ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230605

Highly-Efficient Robinson-Foulds Distance Estimation with Matrix Correction

Fangchen Yu[a], Rui Bao[a], Jianfeng Mao[a,b] and Wenye Li[a,b;*]

[a]The Chinese University of Hong Kong, Shenzhen, China
[b]Shenzhen Research Institute of Big Data, Shenzhen, China
fangchenyu@link.cuhk.edu.cn, ruibao@link.cuhk.edu.cn, jfmao@cuhk.edu.cn, wyli@cuhk.edu.cn

Abstract. Phylogenetic trees are essential in studying evolutionary relationships, and the Robinson-Foulds (RF) distance is a widely used metric to calculate pairwise dissimilarities between phylogenetic trees, with various applications in both the biology and computing communities. However, generating a precise RF distance matrix becomes difficult or even intractable when tree information is partially missing. To address this issue, we introduce a novel distance correction algorithm for estimating the RF distance matrix of incomplete phylogenetic trees. Our method innovatively harnesses the assumption of Euclidean embedding, correcting an approximate distance matrix into a valid distance metric, guaranteed to be closer to the unknown ground-truth. Despite its simplicity, our approach exhibits robust performance, efficiency, and scalability in empirical evaluations, outperforming classical distance correction algorithms and holding potential benefits in downstream applications. Our code is available at https://github.com/CUHKSZ-Yu/EMC.

1 Introduction

Phylogenetic trees play a crucial role in biology as they provide a visual representation of evolutionary relationships among various biological entities through an acyclic graph with labeled leaves. Effective use of such trees in practical applications, including supertree construction [13, 21, 30], phylogenetic database searching [26, 37, 38], and gene tree clustering [1, 35, 41], requires distance calculation across trees. Several distance measures have been developed to systematically compare different phylogenetic trees, with the most popular distance metric being the Robinson-Foulds (RF) distance [29]. Other essential distance metrics include quartet distance [8], nearest neighbor interchange distance [18], and subtree-prune-and-regraft operations [33]. These tree distance measures facilitate accurate and efficient analysis of evolutionary relationships between biological entities represented by phylogenetic trees.

Calculating pairwise tree distance traditionally presupposes that two trees share identical leaf sets. When trees have non-identical leaf sets, two pre-processing techniques can be used: restriction on shared leaf sets [10] and completion on total leaf sets [3, 40]. While comprehensive research exists on complete phylogenetic trees with full label information, the comparison of two incomplete trees with unknown leaf labels, illustrated in Fig. 1, remains a substantial challenge in phylogenetics and evolutionary biology, particularly when considering uncertainty and noise. Despite its importance, this problem has received relatively little attention, necessitating further exploration to improve the accuracy of phylogenetic tree analysis.

* Corresponding Author. Email: wyli@cuhk.edu.cn.

In this study, we focus on estimating a high-quality distance metric using the widely used RF distance measure for incomplete phylogenetic trees. Conventionally, completion techniques are deployed to address missing labels by exploring every permutation of these labels to minimize the distance between two incomplete trees. However, these methods heavily rely on the tree structure and leaf information, and they can be computationally demanding and impractical for large leaf sets due to the vast search space. Alternatively, pairwise distance estimation approximates the distance between two incomplete trees by treating known leaf positions as complete trees, but this approach might produce a non-metric distance matrix that falls short of satisfying metric properties like triangle inequalities. Consequently, the difficulties imposed by incomplete observations and metric requisites render this problem challenging to solve.

To provide a dependable solution, we suggest a fundamentally different approach, i.e., *matrix correction*. Rather than completing the missing information, we initiate with an approximate estimation of the tree distance matrix, subsequently correcting it to fulfill certain metric properties [7, 15, 25, 34]. Specifically, our work utilizes the property of Euclidean embedding [32] and ingeniously devises a novel matrix correction method [24]. Under a very mild assumption, the corrected matrix is proven to be closer to the unknown ground-truth with a solid guarantee. Aligned with the theoretical observation, our method demonstrates better efficiency and effectiveness than classical correction methods [7, 15] in empirical evaluations and has promising potential to enhance downstream applications.

In short, our motivations and contributions are listed as follows:

- **[Motivation]** Our work is motivated by the practical problem of estimating tree distances for incomplete phylogenetic trees. To address this issue, we develop a new distance correction method and validate its performance on a widely used RF distance metric, providing practical tools and useful insights to both the biology and computing communities.
- **[Innovation]** We engineer a universal and completion-free strategy that directly corrects RF distance matrices by leveraging Euclidean embedding and optimization techniques. Compared with the standard completion and correction methods, our proposed approach is fundamentally different and provides a theoretically reliable estimate of the unknown ground-truth.
- **[Performance]** Simple yet efficient, our approach excels in empirical evaluations of RF distance estimation, achieves an improved estimate over classical methods with better scalability and stability, and showcases its valuable impact in downstream applications, such as distance denoising and tree clustering.

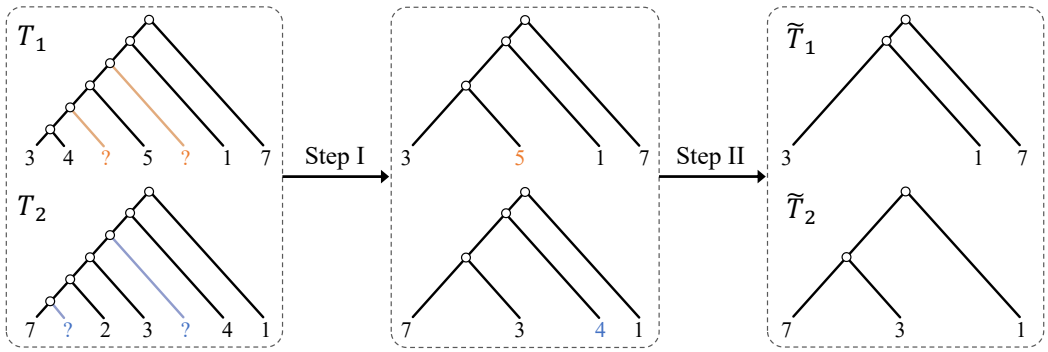

Figure 1. Two-step approximation of the pairwise Robinson-Foulds (RF) distance for incomplete phylogenetic trees. Step I is to restrict the incomplete trees T_1, T_2 on commonly known leaves. Step II is to compute the RF distance on the shared leaf sets. Eq. (2) gives the approximated RF distance $d^0_{\mathrm{RF}}(T_1, T_2)$.

2 Background

2.1 The Robinson-Foulds (RF) Distance

The RF distance [29] is a standard metric for comparing phylogenetic trees, which can be computed in linear time in the size of leaf sets [27]. Given a phylogenetic tree T, let $E(T)$ be the edge set and $L(T)$ be the set of leaf labels. In this paper, all phylogenetic trees considered are *rooted leaf-labeled binary trees* with l leaves, whose leaves are uniquely labeled with integers from 1 to l, i.e., $L(T) = \{1, 2, \ldots, l\}$. Denote $B(T)$ as the set of all bipartitions producible by removing any edge $e \in E(T)$. For any phylogenetic trees T_i and T_j on the same label set, the RF distance is defined as the size of the symmetric difference between the sets of bipartitions:

$$d_{\mathrm{RF}}(T_i, T_j) = |B(T_i) \oplus B(T_j)|, \tag{1}$$

where the symmetric difference defined on two sets X and Y is $X \oplus Y = (X \backslash Y) \cup (Y \backslash X)$ and thus $|X \oplus Y| = |X| + |Y| - 2|X \cap Y|$.

