An information theory perspective on the Balanced Minimum Evolution Problem

Daniele Catanzaro[a,*], Martin Frohn[a], Raffaele Pesenti[b]

[a]*Center for Operations Research and Econometrics (CORE)*
Université Catholique de Louvain, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium.
[b]*Department of Management, Università Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venezia, Italy.*

Abstract

We investigate the *Balanced Minimum Evolution Problem* (BMEP) from an information theory perspective and we show that, under specific hypotheses, the BMEP can be considered as a cross-entropy minimization problem. This perspective contributes to bridge the gap between phylogenetics and information theory and enables the development of new lower bounds on the value of the optimal solution to the BMEP that can be efficiently computed by exploiting some analogies between the BMEP and Huffman coding.

Keywords: combinatorial optimization, the balanced minimum evolution problem, phylogenetics, distance methods, Huffman coding, entropy encoding, cross-entropy, Kulleback-Leibler divergence, Pinsker's inequality.

1. Introduction

A *phylogeny* of a set Γ of $n \geq 3$ species (also called *taxa*) is an unrooted binary tree having Γ as leafset [1, 2, 3, 4]. Consider a $n \times n$ symmetric distance matrix $\mathbf{D} = \{d_{ij}\}$, whose generic entry d_{ij} represents a measure of dissimilarity between the pair of taxa $i, j \in \Gamma$. Then, the *Balanced Minimum Evolution Problem* (BMEP) consists in finding a phylogeny T of Γ that minimizes the following *length function*

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

where τ_{ij} represents the *topological distance between taxa i and j* in T, i.e., the number of edges belonging to the (unique) path connecting taxon i to taxon j [2, 5]. The BMEP is a statistically consistent phylogenetic estimation model that belongs to the family of the distance matrix methods [1, 6, 7, 8, 9]. It has been introduced by Pauplin [10] in 2000 and systematically investigated from a biological perspective in [11, 12]. The problem is proved to be $\mathcal{NP}$-hard and inapproximable within c^n, for some positive constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$ [13]. This fact has justified the development of a number of implicit enumeration algorithms to exactly solve it (e.g., [2, 5, 4]) as well as a number of heuristics to approximate its optimal solution (e.g., [5, 6, 8]). The current state-of-the-art exact solution algorithm for the BMEP, described in [2], is based on an integer linear programming model that exploits a number of combinatorial properties that the BMEP shares with the *Huffman Coding Problem* [14, 15, 16]. Such properties suggest the existence of a deep connection between both problems and, more in general, between phylogenetics and coding theory. However, these properties have never been systematically investigated in the literature. In this article, we address this issue more in detail. In particular, we show that, under specific hypotheses, the BMEP can be seen as a *cross-entropy* minimization problem [17, 18].

The concept of information entropy is not new in the literature of phylogenetics. For example, in the context of Bayesian inference, information entropy provides a natural way to measure the *information content* (or, equivalently, the *noise*) of systematic data [19, 20, 21, 22, 23]. In the context of evolution of biological systems, information entropy is proposed as a conceptual bond to relate microevolutionary patterns (i.e., changes occurring on time-scales shorter than speciation rates) to macroevolutionary patterns (i.e., changes occurring on time-scales longer than speciation rates) [24]. More importantly, in the context of phylogenetic diversity, information entropy is used to describe possible methods to compute a measure of dissimilarity within taxa belonging to the same population of individuals or between taxa from different populations of individuals [25]. Further insights on the measures of dissimilarity among taxa (with particular emphasis on the *quadratic entropy* [26]), new methods to improve them, and efficient algorithms to compute them are presented in [22, 27, 28, 29, 30, 31]. A recent survey on entropy measures in phylogenetic diversity is proposed by Chao et al. [26]. This article further extends the use of information entropy in phylogenetics and brings to light the existence of deep connections between information theory and the class of phylogenetic estimation models based on the *minimum evolution criterion* (see [1, 6, 7]). Such class includes e.g., the BMEP and the *Minimum Evolution Problem* (MEP) [1, 32] from among distance methods and the *Parsimonious Phylogeny Estimation Problem* (PPEP) and its versions from among the character-based methods [7]. Because some of these problems are strictly related (e.g., the BMEP is known to be a restriction of the MEP [32], which in turn shares strong combinatorial aspects with the PPEP [33]) it is plausible to think that

[*] Corresponding author

Email addresses: `daniele.catanzaro@uclouvain.be` (Daniele Catanzaro), `martin.frohn@uclouvain.be` (Martin Frohn), `pesenti@unive.it` (Raffaele Pesenti) the relationships between information entropy and the BMEP presented in this article may be further extended also to other problems belonging to this class. Our belief is that having a better mathematical understanding of these connections may bring new ideas in both the theoretical and computational aspects of phylogenetics and ultimately pay off in terms of improved estimation algorithms.

