 \\ 6 & 6 & 5 & 4 & 2 & 0 & 5 & 6 & 7 & 8 & 8 \\ 5 & 5 & 4 & 4 & 5 & 5 & 0 & 3 & 4 & 5 & 5 \\ 6 & 6 & 5 & 5 & 6 & 6 & 3 & 0 & 3 & 4 & 4 \\ 7 & 7 & 6 & 6 & 7 & 7 & 4 & 3 & 0 & 3 & 3 \\ 8 & 8 & 7 & 7 & 8 & 8 & 5 & 4 & 3 & 0 & 2 \\ 8 & 8 & 7 & 7 & 8 & 8 & 5 & 4 & 3 & 2 & 0 \end{bmatrix} \qquad \tau' = \begin{bmatrix} 0 & 2 & 3 & 5 & 6 & 6 & 6 & 6 & 6 & 7 & 7 \\ 2 & 0 & 3 & 5 & 6 & 6 & 6 & 6 & 6 & 7 & 7 \\ 3 & 3 & 0 & 4 & 5 & 5 & 5 & 5 & 5 & 6 & 6 \\ 5 & 5 & 4 & 0 & 3 & 3 & 5 & 5 & 5 & 6 & 6 \\ 6 & 6 & 5 & 3 & 0 & 2 & 6 & 6 & 6 & 7 & 7 \\ 6 & 6 & 5 & 3 & 2 & 0 & 6 & 6 & 6 & 7 & 7 \\ 6 & 6 & 5 & 5 & 6 & 6 & 0 & 2 & 4 & 5 & 5 \\ 6 & 6 & 5 & 5 & 6 & 6 & 2 & 0 & 4 & 5 & 5 \\ 6 & 6 & 5 & 5 & 6 & 6 & 4 & 4 & 0 & 3 & 3 \\ 7 & 7 & 6 & 6 & 7 & 7 & 5 & 5 & 3 & 0 & 2 \\ 7 & 7 & 6 & 6 & 7 & 7 & 5 & 5 & 3 & 2 & 0 \end{bmatrix}$$

Figure 5: Two path-length matrices τ and τ' encoding the two distinct phylogenies T and T' of Figure 4, respectively.

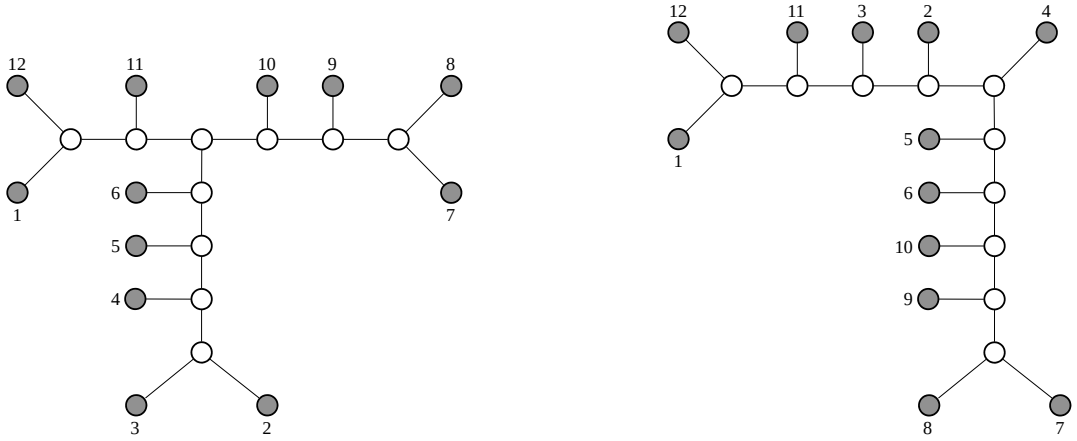


Figure 6: The optimal solutions T (left) and T' (right) to the BMEP and the RBMEP for $k = 2$, respectively, for the instance "Primates12" of 12 taxa.