The pairwise distance computation for a set of complete phylogenetic trees is quite straightforward. Meanwhile, the RF distance matrix calculated by Eq. (1) is shown to be a distance metric [29]. In particular, a *distance metric* is defined as a matrix $D \in \mathbb{R}^{n \times n}$ satisfying $d_{ii} = 0, d_{ij} = d_{ji} \geq 0, d_{ij} \leq d_{ik} + d_{kj}, \forall\, 1 \leq i, j, k \leq n$. However, for a set of incomplete phylogenetic trees, calculating the accurate pairwise distance becomes infeasible due to unknown labels. The approximation has to be sought, and the most straightforward way is to first restrict the incomplete trees on commonly known leaves. Then the pairwise distance between these two pruned complete trees with the same tree structure can be computed by Eq. (1) on shared known label sets. As shown in Fig. 1, the approximate distance for the two incomplete trees can be reasonably defined as

$$d^0_{ij} = d^0_{\mathrm{RF}}(T_i, T_j) = d_{\mathrm{RF}}(\widetilde{T}_i, \widetilde{T}_j) \cdot \frac{|L(T_i)|}{|L(\widetilde{T}_i)|}. \tag{2}$$

Unfortunately, the approximate RF distance matrix $D^0 = \{d^0_{ij}\}^n_{i,j=1}$ may no longer satisfy the criteria of a distance metric and often breaches some triangle inequalities due to the incomplete label information, which motivates us for further correction and improvement.

2.2 Distance Metric Correction

Significant advancements [2, 28, 31] have been made in the computing community towards correcting a non-metric distance to a high-quality metric distance. Notably, two classical strategies, each employing distinct optimization techniques, are taken into account.

The first strategy involves the correction of the non-metric distance matrix D^0 to ensure the satisfaction of all triangle inequalities. This process is referred to as the metric nearness model [7, 25] and is formulated as:

$$\min_{D \in \mathbb{R}^{n \times n}} \|D - D^0\|^2_F, \tag{3}$$

subject to $d_{ii} = 0, d_{ij} = d_{ji} \geq 0, d_{ij} \leq d_{ik} + d_{kj}$ for all $1 \leq i, j, k \leq n$, which seeks the closest distance matrix in the metric space as the optimal approximation.

Modern optimization solvers such as CPLEX and MOSEK can be used to solve this strongly convex problem [5] in Eq. (3) , but they only work on a small scale. Besides, the Triangle Fixing (TRF) algorithm [7] was proposed to efficiently iterate over the triangle inequalities based on a primal-dual method. Despite the partial progress, the scalability of these solutions is severely limited by a large number of $O(n^3)$ triangle inequality constraints inherent in the feasible region.

The second strategy aims to correct the non-metric distance matrix into a *Euclidean distance matrix* (EDM) [20]. An EDM is defined as a $n \times n$ real symmetric (squared) matrix that guarantees the existence of $x_1, \cdots, x_n$ in a Euclidean space and satisfies

$$d_{ij} = \|x_i - x_j\|^2_2, \ \forall\, 1 \leq i, j \leq n. \tag{4}$$

Intrinsically, the square root of the EDM naturally satisfies the triangle inequalities, thereby becoming a valid distance metric. A well-known property provides a sufficient and necessary condition [17]:

Lemma 1. *$D \in \mathbb{R}^{n \times n}$ is an EDM if and only if $S = -\frac{1}{2} JDJ$ is positive semi-definite (PSD), where $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix with an identity matrix I and a vector of ones $\mathbf{1}$.*

In general, these EDM-based methods [15, 28] initially convert the non-metric squared distance matrix D to the similarity matrix $S = -\frac{1}{2} JDJ$ by the Double-Centering (DC) algorithm [15]. They then execute truncation operations on the eigenvalues of S to ensure the positive semi-definiteness of the matrix S. The similarity matrix is finally reverted to the distance matrix, resulting in an EDM and laying the foundation for the multi-dimensional scaling (MDS) method [36]. It can be seen that both matrices D and S are closely related to each other, however, the conversion between D and S may be inaccurate, and significant information could be lost by simply truncating the negative eigenvalues [15].

Straightforward as they are, both correction strategies have certain limitations, including high computational complexity, limited scalability, and low accuracy, which potentially hinder their wide applications in practice.

3 Methodology

Our proposed approach [24] stands apart from previous work by developing a brand-new framework that utilizes the powerful embedding property of Euclidean space. Leveraging this, we formulate a new matrix correction problem and adopt an efficient alternating projection algorithm [12] to obtain the optimal solution with a robust theoretical guarantee. Through the development of a unique methodology, we have created a highly effective strategy that marks a substantial leap forward from existing works.

3.1 Embedding Technique

Our approach [24] differs from the classical strategy of assuming a Euclidean distance matrix (EDM). Instead, we adopt a more flexible assumption: the distance matrix can be isometrically embedded in Euclidean space given an underlying data representation in the vector space. To be specific, we introduce a definition that allows for greater adaptability in our approach:

Definition 1. *A distance matrix $D = \{d_{ij}\}_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is said to be isometrically embeddable in Euclidean space if there exists a set of data points $\{x_1, \ldots, x_n\}$ in the vector space with a distance function ρ, having the properties that $\rho(x_i, x_i) = 0$ and $\rho(x_i, x_j) = \rho(x_j, x_i) \geq 0$ for all points x_i and x_j in the set, such that*

$$d_{ij} = \rho(x_i, x_j),\ \forall\, 1 \leq i, j \leq n.$$

It is worth noting that the squared Euclidean distance used in the EDM-based strategy is actually a specific instance of the generalized distance function ρ, where $\rho(x_i, x_j) = \|x_i - x_j\|_2^2$. Our approach, based on isometrical embedding, enables us to utilize Schoenberg's characterization [32], a classical result [39] that provides an equivalent transformation for isometrically embeddable matrices:

Lemma 2. *$D = \{d_{ij}\}_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is isometrically embeddable in Euclidean space if and only if the kernel matrix $K = \exp(-\gamma D) = \{\exp(-\gamma d_{ij})\}_{i,j=1}^n$ is positive semi-definite for any $\gamma > 0$.*

Similar to the EDM-based strategy, we establish a connection between the distance matrix D and kernel matrix K. This enables us to formulate the matrix correction problem in a natural way, thereby promoting efficient application of our method.