The article is organized as follows. In Section 2, we introduce some notation and briefly recall a number of fundamental results from information theory and phylogenetics that will prove useful throughout the articles. In Section 3, we show that the BMEP can be seen as a cross-entropy minimization problem, in which probabilities are restricted to encode unrooted binary trees. We exploit the structure of unrooted binary trees to connect the BMEP with the *Kullback-Leibler divergence*, thereby providing a specific lower bound on the value of the optimal solution to the BMEP via an adaptation of *Pinsker's inequality* [17, 18]. Finally, in Section 4 we draw a parallel between the Kullback-Leibler divergence and Huffman coding in order to both strengthen the proposed lower bound and develop an efficient algorithm to compute it.

2. Background and Notation

Given a phylogeny T of Γ and a taxon $i \in \Gamma$, we denote Γ_i as the set $\Gamma \setminus \{i\}$ and $\tau_i = (\tau_{i1}, \ldots, \tau_{in})$ as the *path-length sequence seen from taxon i*, i.e., the vector of the topological distances relative to the $(n-1)$ paths in T from taxon i to the remaining taxa in Γ_i. We define $\tau = \{\tau_i : i \in \Gamma\}$ as the *path-length sequence collection* of the topological distances in T. Note that a path-length sequence collection τ encodes all topological distances of a phylogeny T. Finally, we define $\mathcal{T}$ as the set of all possible phylogenies for Γ, Θ as the set of all path-length sequences τ_i from a generic taxon $i \in \Gamma$ associated to the phylogenies in $\mathcal{T}$, and $\bar{\Theta}$ as the set of path-length sequence collections τ associated to the phylogenies in $\mathcal{T}$. Throughout the article, we will explicit the dependence of the above entities on the number of taxa considered whenever the context can be ambiguous. For example, we will write Θ_n to denote the set Θ for a given set Γ of n taxa.

We recall thats the sets Θ_i, $i \in \Gamma$, and Θ are characterized by specific equalities [2]. In particular, the first equality, called *Kraft equality* [15], states that $\tau_i \in \Theta_n$ if and only if

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

Stott-Parker and Ram [15] showed that Kraft equality plays a central role in the combinatorics of the *Huffman Coding Problem* (HCP) [16], as it contributes to explain e.g., why the HCP can be solved in polynomial-time. The second equality, constrains the path-length sequence collections associated to the phylogenies of a given set Γ of taxa to satisfy the *the phylogenetic manifold equation* [34], i.e.,

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = 2n - 3, \quad \forall\, \tau \in \bar{\Theta}. \tag{2}$$

We refer the reader interested in deepening the study of these equalities as well as combinatorial aspects of the BMEP to [34]. In the next sections we will see that the phylogenetic manifold plays a central role in determining possible lower bounds for the optimal solution to the BMEP.

Fixed a positive integer m, consider a source that sends out data as a sequence of symbols over an alphabet $\Sigma = \{a_1, a_2, \ldots, a_m\}$. Let $p(a_j)$ be the probability of the occurrence of the symbol $a_j \in \Sigma$ in the sequence and let $p = (p_1, \ldots, p_m)$ the discrete distribution of such probabilities. Assume that the symbols in Σ are independent, identically distributed, and that the probabilities of their occurrences are known a priori. Then, the *information entropy* (or *Shannon's entropy*) of the source is defined as [16]

$$\mathcal{H}(p) = -\sum_{j=1}^{m} p(a_j) \log_2 p(a_j). \tag{3}$$

Let C be a *coding scheme* for the source, i.e., a set of codewords that are in a bijective relationship with the symbols in Σ. Let l_j be the *length of the codeword* associated to a_j with respect to a given measuring system (usually the binary system). Then, the *Average Codeword Length* (ACL) of C is defined as $\mathrm{ACL}(C) = \sum_{j=1}^{m} p(a_j) l_j$ [16]. Finding a coding scheme for the given source having minimum ACL is a central problem in information theory [16]. The fundamental *Shannon's source coding theorem* provides a lower bound for such a minimum, by stating that no coding scheme for the source can have an ACL smaller than the relative information entropy [35].

We recall two supplementary definitions from information theory that will be frequently used in the remainder of the article. In particular, given a second probability distribution $q = (q_1, \ldots, q_m)$ associated to the same set of symbols $a_j \in \Sigma$, the *cross-entropy* is defined as [17, 18]

$$\mathcal{H}(p, q) = -\sum_{j=1}^{m} p_j \log_2(q_j)$$

while the *Kullback-Leibler Divergence* (KLD), also called *relative entropy*, is defined as

$$D_{KL}(p \| q) = \mathcal{H}(p, q) - \mathcal{H}(p).$$

The cross-entropy can be interpreted as the information entropy of the source when data follow a true discrete probability distribution p while their lengths are encoded with respect to a "wrong" discrete probability distribution q. In this context, the KLD measures of the dissimilarity between the proper information entropy of the source and the "wrong one". We refer the reader interested in deepening the interpretations of such definitions to [16].