Acknowledgments

The first author acknowledges support the Université Catholique de Louvain via the "Fonds Spéciaux de Recherche" (FSR) 2017-2021, and the Fondation Louvain via the grant "COALESCENS".

References

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

[2] Auch, A.F., R., H.S., Holland, B.R., M., G., 2006. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinformatics 7, 1–5.

[3] Bordewich, M., Gascuel, O., Huber, K., Moulton, V., 2009a. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 110–117.

[4] Bordewich, M., Gascuel, O., Huber, K.T., Moulton, V., 2009b. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 110–117.

[5] Catanzaro, D., 2011. Estimating phylogenies from molecular data, in: Bruni, R. (Ed.), Mathematical approaches to polymer sequence analysis and related problems. Springer, NY, pp. 149–176.

[6] Catanzaro, D., Aringhieri, R., di Summa, M., Pesenti, R., 2015. A branch-price-and-cut algorithm for the minimum evolution problem. European Journal of Operational Research 244, 753–765.

[7] Catanzaro, D., Frohn, M., Gascuel, O., Pesenti, R., submitted 2020. A tutorial on the balanced minimum evolution. European Journal of Operational Research .

[8] Catanzaro, D., Frohn, M., Pesenti, R., 2020a. An information theory perspective on the balanced minimum evolution problem. Operations Research Letters 48, 362–367.

[9] Catanzaro, D., Labbé, M., Pesenti, R., 2013. The balanced minimum evolution problem under uncertain data. Discrete Applied Mathematics 161, 1789–1804.

[10] Catanzaro, D., Labbé, M., Pesenti, R., Salazar-Gonzáles, J.J., 2009. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. Networks 53, 126–140.

[11] Catanzaro, D., Labbé, M., Pesenti, R., Salazar-Gonzáles, J.J., 2012. The balanced minimum evolution problem. INFORMS Journal on Computing 24, 276–294.

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

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

[14] Cauchy, A.L., 1821. Cours d'Analyse de l'École Royale Polytechnique. L'Imprimerie Royale, Debure frères, Libraires du Roi et de la Bibliothèque du Roi.

[15] Cheng, X., Du, D.Z., 2001. Steiner Trees in Industry. Kluwer Academic Publishers, Boston, MA.