3.2 Distance Correction Algorithm

Given a non-metric distance matrix D^0, we naturally aim to correct it to an isometrically embeddable distance matrix $\hat{D}$ that better approximates the true pairwise distances between the data points. Thus, we utilize the approach [24] with the following optimization problem:

$$\min_{D\in\mathbb{R}^{n\times n}} \|D - D^0\|_F^2, \tag{5}$$

subject to the constraints:

$$\begin{cases} d_{ii} = 0,\ \forall\, 1 \leq i \leq n, \\ d_{ij} = d_{ji} \geq 0,\ \forall\, 1 \leq i \neq j \leq n, \\ \exp(-\gamma D) \succeq 0, \end{cases}$$

where $\succeq 0$ denotes the positive semi-definiteness (PSD), and the PSD constraint ensures the corrected matrix is isometrically embeddable. Unfortunately, solving the above-defined optimization problem is challenging due to the PSD constraint in exponential form. To address this issue, we change the decision variable from D to $K = \exp(-\gamma D)$ and reformulate the problem in a more tractable form under an efficient approximation:

$$\min_{K\in\mathbb{R}^{n\times n}} \|K - K^0\|_F^2, \tag{6}$$

subject to the constraints:

$$\begin{cases} k_{ii} = 1,\ \forall\, 1 \leq i \leq n, \\ k_{ij} = k_{ji} \in [0, 1],\ \forall\, 1 \leq i \neq j \leq n, \\ K \succeq 0. \end{cases}$$

We define the feasible region in Eq. (6) as C, a closed convex set, from which the optimal solution $\hat{K}$ is the projection of K^0 onto the set C. Consequently, the optimal embeddable matrix is derived as $\hat{D} \approx -\frac{1}{\gamma}\log(\hat{K})$ with $\gamma = \frac{0.02}{\max\{d_{ij}^0\}}$ (default value in [24]).

Performing a direct projection from K^0 onto C is complex and computationally demanding. However, the elegant structure of this reformulation leads us to the well-established *Alternating Projection* algorithm [16] from the optimization community. In this context, the feasible region C can be seen as the intersection of two less complex, closed convex subsets C_1 and C_2:

$$\begin{cases} C_1 = \{X \in \mathbb{R}^{n\times n} | X \succeq 0\}, \\ C_2 = \{X \in \mathbb{R}^{n\times n} | x_{ii} = 1, x_{ij} = x_{ji} \in [0, 1],\ \text{for all } i, j\}. \end{cases}$$

The optimal solution $\hat{K}$ can subsequently be efficiently computed by iteratively projecting K^0 onto C_1 and C_2, with assured convergence [9]. Denote P_1 and P_2 as the projection onto C_1 and C_2, respectively:

$$\begin{cases} P_1(K) = U\{\max\{\Sigma_{ij}, 0\}\}_{i,j=1}^n V^\top, \\ P_2(K) = \{\text{median}\{0, k_{ij}, 1\}\}_{i,j=1}^n. \end{cases}$$

where $U\Sigma V^\top$ gives the singular value decompostion (SVD) of K.

We employ Dykstra's alternating projection algorithm [12, 22, 23, 42] to find the optimal solution, expressed as follows:

$$\begin{cases} X_0^{(t)} = X_2^{(t-1)}, \\ Z = X_{i-1}^{(t)} + Y_i^{(t-1)}, \\ X_i^{(t)} = P_i(Z), \\ Y_i^{(t)} = Z - P_i(Z), \end{cases} \tag{7}$$

for $i = 1, 2$ and $t = 1, 2, \cdots$, where $Y_1^{(0)} = Y_2^{(0)} = \mathbf{0}$, $X_2^{(0)} = K^0$, and $\mathbf{0}$ is an all-zero matrix of suitable size. Based on the Boyle-Dykstra convergence result [6], both $\{X_1^{(t)}\}$ and $\{X_2^{(t)}\}$ generated by Eq. (7) converge in the Frobenius norm to the unique optimal solution $\hat{K}$ of $\min_{K\in C=C_1\cap C_2} \|K - K^0\|_F^2$.

In such cases, our new matrix correction problem in Eq. (5) is solved efficiently with guaranteed and fast convergence, the details of which are summarized in Algorithm 1.

3.3 Theoretical Analysis

Performance Guarantee. Beyond the guaranteed convergence described earlier, our proposed approach also offers a compelling theoretical guarantee [24] regarding the correction performance, as captured in Theorem 1. Essentially, if the initial estimate D^0 is not isometrically embeddable, our algorithm can enhance it to a superior one, denoted by $\hat{D}$, closer to the unknown ground-truth D^*.

Algorithm 1 Embedding-based Matrix Correction (EMC)

Input: $D^0 \in \mathbb{R}^{n\times n}$: a real symmetric non-metric matrix; $maxiter$: maximum of iterations; γ: hyper-parameter (default $\frac{0.02}{\max\{d_{ij}^0\}}$); tol: tolerence (default 10^{-5}).

Output: $\hat{D} \in \mathbb{R}^{n\times n}$: optimal corrected distance metric.

Set $K^0 = \exp(-\gamma D^0), Y_1^{(0)} = Y_2^{(0)} = \mathbf{0}, X_2^{(0)} = K^0$.
for $t = 1, 2, \ldots, maxiter$ **do**
  $X_0^{(t)} \leftarrow X_2^{(t-1)}$.
  for $i = 1, 2$ **do**
    $Z \leftarrow X_{i-1}^{(t)} + Y_i^{(t-1)}$;
    $X_i^{(t)} \leftarrow P_i(Z)$;
    $Y_i^{(t)} \leftarrow Z - P_i(Z)$.
  end for
  if $\|X_1^{(t)} - X_1^{(t-1)}\|_F < tol$ **then**
    break
  end if
end for
Set $\hat{K} = X_1^{(t)}, \hat{D} = -\frac{1}{\gamma}\log(\hat{K})$.

Theorem 1. *Let D^0 be the initial non-metric matrix, $\hat{D}$ be the corrected distance matrix obtained from Eq. (5), D^* be the unknown ground-truth assumed to be isometrically embeddable, then we have*

$$\|D^* - \hat{D}\|_F^2 \le \|D^* - D^0\|_F^2.$$

Proof. Denote by D^* the true but unknown metric and $K^* = \exp(-\gamma D^*)$. It can be observed that $K^* \in C$ and K^* satisfies

$$\begin{aligned}\|K^* - \hat{K}\|_F^2 &\le \|K^* - \hat{K}\|_F^2 - 2\langle K^* - \hat{K}, K^0 - \hat{K}\rangle \\ &\le \|(K^* - \hat{K}) - (K^0 - \hat{K})\|_F^2 \\ &= \|K^* - K^0\|_F^2,\end{aligned} \tag{8}$$

where $\langle A, B\rangle = trace(A^T B)$ is an inner product defined on the closed convex set C and the first "$\le$" holds due to Kolmogrov's criterion [11, 24]. When $\gamma = \frac{\epsilon}{\max\{d_{ij}^0\}}$ and ϵ is sufficiently small, we can employ the Taylor-series expansion of exponential data to derive:

$$\begin{cases} k_{ij} = \exp(-\gamma d_{ij}) = 1 - \gamma d_{ij} + O(\epsilon^2), \\ k_{ij}^0 = \exp(-\gamma d_{ij}^0) = 1 - \gamma d_{ij}^0 + O(\epsilon^2). \end{cases}$$

Next, we connect the elements between D and K:

$$(d_{ij} - d_{ij}^0)^2 = \frac{1}{\gamma^2}(k_{ij} - k_{ij}^0)^2 + O(\epsilon^2),$$

which means

$$\|D - D^0\|_F^2 = \frac{1}{\gamma^2}\|K - K^0\|_F^2 + O(\epsilon^2).$$

For a sufficiently small ϵ, the inequality in Eq. (8) is equivalent to

$$\|D^* - \hat{D}\|_F^2 \le \|D^* - D^0\|_F^2, \tag{9}$$

where "$=$" holds if and only if $\hat{D} = D^0$. □

Complexity Analysis. The per-iteration time complexity of Algorithm 1 is $O(n^3)$, mainly derived from the singular value decomposition (SVD) in the projection operation P_1. This can be expedited further using a randomized SVD [19] or parallel SVD [4]. The storage complexity is $O(n^2)$ to store the whole distance matrix.