3. The BMEP as a Cross-Entropy Minimization Problem

We will show now that the BMEP can be seen as a cross-entropy minimization problem in which the probabilities must satisfy the phylogenetic manifold equation (2). Before proceeding, we observe that the following result holds:

Proposition 1. *The input distance matrix* $\mathbf{D}$ *of the BMEP is an* Exponentially Double-Stochastic *(EDS) matrix, i.e., it is such that its component-wise exponential matrix* $2^{-\mathbf{D}} = \{2^{-d_{ij}} : i, j \in \Gamma\}$ *enjoys the following properties:*

- $(2^{-\mathbf{D}})_{ii} = 1$ *for all* $i \in 1, \ldots, n$*;*
- $2^{-\mathbf{D}} = \mathbf{S} + \mathbf{I}$ *where* $\mathbf{S}$ *is a double-stochastic matrix.*

Proof. We first observe that $2^{-\mathbf{D}} - \mathbf{I}$ is a nonnegative symmetric matrix. Then, by [36, 37], there exists an unique diagonal matrix $\Pi = diag(\pi_1, \ldots, \pi_n)$ with positive entries such that matrix $\mathbf{S} = \Pi(2^{-\mathbf{D}} - \mathbf{I})\Pi$ has rows/columns that sum to one if and only if there exists a symmetric nonnegative matrix with the same support as $\mathbf{S}$ and rows that sum to 1. As any matrix $2(2^{-\tau} - \mathbf{I})$, with $\tau \in \bar{\Theta}$, satisfies the latter requirement, we can conclude that matrix $\mathbf{S}$ exists. □

Now, consider the positive matrix $\mathbf{S} + \mathbf{I}$, and its component wise logarithmic matrix $\hat{\mathbf{S}}$. We have that

$$(\mathbf{S}+\mathbf{I})_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \pi_i \pi_j 2^{-d_{ij}} & \text{if } i \neq j \end{cases}$$
$$\Leftrightarrow$$
$$\hat{\mathbf{S}}_{ij} = \begin{cases} 0 & \text{if } i = j, \\ -\log_2(\pi_i) - \log_2(\pi_j) + d_{ij} & \text{if } i \neq j. \end{cases} \tag{4}$$

In particular, for each $\tau \in \bar{\Theta}$, it holds that

$$\begin{aligned} \mathrm{Tr}(2^{-\tau}\hat{\mathbf{S}}) &= \mathrm{Tr}(2^{-\tau}\mathbf{D}) - \sum_{\substack{i,j\in\Gamma \\ i\neq j}} \frac{\log_2(\pi_i)}{2^{-\tau_{ij}}} - \sum_{\substack{i,j\in\Gamma \\ i\neq j}} \frac{\log_2(\pi_j)}{2^{-\tau_{ij}}} \\ &= \mathrm{Tr}(2^{-\tau}\mathbf{D}) - \sum_{i\in\Gamma} \frac{\log_2(\pi_i)}{2} - \sum_{j\in\Gamma} \frac{\log_2(\pi_j)}{2} \\ &= \mathrm{Tr}(2^{-\tau}\mathbf{D}) - \sum_{i\in\Gamma} \log_2(\pi_i) \end{aligned} \tag{5}$$

where $\sum_{i\in\Gamma} \log_2(\pi_i)$ is independent of the choice of $\tau \in \bar{\Theta}$ and consequently,

$$\underset{2^{-\tau}\in 2^{-\bar{\Theta}}}{\arg\min}\, \mathrm{Tr}(2^{-\tau}\mathbf{D}) = \underset{2^{-\tau}\in 2^{-\bar{\Theta}}}{\arg\min}\, \mathrm{Tr}(2^{-\tau}\hat{\mathbf{S}}). \tag{6}$$

We exploit the above results to rewrite (1) as a sum of cross-entropies through the following series of equalities:

$$\begin{aligned} L(T) &= \sum_{i\in\Gamma}\sum_{j\in\Gamma_i} \frac{d_{ij}}{2^{\tau_{ij}}} = \mathrm{Tr}(2^{-\tau}\mathbf{D}) = \frac{1}{2}\mathrm{Tr}(2\cdot 2^{-\tau}\mathbf{D}) \\ &= \frac{1}{2}\sum_{i\in\Gamma}\sum_{j\in\Gamma_i} \frac{d_{ij}}{2^{\tau_{ij}-1}} = \frac{1}{2}\sum_{i\in\Gamma}\left(-\sum_{j\in\Gamma_i} \frac{1}{2^{\tau_{ij}-1}} \log_2(2^{-d_{ij}})\right) \\ &= \frac{1}{2}\sum_{i\in\Gamma}\left(-\sum_{j\in\Gamma_i} p_{ij}(\tau)\log_2(q_{ij})\right) = \frac{1}{2}\sum_{i\in\Gamma}\mathcal{H}(p_i(\tau), q_i) \end{aligned} \tag{7}$$

where:

- $p_{ij}(\tau) = 2^{1-\tau_{ij}} > 0$ and such that $\sum_{j\in\Gamma_i} p_{ij}(\tau) = 1$; hence, $p_i(\tau) = \{p_{ij}(\tau) : j \in \Gamma_i\}$ can be seen as a probability distribution;
- $q_{ij} = 2^{-d_{ij}} > 0$ and such that $\sum_{j\in\Gamma_i} q_{ij} = 1$; hence, $q_i = \{q_{ij} : j \in \Gamma_i\}$ can be seen as a probability distribution.
- $\mathcal{H}(p_i(\tau), q_i)$ is the cross-entropy between the two probability distributions $p_i(\tau)$ and q_i over the same underlying set Γ_i.

We also observe that

- The cross-entropy is always greater than or equal to the entropy, i.e.,

  $$\mathcal{H}(p_i(\tau), q_i) \geq \mathcal{H}(p_i(\tau), p_i(\tau)) = \mathcal{H}(p_i(\tau)).$$

  and, in particular,

  $$\begin{aligned} &\min_q \mathcal{H}(p, q) = \mathcal{H}(p, p), \\ &\min_p \mathcal{H}(p, q) = \log\max\{q_j\}. \end{aligned}$$

- The KLD in the context of the BMEP is

  $$D_{KL}(p_i(\tau)||q_i) = \mathcal{H}(p_i(\tau), q_i) - \mathcal{H}(p_i(\tau))$$

  and, in particular,

  $$\begin{aligned} &\min_q D_{KL}(p||q) = D_{KL}(p||p) = 0 \\ &= \min_p D_{KL}(p||q) = D_{KL}(q||q) \end{aligned}$$

- The Pinsker's inequality holds on $D_{KL}(p_i(\tau)||q_i)$, together with its reversed expression [38], i.e.,

  $$\alpha \leq D_{KL}(p_i(\tau)||q_i) \leq \log_2(1+\beta) - \delta$$

  where

  $$\begin{aligned} \alpha &= \frac{1}{\log 4}||p_i(\tau) - q_i||_1^2, \quad \beta = \frac{||p_i(\tau) - q_i||_1^2}{2\min_{j\in\Gamma_i} q_{ij}}, \\ \delta &= \frac{\gamma}{\log 4}||p_i(\tau) - q_i||_1^2, \quad \gamma = \min_{j\in\Gamma_i}\left\{\frac{p_{ij}(\tau)}{q_{ij}}\right\} \end{aligned}$$

  being $||p_i(\tau) - q_i||_1^2 = \left(\sum_{j\in\Gamma_i} |p_{ij}(\tau) - q_{ij}|\right)^2$.

Specific bounds on the value of the KLD for rooted trees can be found in [17, 18]. Since all phylogenies satisfy (2), we can reformulate further the KLD as follows:

$$\begin{aligned} &D_{KL}(p_i(\tau)||q_i) \\ &\quad = -\sum_{j\in\Gamma_i}(p_{ij}(\tau)\cdot\log_2(q_{ij})) + \sum_{j\in\Gamma_i}(p_{ij}(\tau)\cdot\log_2(p_{ij}(\tau))) \\ &\quad = 2\cdot\sum_{j\in\Gamma_i}\frac{d_{ij}}{2^{\tau_{ij}}} + \sum_{j\in\Gamma_i}\left(\frac{1}{2^{\tau_{ij}-1}}\cdot\log_2\left(2^{1-\tau_{ij}}\right)\right) \\ &\quad = 2\cdot\sum_{j\in\Gamma_i}\left(\frac{d_{ij}-\tau_{ij}}{2^{\tau_{ij}}}\right) + 1 \end{aligned}$$

to obtain

$$\begin{aligned}
&\min_{2^{-\tau}\in 2^{-\bar{\Theta}}} \frac{1}{2}\sum_{i\in\Gamma} \mathcal{H}(p_i(\tau), q_i) \\
&= \min_{2^{-\tau}\in 2^{-\bar{\Theta}}} \frac{1}{2}\sum_{i\in\Gamma} \left(D_{KL}(p_i(\tau)\|q_i) + 2\sum_{j\in\Gamma_i} \frac{\tau_{ij}}{2^{\tau_{ij}}} - 1 \right) \\
&= \min_{2^{-\tau}\in 2^{-\bar{\Theta}}} \frac{1}{2}\sum_{i\in\Gamma} D_{KL}(p_i(\tau)\|q_i) + \frac{3n-6}{2}
\end{aligned}$$

which in turn leads to the following lower bound on the value of the optimal solution to the BMEP:

$$\frac{3n-6}{2} + \frac{1}{2}\sum_{i\in\Gamma} \min_{\tau_i\in\Theta_n} D_{KL}(p_i(\tau), q_i). \tag{8}$$