[16] Cieslik, D., 1998. Steiner minimal trees. Springer, Boston, MA, USA.
[17] Criscuolo, A., Michel, C.J., 2009. Phylogenetic inference with weighted codon evolutionary distances. Journal of Molecular Evolution 68, 377–392.
[18] Cueto, M.A., Matsen, F.A., 2011. Polyhedral geometry of phylogenetic rogue taxa. Bulletin of Mathematical Biology 73, 1202–1226.
[19] Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. Journal of Computational Biology 9, 687–705.
[20] Desper, R., Gascuel, O., 2004. 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, 587–598.
[21] Desper, R., Gascuel, O., 2008. Encyclopedia of Algorithms. Springer. chapter Distance-based Phylogeny Reconstruction (Optimal Radius). pp. 1–99.
[22] Du, D.Z., Hu, X., 2008. Steiner tree problems in computer communication networks. World Scientific Publishing Company, Singapore.
[23] Du, D.Z., Smith, J.M., Rubinstein, J.H., 2000. Advances in Steiner trees. Kluwer Academic Publishers, Boston, MA, USA.
[24] Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.
[25] Fiorini, S., Joret, G., 2012. Approximating the balanced minimum evolution problem. Operations Research Letters 40, 31–35.
[26] Forcey, S., Keefe, L., Sands, W., 2015. Facets of the balanced minimal evolution polytope. Journal of Mathematical Biology 73, 447–468.
[27] Forcey, S., Keefe, L., Sands, W., 2017. Split-facets for balanced minimal evolution polytopes and the permutoassociahedron. Bulletin of Mathematical Biology, in press 79, 975–994.
[28] Gascuel, O., 2005. Mathematics of evolution and phylogeny. Oxford University Press, New York, NY.
[29] Gascuel, O., Steel, M.A., 2006. Neighbor-joining revealed. Molecular Biology and Evolution 23, 1997–2000.
[30] Golub, G., van Loan, C., 1996. Matrix computations. The Johns Hopkins University Press, London.
[31] Gusfield, D., 1984. The Steiner Tree Problem in Phylogeny. Technical Report 334. Yale University, New Haven, CT.
[32] Haws, D.C., Hodge, T.L., Yoshida, R., 2011. Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. Bulletin of Mathematical Biology 73, 2627–2648.
[33] Hayasaka, K., Gojobori, T., Horai, S., 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Molecular Biology and Evolution 5, 626–644.
[34] Hwang, F.K., Richards, D.S., Winter, P., 1992. The Steiner tree problem. North-Holland, Amsterdam, The Netherlands.
[35] Johnson, D.S., Lenstra, J.K., Kan, A.H.G.R., 1978. The complexity of the network design problem. Networks 8, 279–285.
[36] Jung, M., 2012. Évolution du VIH: méthodes, modèles et algorithmes. Ph.D. thesis. Université Montpellier II - Sciences et Techniques du Languedoc, France.
[37] Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20, 86–93.
[38] Lefort, V., Desper, R., Gascuel, O., 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Molecular Biology and Evolution 32, 2798–2800.
[39] Li, X., Mao, Y., 2016. Generalized connectivity of graphs. Springer, Boston, MA, USA.
[40] Lu, C.L., Tang, C.Y., Lee, R.C.T., 2003. The full Steiner tree problem. Theoretical Computer Science 306, 55–67.
[41] Mitrinović, D.S., Pečarić, J.E., Fink, A.M., 1993. Classical and new inequalities in analysis. Kluwer Academic Publishers Group, Dordrecht.
[42] Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford, UK.
[43] Pardi, F., 2009. Algorithms on Phylogenetic Trees. Ph.D. thesis. University of Cambridge, UK.
[44] Pardi, F., Gascuel, O., 2012. Combinatorics of distance-based tree inference. PNAS 109, 16443–16448.
[45] Pardi, F., Gascuel, O., 2016. Encyclopedia of Evolutionary Biology. Elsevier. chapter Distance-based methods in phylogenetics. pp. 458–465.
[46] Pardi, F., Guillemot, S., Gascuel, O., 2010. Robustness of phylogenetic inference based on minimum evolution. Bulletin of Mathematical Biology 72, 1820–1839.
[47] Pauplin, Y., 2000. Direct calculation of a tree length using a distance matrix. Journal of Molecular Evolution 51, 41–47.
[48] Pop, P.C., 2012. Generalized network design problems: Modeling and optimization. De Gruyter, Berlin, Germany.
[49] Prömel, H.J., Steger, A., 2002. The Steiner tree problem: A tour through graphs, algorithms, and complexity. Vieweg+Teubner Verlag, Berlin.
[50] Schwartz, R., 2019. Computational models for cancer phylogenetics, in: Warnow, T. (Ed.), Bioinformatics and Phylogenetics. Computational Biology. Springer, Cham. volume 29, pp. 243–275.
[51] Semple, C., Steel, M.A., 2004. Cyclic permutations and evolutionary trees. Advances in Applied Mathematics 32, 669–680.
[52] Tavare, S., 1987. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17, 57–86.
[53] Waddell, P.J., Steel, M.A., 1997. General time-reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. Molecular Phylogenetics and Evolution 8, 398–414.
[54] Wu, B.Y., Chao, K.M., 2004. Spanning trees and optimization problems. Chapman and Hall/CRC, Boca Raton, FL.
[55] Yang, Z., 2014. Molecular Evolution: A Statistical Approach. Oxford University Press, Oxford, UK.