Algorithm Extensions. We can augment our work further through two types of scalable extensions. (i) A divide-and-conquer strategy can be utilized to break a large $n \times n$ matrix into smaller principal sub-matrices, thereby approximating the projection result. (ii) Dykstra's projection could be transformed into a cyclic projection [14] on multiple closed convex subsets using a dual-primal method with a faster convergence rate. These techniques can significantly reduce the time and space complexity of the algorithm, thereby improving its execution speed and scalability.

Application Scopes. Our method offers a general approach for distance correction, applicable to tree distance, Euclidean distance, and other distance metrics, extending its scope significantly. In this work, we focus on the RF distance, a commonly utilized metric for tree comparison, applicable not only to phylogenetic trees but also to binary trees. We anticipate that the more accurate distance metrics corrected by our method could enhance distance-based algorithms in downstream applications, such as tree clustering and tree retrieval.

4 Experiments

To evaluate the effectiveness of the proposed method, we carry out a series of empirical studies focusing on:

- the quality of the corrected distance matrix from incomplete phylogenetic trees;
- the noise reduction effect on the noisy RF distance matrix;
- the improvement in sensitivity and scalability;
- the benefit for the tree clustering application.

We compare our proposed **E**mbedding-based **M**atrix **C**orrection algorithm, referred to as **EMC**, with two widely used correction algorithms in practice: the Triangle Fixing (**TRF**) algorithm [7] and the Double-Centering (**DC**) algorithm in tandem with clip-operation [15]. The specifics of these baseline methods, which represent two classical correction strategies — the metric nearness strategy and the EDM-based strategy respectively, are detailed in Section 2.2. All experiments are conducted five times using MATLAB on a ThinkStation P360 workstation equipped with a 2.1 GHz Intel i7-12700 Core and 32 GB of RAM.

4.1 Correction on Incomplete Trees

We apply the proposed approach to dealing with incomplete phylogenetic trees with some unknown leaf labels. As depicted in Fig. 1, the pairwise RF distance for two incomplete trees is initially approximated by Eq. (2) through pruning and restricting operations. However, due to incomplete observation, the approximate distance matrix usually fails to satisfy metric properties, which is supposed to be a metric through correction methods for downstream applications.

In this experiment, we randomly generate a set of n complete phylogenetic trees $\{T_1^*, T_2^*, \ldots, T_n^*\}$ on the same label set $L = \{1, 2, \ldots, l\}$ with the same tree structure. Accordingly, we generate the set of incomplete phylogenetic trees $\{T_1^0, T_2^0, \ldots, T_n^0\}$ by removing the label information of leaves completely at random for a given missing ratio r varying from 40% to 80%. The RF distance matrix $D^* \in \mathbb{R}^{n\times n}$ calculated by the set of complete trees is regarded as the ground-truth, which is definitely a distance metric. The RF distance matrix $D^0 \in \mathbb{R}^{n\times n}$ approximated from incomplete trees[1] is set as our input matrix, which usually dissatisfies the metric properties.

[1] Note that for any two incomplete trees T_i^0, T_j^0 with no common known labels, $d_{\mathrm{RF}}^0(T_i^0, T_j^0) = |L(T_i^0)| + |L(T_j^0)| = 2l$.

Table 1. Distance metric correction for incomplete phylogenetic trees. Given a set of n incomplete phylogenetic trees with l leaves and a missing ratio r, the proposed EMC approach obtains the smallest Mean Squared Error (MSE), showing evident superiority over baselines. Best results are highlighted in **bold**.

Missing Ratio r	# leaves # trees	$l = 10$				$l = 20$			
		$n = 100$	$n = 200$	$n = 500$	$n = 1,000$	$n = 100$	$n = 200$	$n = 500$	$n = 1,000$
40%	D^0	166.6 ± 1.5	167.8 ± 0.9	168.2 ± 0.3	168.5 ± 0.1	705.0 ± 8.6	712.8 ± 2.2	713.0 ± 1.4	713.3 ± 0.4
	DC	280.4 ± 1.9	522.4 ± 2.5	1064.1 ± 0.8	1728.3 ± 1.1	701.4 ± 6.0	1490.0 ± 4.6	3337.8 ± 2.4	5693.9 ± 7.2
	TRF	81.9 ± 1.7	81.6 ± 0.5	81.1 ± 0.4	81.0 ± 0.1	418.1 ± 9.7	422.9 ± 2.4	418.5 ± 2.1	418.2 ± 0.4
	EMC	$\mathbf{53.2 \pm 2.0}$	$\mathbf{57.8 \pm 1.0}$	$\mathbf{64.0 \pm 0.4}$	$\mathbf{68.6 \pm 0.2}$	$\mathbf{314.5 \pm 10.0}$	$\mathbf{333.6 \pm 3.4}$	$\mathbf{355.9 \pm 1.8}$	$\mathbf{373.2 \pm 0.4}$
60%	D^0	220.6 ± 1.4	221.3 ± 0.4	222.3 ± 0.2	222.5 ± 0.1	1148.9 ± 2.4	1155.5 ± 3.0	1157.8 ± 0.7	1158.2 ± 0.4
	DC	189.9 ± 8.3	348.9 ± 4.8	659.1 ± 2.8	1002.5 ± 0.8	973.4 ± 16.3	1216.4 ± 21.9	2836.6 ± 9.3	4894.1 ± 6.8
	TRF	160.7 ± 3.3	158.7 ± 0.7	159.4 ± 0.5	161.6 ± 0.2	899.3 ± 8.5	899.8 ± 6.9	896.5 ± 1.4	913.0 ± 0.8
	EMC	$\mathbf{140.3 \pm 3.8}$	$\mathbf{147.3 \pm 1.1}$	$\mathbf{156.7 \pm 0.5}$	$\mathbf{159.3 \pm 0.2}$	$\mathbf{808.9 \pm 5.9}$	$\mathbf{849.3 \pm 7.2}$	$\mathbf{887.6 \pm 1.3}$	$\mathbf{895.1 \pm 0.9}$
80%	D^0	107.2 ± 0.8	106.5 ± 0.3	105.8 ± 0.4	105.3 ± 0.1	791.4 ± 2.5	805.9 ± 4.1	806.3 ± 3.0	807.3 ± 0.8
	DC	27.5 ± 0.5	35.3 ± 0.3	39.3 ± 0.5	40.7 ± 0.1	491.1 ± 11.2	882.5 ± 12.5	1544.6 ± 15.8	2065.9 ± 8.1
	TRF	37.7 ± 1.2	36.9 ± 0.2	36.3 ± 0.2	36.1 ± 0.1	398.5 ± 4.2	409.6 ± 3.2	405.2 ± 2.8	405.1 ± 0.9
	EMC	$\mathbf{21.5 \pm 1.2}$	$\mathbf{20.7 \pm 0.4}$	$\mathbf{22.0 \pm 0.3}$	$\mathbf{23.1 \pm 0.1}$	$\mathbf{263.0 \pm 1.6}$	$\mathbf{286.6 \pm 4.9}$	$\mathbf{302.4 \pm 3.1}$	$\mathbf{312.0 \pm 1.0}$