4. Relative Entropy Minimization of Topological Distances

A question that natural arises is how to efficiently compute (8) for a given instance of the BMEP. In this section, we will develop a polynomial-time algorithm to address this question. Before proceeding, we introduce some notation and definitions that will prove useful throughout the remainder of the article. In particular, given a vector $x \in \mathbb{R}^n_{0^+}$, we define sort ↑ (x) as the vector obtained from x by sorting its entries in non-decreasing order. For example, if $x = (9, 7, 8, 10)$, then sort ↑ $(x) = (7, 8, 9, 10)$. Given a second vector $y \in \mathbb{R}^n_{0^+}$, we define sort$(x \mid \uparrow y)$ as the vector obtained from x by reordering its entries according to the permutation induced by sort ↑ (y). For example, let $x = (9, 7, 8, 10)$ and $y = (5, 1, 6, 2)$; the permutation of indices of the entries of y yielding sort ↑ (y) is $(2, 4, 1, 3)$; then, sort$(x \mid \uparrow y) = (7, 10, 9, 8)$. Fixed a generic n-dimensional vector $d \in \mathbb{R}^n_{0^+}$, we define d° as the $(n-1)$-dimensional vector having entries $d^\circ = (d_1, \ldots, d_{n-2}, (d_{n-1} + d_n)/2 - 1)$. Similarly, we define d^- as the $(n-1)$-dimensional vector having entries $d^- = (d_1, \ldots, d_{n-2}, d_n - 1)$. Given a path-length sequence of a phylogeny of n taxa $\tau_i \in \Theta_n$ and a generic n-dimensional vector $d \in \mathbb{R}^n_{0^+}$, we consider the support function $f_n : \Theta_n \times \mathbb{R}^n_{0^+} \to \mathbb{R}$ defined as

$$f_n(\tau_i, d) = \sum_{j=1}^{n} (d_j - \tau_{ij}) 2^{-\tau_{ij}}. \tag{9}$$

Finally, for a fixed n-dimensional vector $d \in \mathbb{R}^n_{0^+}$, we define the *Minimum Relative Entropy Problem* (MREP) associated to d as the problem of finding the path-length sequence $\tau_i \in \Theta_n$ that minimizes (9), i.e.,

$$\min_{\tau_i\in\Theta_n} f_n(\tau_i, d). \tag{10}$$

We first observe that

$$\min_{\tau_i\in\Theta_n} f_n(\tau_i, d) = \min_{\tau_i\in\Theta_n} \frac{1}{2} D_{KL}(p_i(\tau_i), q_i) - 1. \tag{11}$$

We also observe that for $n = 3$, $\Theta_n = \{(2, 3, 3)\}$. Hence, $(2, 3, 3)$ is an optimal solution to (11). Moreover, the following proposition holds:

Proposition 2. *Let $d \in \mathbb{R}^n_{0^+}$ be a generic n-dimensional vector and let x be a path-length sequence in Θ_n. Then*

(i) $f_n(x, d) \geq f_n(sort \uparrow (x), sort \uparrow (d))$.

(ii) $\min\{f_n(x, d) : x \in \Theta_n\} = \min\{f_{n-1}(y, d^\circ) : y \in \Theta_{n-1}\}$.

Proof.

(i) Let $d_i < d_j$ with $1 \leq i, j \leq n$. Then

$$\begin{aligned}
&(d_i - x_i)2^{-x_i} + (d_j - x_j)2^{-x_j} \\
&\leq (d_i - x_j)2^{-x_j} + (d_j - x_i)2^{-x_i} \;\Leftrightarrow\; x_i \leq x_j
\end{aligned}$$

This implies $x_{i+1} = x_i$ whenever $d_{i+1} < d_i$, hence $f_n(x, d) \geq f_n(\text{sort} \uparrow (x), \text{sort} \uparrow (d))$.