Finally, we correct D^0 to $\hat{D}$ by different approaches, and calculate the Mean Squared Error (**MSE**) from the ground-truth D^* as the evaluation metric:

$$\text{MSE} = \frac{1}{n^2} \sum_{1 \le i,j \le n} (\hat{D}_{ij} - D^*_{ij})^2.$$

We evaluate all methods on our synthetic tree dataset under various settings, i.e., the number of leaves l, number of trees n, and missing ratio r. The average performance is reported in Table 1, demonstrating that our EMC method consistently outperforms the baseline methods in all experiments, achieving a lower MSE.

Performance Comparison. As Table 1 reveals, both the TRF and EMC methods provide improved results with smaller MSEs than D^0. In contrast, the DC method, which lacks a guarantee, occasionally exhibits an MSE exceeding 1,000. This could occur when, particularly for large matrices, the naïve truncation of negative eigenvalues leads to the loss of relevant information, resulting in poorer distance estimation. Our EMC approach, on the other hand, offers a superior solution grounded on a more robust embedding property, guaranteeing performance instead of relying on the Euclidean distance matrix.

Effect of Tree Settings. Analyzing different tree settings, we observe that (i) with an increasing size n of the tree sets, the TRF and EMC methods exhibit stable performance, while the MSE of the DC method significantly escalates; (ii) with a larger size l of leaf sets, the correction algorithms face greater challenges, potentially leading to higher MSEs; (iii) with a higher missing ratio r (e.g., 80%), the element D^0_{ij} approaches to the upper bound $2l$, and D^0 may be closer to the ground-truth D^* and leaves a smaller room for further correction, resulting in lower MSEs.

In short, the proposed EMC method achieves consistently superior results on the MSE evaluation, with no assumptions on the missing ratio or mechanism, which justifies its effectiveness on distance estimation for incomplete phylogenetic trees.

4.2 Correction on Noisy Tree Distance

In real-world scenarios, there commonly exists noise or error during the distance measurement, as well as incompleteness in label sets due to data corruption. The provided distance matrix with noisy observation also faces the similar challenge of not meeting the metric properties, which is another application scenario for matrix correction algorithms in noisy tree distance.

In each run of this experiment, we randomly generate a set of n complete phylogenetic trees as described in Section 4.1. Then, we calculate the RF distance matrix D^* on the complete tree set as the ground-truth. Next, we add Gaussian noise to obtain a noisy distance matrix D^0 whose elements are given by

$$d^0_{ij} = \max\{0, d^*_{ij} + \zeta\mu v\},$$

where ζ denotes the noise level ranging from 0.4 to 0.8, μ is the mean of all elements in D^*, and $v \sim N(0,1)$ is a standard Gaussian random variable. Take D^0 as our input. We correct D^0 by different correction algorithms and also record the average of the Mean Squared Error (MSE) over five runs.

Similar to the results in Table 1, the proposed correction approach reports significantly improved MSE results in Table 2. With different noise levels ζ varying from 0.4 to 0.8, both TRF and EMC correction approaches consistently reduce the MSE values over the approximate matrix D^0 across all settings, between which the proposed EMC approach performs better. More detailed, we can see that the EMC approach brings significant drops in the MSE from the ground-truth D^*, reducing even more than 90% error when $\zeta = 0.6$ or 0.8. In contrast, the corrected matrices from the DC approach always perform worse than D^0, and MSE values even exceed 10,000, which indicates the corrected distance value is far away from the ground-truth and is unsuitable for noise reduction tasks.

All the results, although preliminary, clearly justify the benefits of the proposed approach and demonstrate its high potential in practical tasks at the noisy tree distance.

4.3 Sensitivity Analysis

We conduct a sensitivity analysis experiment on the synthetic tree dataset with varying missing ratios and noise levels. We vary r in $[20\%, 80\%]$ and ζ in $[0.2, 0.8]$ to examine how the correction performance changes. In addition to the MSE, we also assess performance using another evaluation metric, the Mean Absolute Error (**MAE**), defined as

$$\text{MAE} = \frac{1}{n^2} \sum_{1 \le i,j \le n} |\hat{D}_{ij} - D^*_{ij}|.$$

It is evident from the results in Fig. 2 that our EMC method's performance consistently outperforms other correction methods under various settings. The robust performance across a wide range of parameter settings validates the effectiveness of our proposed approach.

Table 2. Noise reduction on noisy Robinson-Foulds distance matrix under Mean Squared Error (MSE) measure on different tree settings (i.e., n trees with l leaves) and various noise levels ζ. The proposed EMC approach shows significant improvement on MSE in all experiments, which also justifies the theoretical guarantee provided in Theorem 1. The best performances are highlighted in **bold**.

Noise	# leaves	$l = 10$				$l = 20$			
Level ζ	# trees	$n = 100$	$n = 200$	$n = 500$	$n = 1,000$	$n = 100$	$n = 200$	$n = 500$	$n = 1,000$
0.4	D^0	37.6 ± 0.1	38.3 ± 0.0	38.6 ± 0.0	38.7 ± 0.0	197.9 ± 0.1	201.5 ± 0.0	203.1 ± 0.0	203.4 ± 0.0
	DC	232.4 ± 1.1	449.1 ± 0.9	942.2 ± 0.7	1564.5 ± 0.3	1191.7 ± 2.4	2311.8 ± 1.1	4874.3 ± 1.3	8108.6 ± 0.7
	TRF	20.5 ± 0.1	19.9 ± 0.0	20.1 ± 0.0	20.3 ± 0.0	108.3 ± 0.1	104.7 ± 0.0	105.6 ± 0.0	106.5 ± 0.0
	EMC	$\mathbf{5.8 \pm 0.0}$	$\mathbf{3.5 \pm 0.0}$	$\mathbf{2.0 \pm 0.0}$	$\mathbf{1.5 \pm 0.0}$	$\mathbf{30.6 \pm 0.1}$	$\mathbf{17.0 \pm 0.0}$	$\mathbf{8.5 \pm 0.0}$	$\mathbf{5.6 \pm 0.0}$
0.6	D^0	81.1 ± 0.2	80.9 ± 0.0	80.9 ± 0.0	80.9 ± 0.0	426.7 ± 0.1	425.5 ± 0.1	425.6 ± 0.0	425.6 ± 0.0
	DC	604.0 ± 1.7	1044.0 ± 1.3	1990.5 ± 0.8	3143.4 ± 0.8	3141.7 ± 1.7	5438.3 ± 2.3	10381.5 ± 0.5	16405.3 ± 1.4
	TRF	34.2 ± 0.1	32.8 ± 0.0	32.6 ± 0.0	32.5 ± 0.0	179.1 ± 0.1	171.6 ± 0.0	170.7 ± 0.0	169.9 ± 0.0
	EMC	$\mathbf{7.9 \pm 0.1}$	$\mathbf{4.9 \pm 0.0}$	$\mathbf{3.0 \pm 0.0}$	$\mathbf{2.3 \pm 0.0}$	$\mathbf{39.6 \pm 0.0}$	$\mathbf{23.3 \pm 0.0}$	$\mathbf{12.6 \pm 0.0}$	$\mathbf{8.7 \pm 0.0}$
0.8	D^0	127.2 ± 0.4	129.0 ± 0.0	129.1 ± 0.0	129.8 ± 0.0	669.4 ± 0.2	678.3 ± 0.2	678.9 ± 0.0	682.4 ± 0.0
	DC	1071.7 ± 5.2	1824.3 ± 1.3	3358.7 ± 0.6	5199.3 ± 1.0	5594.4 ± 2.3	9526.4 ± 2.3	17567.6 ± 1.6	27200.5 ± 2.3
	TRF	49.2 ± 0.2	49.2 ± 0.0	47.2 ± 0.0	47.3 ± 0.0	257.1 ± 0.1	257.3 ± 0.1	246.8 ± 0.0	247.0 ± 0.0
	EMC	$\mathbf{11.7 \pm 0.1}$	$\mathbf{7.9 \pm 0.0}$	$\mathbf{4.8 \pm 0.0}$	$\mathbf{3.9 \pm 0.0}$	$\mathbf{58.9 \pm 0.1}$	$\mathbf{38.4 \pm 0.0}$	$\mathbf{21.5 \pm 0.0}$	$\mathbf{16.5 \pm 0.0}$

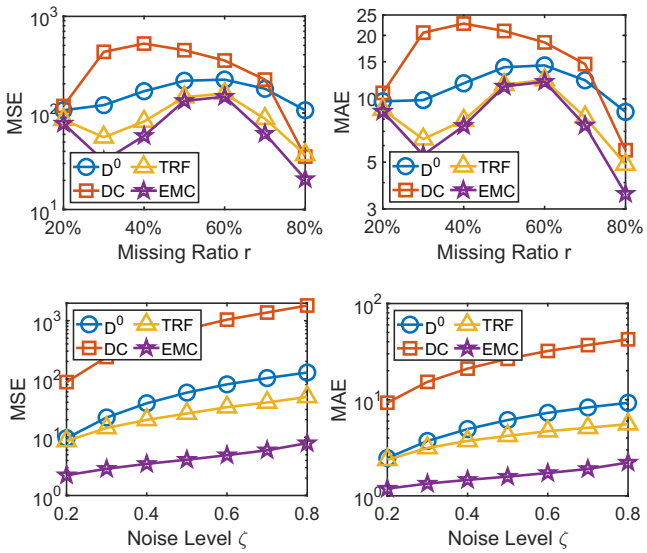

Figure 2. Sensitivity analysis on the synthetic tree dataset under MSE and MAE measurements for $l = 10$ leaves and $n = 200$ trees with different missing ratios r or noise levels ζ.

4.4 *Scalability Analysis*

To evaluate the scalability, we increase the size of tree sets to test the performance of correction approaches, and also record the running time of correcting D^0 to $\hat{D}$. The results are shown in Fig. 3. When the tree set size rises from 500 to 2,000, our EMC method has relatively stable performance and reports the best results on both scenarios (i.e., incomplete trees and noisy distance) with a relatively small running time, demonstrating its good scalability and robustness with high potential for large-scale applications.

Scalability. The TRF approach provides guaranteed performance but encounters a computational bottleneck due to the heavy processing of all $O(n^3)$ triangle inequalities. This limitation makes the TRF algorithm inapplicable when n exceeds 2,000, while the DC and EMC methods can handle larger problems with up to 5,000 trees.

Accuracy. The DC method performs poorly, especially with larger tree sets. Due to its lack of a theoretical guarantee, the DC correction method tends to result in more severe information loss and poorer performance as problem sizes increase. As expected, both the TRF and EMC methods produce smaller MSEs than D^0, validating their performance guarantees.

Efficiency. The DC method is the fastest, but its error is noticeably the largest, without any guarantee of accuracy. Among the two methods that provide performance guarantees, our EMC method can handle larger datasets and offers much faster running speed and significantly lower memory consumption, as shown in Table 3.

Comparatively, the proposed EMC approach has the highest potential to be applied in large-scale scenarios, along with the scalable techniques described in Section 3.3.

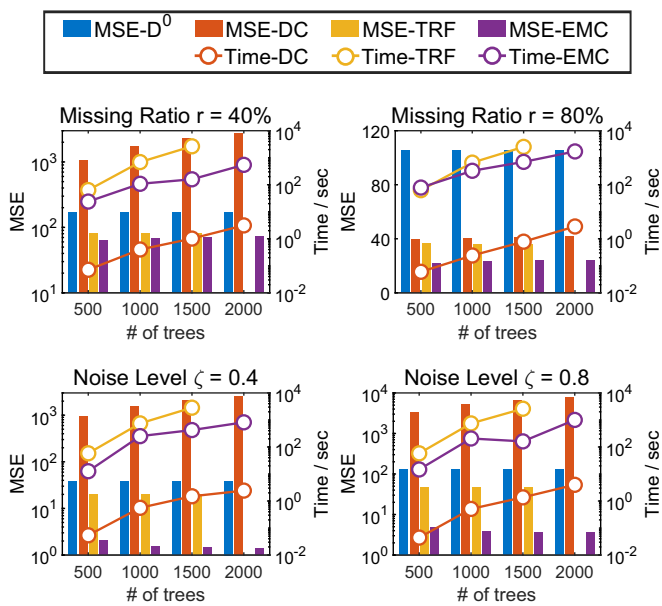

Figure 3. Scalability analysis on the synthetic tree dataset with $l = 10$ leaves and different sizes n of the tree set.

Table 3. Average memory consumption (GB) on the synthetic tree dataset.

# trees	500	1,000	1,500	2,000	5,000
DC	1.1	1.2	1.3	1.4	3.0
TRF	1.6	4.2	25.3	59.7	931.5
EMC	1.1	1.2	1.2	1.3	2.0

4.5 Tree Clustering on Incomplete Trees

We extend our evaluation to the task of clustering incomplete phylogenetic trees generated as described in Section 4.1. We take the K-means clustering results from complete trees as the ground truth, and choose the number of clusters as 50, an empirically determined value. Using the Normalized Mutual Information (**NMI**) metric for quality evaluation, as shown in Table 4 and Fig. 4, our EMC method consistently outperforms the other methods with NMI scores between 0.703 and 0.779 for configurations with 10 leaves and 100 trees. The EMC method provides more accurate clustering results and deeper insights into the underlying data structure, contributing to a better understanding of evolutionary relationships.