(ii) Assume, without loss of generality, that x is sorted non-decreasing and suppose, by contradiction, that

$$\min\{f_n(x, d) : x \in \Theta_n\} > \min\{f_{n-1}(y, d^\circ) : y \in \Theta_{n-1}\}.$$

Let $x' = (y_1, \ldots, y_{n-2}, y_{n-1} + 1, y_{n-1} + 1)$. We first note that $x' \in \Theta_n$. Now, consider the difference $f_n(x', d) - f_{n-1}(y, d^\circ)$. We obtain

$$\begin{aligned}
&f_n(x', d) - f_{n-1}(y, d^\circ) \\
&= (d_{n-1} - (y_{n-1} + 1))2^{-(y_{n-1}+1)} + (d_n - (y_{n-1} + 1))2^{-(y_{n-1}+1)} \\
&\quad - (((d_{n-1} + d_n)/2 - 1) - y_{n-1})\, 2^{-y_{n-1}} = 0
\end{aligned}$$

Hence, $f_n(x', d) < \min\{f_n(x, d) : x \in \Theta_n\}$, which leads to a contradiction. As a result, we can deduce that

$$\min\{f_n(x, d) : x \in \Theta_n\} \leq \min\{f_{n-1}(y, d^\circ) : y \in \Theta_{n-1}\}.$$

Now, suppose by contradiction that the strict inequality holds. Let $y' = (x_1, \ldots, x_{n-2}, x_{n-1} - 1)$. Consider the difference $f_{n-1}(y', d^\circ) - f_n(x, d)$. We obtain

$$\begin{aligned}
&f_{n-1}(y', d^\circ) - f_n(x, d) \\
&= (((d_{n-1} + d_n)/2 - 1) - (x_{n-1} - 1))\, 2^{-(x_{n-1}-1)} \\
&\quad - (d_{n-1} - x_{n-1})2^{-x_{n-1}} - (d_n - x_{n-1})2^{-x_{n-1}} = 0.
\end{aligned}$$

This implies that $\min\{f_{n-1}(y, d^\circ) : y \in \Theta_{n-1}\}$ is not an optimal solution to the MREP of d°, which leads to a contradiction. Thus, also the second claim follows. □

Our observations, together with Proposition 2, imply that for an optimal solution x of the MREP associated to an input n-dimensional vector $d \in \mathbb{R}^n_{0^+}$, the following recursion holds for $k = n, \ldots, 3$:

$$f_k(x, \text{sort} \uparrow (d)) = f_{k-1}(\text{sort}(x^- \mid \uparrow d^\circ), \text{sort} \uparrow (d^\circ)). \tag{12}$$

We explain this recursion by means of then following example.

Example 1. Consider the eight-dimensional vector $d = (3, 4, 4, 4, 4, 4, 5, 5)$ and the corresponding instance of the MREP associated to it. With a little abuse of notation, we write d_8 to indicate that the vector d is eight-dimensional and we denote $d_8(i)$ as its i-th entry. By definition, the seven-dimensional vector d_8° associated to d_8 is $(3, 4, 4, 4, 4, 4, 4)$. By Proposition 2, the optimal solution to $\min_{x\in\Theta_n} f(x, d_8)$ can be computed by solving the MREP associated to d_8°. By setting

$d_{k-1} = \text{sort} \uparrow (d_k^\circ)$, we can recurse this process via (12) until $k = 3$, by giving rise to the following vectors

$$\begin{aligned} & d_7 = (3,4,4,4,4,4,4) \Rightarrow d_7^\circ = (3,4,4,4,4,3) \\ \Rightarrow & d_6 = (3,3,4,4,4,4) \Rightarrow d_6^\circ = (3,3,4,4,3) \\ \Rightarrow & d_5 = (3,3,3,4,4) \Rightarrow d_5^\circ = (3,3,3,3) \\ \Rightarrow & d_4 = (3,3,3,3) \Rightarrow d_4^\circ = (3,3,2) \\ \Rightarrow & d_3 = (2,3,3) \end{aligned}$$

As $(2,3,3) = \arg\min\{f(y, d_3) : y \in \Theta_3\}$, we can easily solve the base case for the recursion, by obtaining

$$\min_{x\in\Theta_n} f_n(x, d_8) = (2-2)\cdot 2^{-2} + (3-3)\cdot 2^{-3} + (3-3)\cdot 2^{-3} = 0.$$

In order to recover the optimal solution to the considered instance, we need to appropriately keep track of the reordering process carried out during the recursion. To this end, we initially associate a tuple $l(i) = (i, \text{null})$ to each entry of vector d. These labels propagate to the entries of vector d° according to the following rule

$$l(i)^\circ = \begin{cases} l(i) & \text{if } i = 1,\ldots,n-2, \\ (l(n-1), l(n)) & \text{if } i = n-1. \end{cases}$$

Then, by considering l_i° as a label of d_i°, we can mimic as follows the reordering process of d° during the recursion:

$$\begin{aligned} & d_8^\circ = (3,4,4,4,4,4,4,4) : l_8^\circ = (1,2,3,4,5,6,(7,8)) \\ \Rightarrow & d_7^\circ = (3,4,4,4,4,3) : l_7^\circ = (1,2,3,4,5,(6,(7,8))) \\ \Rightarrow & d_6^\circ = (3,3,4,4,3) : l_6^\circ = (1,(6,(7,8)),2,3,(4,5)) \\ \Rightarrow & d_5^\circ = (3,3,3,3) : l_5^\circ = (1,(4,5),(6,(7,8)),(2,3)) \\ \Rightarrow & d_4^\circ = (3,3,2) : l_4^\circ = (1,(4,5),((6,(7,8)),(2,3))). \end{aligned}$$