Table 4. NMI of K-means clustering on incomplete phylogenetic trees.

Setting	$l = 10, r = 40\%$			$l = 10, r = 80\%$		
# trees	100	200	500	100	200	500
D^0	0.154	0.091	0.047	0.250	0.157	0.062
DC	0.753	0.568	0.353	0.772	0.564	0.334
TRF	0.653	0.418	0.197	0.736	0.561	0.364
EMC	**0.755**	**0.571**	**0.362**	**0.779**	**0.589**	**0.381**

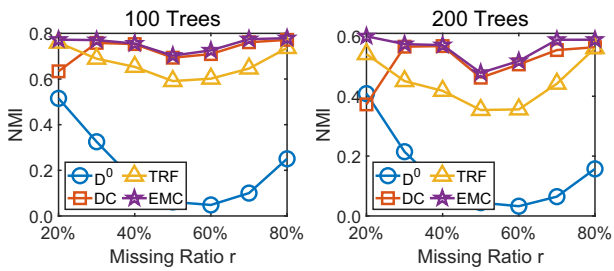

Figure 4. K-means clustering results on incomplete phylogenetic trees for $l = 10$ leaves and $n = 100$ or 200 trees under different missing ratios r.

4.6 Tree Clustering on Noisy Tree Distance

We also conduct K-means clustering on noisy tree distance matrices. Table 5 and Fig. 5 demonstrate the superiority and stability of the EMC method across different noise levels, indicating its potential to improve the reliability of clustering analyses in biological research.

Table 5. NMI of K-means clustering on noisy tree distance.

Setting	$l = 10, \zeta = 0.4$			$l = 10, \zeta = 0.8$		
# trees	100	200	500	100	200	500
D^0	0.767	0.611	0.410	0.530	0.395	0.260
DC	0.694	0.432	0.172	0.714	0.329	0.182
TRF	0.775	0.588	0.337	0.748	0.493	0.277
EMC	**0.795**	**0.616**	**0.424**	**0.787**	**0.593**	**0.405**

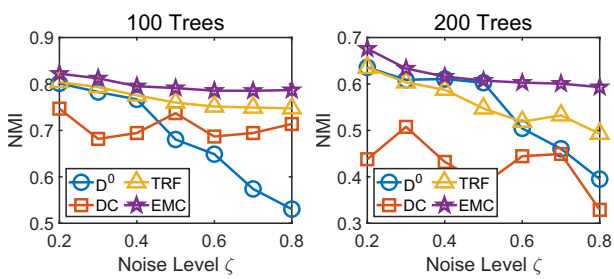

Figure 5. K-means clustering results on noisy tree distance for $l = 10$ leaves and $n = 100$ or 200 trees under different noise levels ζ.

4.7 Summary and Explanation

Statistical Explanation. Consider the noise reduction experiment. As the noise level increases, the initial non-metric matrix D^0 naturally increases the mean squared error (MSE) to the ground-truth and violates more triangle inequality constraints, providing more room for the matrix correction method to improve. Denoting the error reduction ratio (ERR) as $\frac{\text{MSE-}D^0 - \text{MSE-EMC}}{\text{MSE-}D^0}$, we observe a positive correlation among the ERR performance of our correction method, the violated constraint ratio (VCR), and the noise level.

Table 6. Correlation analysis among noise level, violated constraint ratio (VCR), mean squared error (MSE), and error reduction ratio (ERR). Take the setting of $l = 10$ leaves and $n = 200$ trees for example.

Noise Level	0.2	0.3	0.4	0.5	0.6	0.7	0.8
VCR / %	0.5	3.1	7.6	12.1	16.1	18.9	20.8
MSE-D^0	9.6	21.8	38.3	58.3	80.9	103.7	129.0
MSE-EMC	2.2	2.9	3.5	4.1	4.9	6.0	7.9
ERR / %	77.1↓	86.7↓	90.9↓	93.0↓	93.9↓	94.2↓	93.9↓

Algorithmic Advantages. Existing correction approaches for tree distance matrices have practical limitations that prevent them from achieving high-quality corrections. Our work addresses this issue by introducing a novel correction method based on the embedding technique. This method offers a theoretical guarantee of the quality of the corrected distance matrix and can handle large-scale datasets with fast running speed, small memory consumption, and scalable techniques. By overcoming these practical limitations, our method provides a reliable and efficient solution for distance matrix correction.

Application Scenarios. Our correction method has demonstrated superior results on both incomplete trees and noisy tree distance scenarios, and it has the potential to benefit various applications, such as tree clustering, classification, and retrieval. By obtaining more accurate distance metrics through correction, downstream distance-based applications could see improved performance, such as K-means clustering tasks. Overall, our study provides compelling evidence for the superiority of the EMC method and its potential to enhance biological research outcomes.

5 Conclusion

The calculation of the Robinson-Foulds distance between phylogenetic trees is a fundamental problem in both biology and computing communities, with numerous practical applications. However, obtaining an accurate distance metric can be challenging in practice, particularly when dealing with incomplete trees or noisy distances.

To address this issue, our work utilizes the Euclidean embedding technique to propose a new method for obtaining a distance metric. Our approach is based on a mild assumption and differs from classical correction methods that rely on the metric nearness model or Euclidean distance matrix. By ensuring the embeddable property, we have developed a simple yet effective method with several algorithmic advantages, including a theoretical guarantee of distance quality, fast running speed, small memory consumption, and good scalability. Our approach has demonstrated superior results with the smallest error in incomplete trees and noisy distance scenarios over classical methods. Benefiting from this, it can potentially enhance the performance of downstream applications such as tree clustering, classification, and retrieval. We hope our findings can inspire researchers to advance biological applications, and future work is ongoing.

Acknowledgements

We appreciate the anonymous reviewers for their helpful feedback that greatly improved this paper. The work of Fangchen Yu was supported in part by Shenzhen Research Institute of Big Data Scholarship Program. The work of Jianfeng Mao was supported in part by National Natural Science Foundation of China under grant U1733102, in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under grant B10120210117, and in part by CUHK-Shenzhen under grant PF.01.000404. The work of Wenye Li was supported in part by Guangdong Basic and Applied Basic Research Foundation (2021A1515011825) and Shenzhen Science and Technology Program (CUHKSZWDZC0004).

References

[1] Cécile Ané, Bret Larget, David A Baum, Stacey D Smith, and Antonis Rokas, 'Bayesian estimation of concordance among gene trees', *Molecular Biology and Evolution*, **24**(2), 412–426, (2007).

[2] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro, 'A theory of learning with similarity functions', *Machine Learning*, **72**, 89–112, (2008).

[3] Mukul S Bansal, 'Linear-time algorithms for phylogenetic tree completion under robinson-foulds distance', *Algorithms for Molecular Biology*, **15**(1), 1–15, (2020).

[4] Michael W Berry, Dani Mezher, Bernard Philippe, and Ahmed Sameh, 'Parallel algorithms for the singular value decomposition', in *Handbook of Parallel Computing and Statistics*, 133–180, Chapman and Hall/CRC, (2005).

[5] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] James P Boyle and Richard L Dykstra, 'A method for finding projections onto the intersection of convex sets in hilbert spaces', in *Advances in Order Restricted Statistical Inference*, 28–47, Springer, (1986).

[7] Justin Brickell, Inderjit S Dhillon, Suvrit Sra, and Joel A Tropp, 'The metric nearness problem', *SIAM Journal on Matrix Analysis and Applications*, **30**(1), 375–396, (2008).

[8] David Bryant, John Tsang, Paul E Kearney, and Ming Li, 'Computing the quartet distance between evolutionary trees', in *Symposium on Discrete Algorithms: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 9, pp. 285–286, (2000).

[9] Ward Cheney and Allen A Goldstein, 'Proximity maps for convex sets', *Proceedings of the American Mathematical Society*, **10**(3), 448–450, (1959).

[10] James A Cotton and Mark Wilkinson, 'Majority-rule supertrees', *Systematic Biology*, **56**(3), 445–452, (2007).

[11] Frank Deutsch and F Deutsch, *Best Approximation in Inner Product Spaces*, volume 7, Springer, 2001.

[12] Richard L Dykstra, 'An algorithm for restricted least squares regression', *Journal of the American Statistical Association*, **78**(384), 837–842, (1983).

[13] David Fernández-Baca and Lei Liu, 'A new display graph based supertree construction approach', *Supertree Construction with Display Graphs and Dynamic Graph Connectivity*, 103, (2021).

[14] Norbert Gaffke and Rudolf Mathar, 'A cyclic projection algorithm via duality', *Metrika*, **36**(1), 29–54, (1989).

[15] Andrej Gisbrecht and Frank-Michael Schleif, 'Metric and non-metric proximity transformations at linear costs', *Neurocomputing*, **167**, 643–657, (2015).

[16] W Glunt, Tom L Hayden, S Hong, and J Wells, 'An alternating projection algorithm for computing the nearest euclidean distance matrix', *SIAM Journal on Matrix Analysis and Applications*, **11**(4), 589–600, (1990).

[17] John Clifford Gower, 'Euclidean distance geometry', *Mathematical Scientist*, **7**(1), 1–14, (1982).

[18] Bhaskar Das Gupta, Xin He, Tao Jiang, Ming Li, and John Tromp, 'On computing the nearest neighbor interchange distance', in *Discrete Mathematical Problems with Medical Applications: DIMACS Workshop Discrete Mathematical Problems with Medical Applications, December 8-10, 1999, DIMACS Center*, volume 55, p. 125. American Mathematical Society, (2000).

[19] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp, 'Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions', *SIAM Review*, **53**(2), 217–288, (2011).

[20] Nathan Krislock and Henry Wolkowicz, 'Euclidean distance matrices and applications', in *Handbook on Semidefinite, Conic and Polynomial Optimization*, 879–914, Springer, (2012).

[21] Tingting Li, Dongxia Liu, Yadi Yang, Jiali Guo, Yujie Feng, Xinmo Zhang, Shilong Cheng, and Jie Feng, 'Phylogenetic supertree reveals detailed evolution of sars-cov-2', *Scientific Reports*, **10**(1), 1–9, (2020).

[22] Wenye Li, 'Estimating jaccard index with missing observations: a matrix calibration approach', *Advances in Neural Information Processing Systems*, **28**, (2015).

[23] Wenye Li, 'Scalable calibration of affinity matrices from incomplete observations', in *Asian Conference on Machine Learning*, pp. 753–768, Bangkok, Thailand, (2020). PMLR.

[24] Wenye Li and Fangchen Yu, 'Calibrating distance metrics under uncertainty', in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 219–234. Springer, (2022).

[25] Wenye Li, Fangchen Yu, and Zichen Ma, 'Metric nearness made practical', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8648–8656, (2023).

[26] Michelle M McMahon, Akshay Deepak, David Fernández-Baca, Darren Boss, and Michael J Sanderson, 'Stbase: one million species trees for comparative biology', *PloS One*, **10**(2), e0117987, (2015).

[27] Nicholas D Pattengale, Eric J Gottlieb, and Bernard ME Moret, 'Efficiently computing the robinson-foulds metric', *Journal of Computational Biology*, **14**(6), 724–735, (2007).

[28] Elzbieta Pekalska and Robert PW Duin, 'The dissimilarity representation for pattern recognition - foundations and applications', *Series in Machine Perception and Artificial Intelligence*, **64**, (2005).

[29] David F Robinson and Leslie R Foulds, 'Comparison of phylogenetic trees', *Mathematical Biosciences*, **53**(1-2), 131–147, (1981).

[30] Michael J Sanderson, Andy Purvis, and Chris Henze, 'Phylogenetic supertrees: assembling the trees of life', *Trends in Ecology & Evolution*, **13**(3), 105–109, (1998).

[31] Frank-Michael Schleif and Peter Tino, 'Indefinite proximity learning: A review', *Neural Computation*, **27**(10), 2039–2096, (2015).

[32] Isaac J Schoenberg, 'Metric spaces and positive definite functions', *Transactions of the American Mathematical Society*, **44**(3), 522–536, (1938).

[33] Yun S Song, 'Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees', *Annals of Combinatorics*, **10**(1), 147–163, (2006).

[34] Rishi Sonthalia and Anna C Gilbert, 'Project and forget: Solving large-scale metric constrained problems', *arXiv preprint arXiv:2005.03853*, (2020).

[35] Gergely J Szöllősi, Eric Tannier, Vincent Daubin, and Bastien Boussau, 'The inference of gene trees with species trees', *Systematic Biology*, **64**(1), e42–e62, (2015).

[36] Warren S Torgerson, 'Multidimensional scaling: I. theory and method', *Psychometrika*, **17**(4), 401–419, (1952).

[37] Jason TL Wang, Huiyuan Shan, Dennis Shasha, and William H Piel, 'Fast structural search in phylogenetic databases', *Evolutionary Bioinformatics*, **1**, 117693430500100009, (2005).

[38] Jason Tsong-Li Wang, Huiyuan Shan, Dennis Shasha, and William H Piel, 'Treerank: a similarity measure for nearest neighbor searching in phylogenetic databases', in *15th International Conference on Scientific and Statistical Database Management*, pp. 171–180. IEEE, (2003).

[39] James Howard Wells and Lynn R Williams, *Embeddings and extensions in analysis*, volume 84, Springer Science & Business Media, 2012.

[40] Keegan Yao and Mukul S Bansal, 'Optimal completion and comparison of incomplete phylogenetic trees under robinson-foulds distance', in *32nd Annual Symposium on Combinatorial Pattern Matching (CPM 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, (2021).

[41] Ruriko Yoshida, Kenji Fukumizu, and Chrysafis Vogiatzis, 'Multilocus phylogenetic analysis with gene tree clustering', *Annals of Operations Research*, **276**(1), 293–313, (2019).

[42] Fangchen Yu, Yicheng Zeng, Jianfeng Mao, and Wenye Li, 'Online estimation of similarity matrices with incomplete data', in *Uncertainty in Artificial Intelligence*, pp. 2454–2464. PMLR, (2023).