When the base case is reached, from the sequence of labels $l_3^\circ = (((6,(7,8)),(2,3)),1,(4,5))$ associated to $d_3 = (2,3,3)$ we can deduce that sort↑ $(4,5,5,4,4,3,4,4)$ is an optimal solution to the considered instances of the MREP (see Figure 1). Example 1 shows that $D_{KL}(p_i(\tau), q_i) = 1$ whenever the ith row $\mathbf{D}_{i,\cdot}$ of a given distance matrix $\mathbf{D}$ is a path-length sequence. Moreover, if $\mathbf{D}$ is a path-length sequence collection, then the lower bound (8) is equal to the right-hand side of (2) and therefore strict.

Introducing the symmetry constraint $\tau_{ij} = \tau_{ji}$, for all $i, j \in \Gamma$, $i < j$, in the MREP may potentially improve the lower bound (8). This task can be achieved by considering the following problem:

$$\min_{\tau_i,\tau_j\in\Theta_n} f_n(\tau_i, \mathbf{D}_{i,\cdot}) + f_n(\tau_j, \mathbf{D}_{j,\cdot}) \text{ subject to } \tau_{ij} = \tau_{ji}. \quad (13)$$

The Lagrangian dual of (13) is

$$\begin{aligned} & \max_{\lambda\in\mathbb{R}} \min_{\tau_i,\tau_j\in\Theta_n} \left(f_n(\tau_i, \mathbf{D}_{i,\cdot}) + f_n(\tau_j, \mathbf{D}_{j,\cdot}) + \lambda\,(2^{-\tau_{ij}} - 2^{-\tau_{ji}}) \right) \quad (14) \\ & = \max_{\lambda\in\mathbb{R}}(\min_{\tau_i\in\Theta_n} f_n(\tau_i, \mathbf{D}'_{i,\cdot}) + \min_{\tau_j\in\Theta_n} f_n(\tau_j, \mathbf{D}'_{j,\cdot})) \end{aligned}$$

where $\mathbf{D}'_{i,j} = \mathbf{D}_{i,j} + \lambda$, $\mathbf{D}'_{i,k} = \mathbf{D}_{i,k}$ for $k \neq i, j$, and $\mathbf{D}'_{j,i} = \mathbf{D}_{j,i} - \lambda$, $\mathbf{D}'_{j,k} = \mathbf{D}_{j,k}$ for $k \neq \{i, j\}$. Problem (14) naturally splits into two separate optimization subproblems each of which can be solved by means of the recusion (12). Hence, the Lagrangian dual can be solved, at least sub-optimally, by iteratively solving the two subproblems while optimizing with respect to λ according to the standard subgradient method. It is worth noting that a similar result can be derived for multiple symmetry constraints and for any constraint on a topological distance like $\tau_{ij} \geq k$, $k \in \mathbb{R}_{0^+}$. The latter is typically imposed in a branch and bound framework when a partial description of a phylogeny is present (see e.g., [2]).

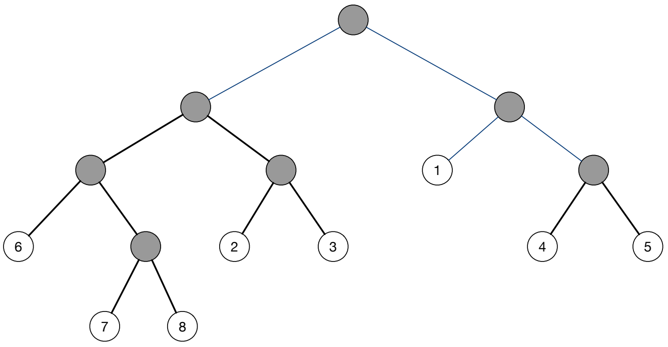

Figure 1: Unsorted optimal path-lengths for the instance of the MREP constituted by the vector $(3,4,4,4,4,4,5,5)$.

Acknowledgements

D. Catanzaro acknowledges support from the FRS-FNRS via the grant "Crédit de Recherche" ref. S/25-MCF/OL J.0026.17, the Université Catholique de Louvain via the "Fonds Spéciaux de Recherche" (FSR) 2017-2019, and the Fondation Louvain via the grant "Le numérique au service de l'humain". Part of this work has been developed when D. Catanzaro was Invited Professor at the Department of Management of University Ca Foscari of Venice, Italy, during the academic year 2018.

References

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

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

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

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

[5] F. Pardi, Algorithms on Phylogenetic Trees, Ph.D. thesis, University of Cambridge, UK, 2009.

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

[7] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2004.

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

[9] C. Semple, M. Steel, Phylogenetics, Oxford University Press, New York, 2003.

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

[11] R. Desper, 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 (2004) 587–598.

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

[13] S. Fiorini, G. Joret, Approximating the balanced minimum evolution problem, Operations Research Letters 40 (2012) 31–35.

[14] D. A. Huffman, A method for the construction of minimum redundancy codes, in: Proceedings of the IRE, 1952.

[15] D. S. Parker, P. Ram, The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees, SIAM Journal on Computing 28 (1996) 1875–1905.

[16] K. Sayood, Introduction to Data Compression, 5th ed., Morgan Kaufmann, San Francisco, CA, 2017.

[17] G. Böcherer, R. A. Amjad, Informational divergence and entropy rate on rooted trees with probabilities, in: IEEE International Symposium on Information Theory, IEEE Computer Society, 2014.

[18] T. Hirschler, W. Woess, Comparing entropy rates on finite and infinite rooted trees, IEEE Transactions on Information Theory (2017) 1–1.

[19] M. V. A. Batista, T. A. E. Ferreira, A. C. Freitas, V. Q. Balbino, An entropy-based approach for the identification of phylogenetically informative genomic regions of papillomavirus, Infection, Genetics and Evolution 11 (2011) 2026–2033.

[20] F. Bay, J. Xu, L. Liu, Weighted relative entropy for phylogenetic tree based on 2-step markov model, Mathematical Biosciences 246 (2013) 8–13.

[21] P. O. Lewis, M.-H. Chen, L. Kuo, L. A. Lewis, K. Fučikova, S. Neupane, Y.-B. Wang, D. Shi, Estimating Bayesian phylogenetic information content, Systematic Biology 65 (2016) 1009–1023.

[22] S. Pavoine, S. Ollier, A. B. Dufour, Is the originality of a species measurable?, Ecology Letters 8 (2005) 579–586.

[23] J. P. Townsed, Z. Su, Y. I. Tekle, Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny, Systematic Biology 61 (2012) 835–849.

[24] D. R. Brooks, J. Collier, B. A. Maurer, J. D. H. Smith, E. O. Wiley, Entropy and information in evolving biological systems, Biology and Philosophy 4 (1989) 407–432.

[25] C. R. Rao, Diversity and dissimilarity coefficients: A unified approach, Journal of Theoretical Population Biology 21 (1982) 24–43.

[26] A. Chao, C.-H. Chiu, L. Jost, Phylogenetic diversity measures and their decomposition: A framework based on Hill numbers, in: R. Pellens, P. Grandcolas (Eds.), Biodiversity Conservation and Phylogenetic Systematics, Springer, NY, 2016.

[27] B. Allen, M. Kon, Y. Bar-Yam, A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats, The American Naturalist 174 (2009) 236–243.

[28] D. Bokal, M. DeVos, S. Klavžar, A. Mimoto, A. O. Mooers, Computing quadratic entropy in evolutionary trees, Computers and Mathematics with Applications 62 (2011) 3821–3828.

[29] L. Brocchieri, Phylogenetic diversity and the evolution of molecular sequences, Journal of phylogenetics and evolutionary biology 3 (2015) 1000e109.

[30] M. A. Mouchet, D. Mouillot, Decomposing phylogenetic entropy into α, β, and γ components, Biology Letters 7 (2011) 205–209.

[31] S. Pavoine, S. Ollier, D. Pontier, Measuring diversity from dissimilarities with rao's quadratic entropy: Are any dissimilarities suitable?, Journal of Theoretical Population Biology 67 (2005) 231–239.

[32] D. Catanzaro, C. Engelbeen, An integer linear programming formulation for the minimum cardinality segmentation problem, Algorithms 8 (2015) 999–1020.

[33] D. Catanzaro, R. Pesenti, The parsimonious phylogeny estimation problem, Technical Report TR-20180108, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Belgium, 2018.

[34] D. Catanzaro, R. Pesenti, L. A. Wolsey, On the balanced minimum evolution polytope, Technical Report TR-20190114, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Belgium, https://perso.uclouvain.be/daniele.catanzaro/myarticles/TR-20190114.pdf, submitted, 2019.

[35] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423.

[36] R. A. Brualdi, The DAD theorem for arbitrary row sums, in: Proceedings of the American Mathematical Society, volume 45, 1974, pp. 189–194.

[37] M. Idel, A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps, arXiv: 1609.06349, 2016.

[38] I. Sason, S. Verdú, Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets, in: IEEE Information Theory Workshop, 2015, pp. 214–218.