BIOLOGY
Methods & Protocols

METHODS ARTICLE

Novel metric for hyperbolic phylogenetic tree embeddings

Hirotaka Matsumoto [1,2,*], Takahiro Mimori[3] and Tsukasa Fukunaga[4]

[1]School of Information and Data Sciences, Nagasaki University, Nagasaki, Japan, [2]Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research, Saitama, Japan, [3]Medical Image Analysis Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan and [4]Department of Computer Science, Graduate School of Information Science and Engineering, The University of Tokyo, Tokyo, Japan

*Correspondence address. School of Information and Data Sciences, Nagasaki University, Nagasaki, Japan. E-mail: hirotaka.matsumoto@nagasaki-u.ac.jp

Abstract

Advances in experimental technologies, such as DNA sequencing, have opened up new avenues for the applications of phylogenetic methods to various fields beyond their traditional application in evolutionary investigations, extending to the fields of development, differentiation, cancer genomics, and immunogenomics. Thus, the importance of phylogenetic methods is increasingly being recognized, and the development of a novel phylogenetic approach can contribute to several areas of research. Recently, the use of hyperbolic geometry has attracted attention in artificial intelligence research. Hyperbolic space can better represent a hierarchical structure compared to Euclidean space, and can therefore be useful for describing and analyzing a phylogenetic tree. In this study, we developed a novel metric that considers the characteristics of a phylogenetic tree for representation in hyperbolic space. We compared the performance of the proposed hyperbolic embeddings, general hyperbolic embeddings, and Euclidean embeddings, and confirmed that our method could be used to more precisely reconstruct evolutionary distance. We also demonstrate that our approach is useful for predicting the nearest-neighbor node in a partial phylogenetic tree with missing nodes. Furthermore, we proposed a novel approach based on our metric to integrate multiple trees for analyzing tree nodes or imputing missing distances. This study highlights the utility of adopting a geometric approach for further advancing the applications of phylogenetic methods.

Keywords: phylogenetic tree; phylogenetic method; hyperbolic geometry; Poincaré embeddings

Introduction

The advancement of DNA sequencing technologies has enabled the determination of the genome sequences of different organisms, resulting in the reconstruction of various types of gene and species trees [1–3]. Phylogenetic analyses based on these trees have contributed substantially to research fields, such as the identification of gene–gene associations and gene functions [4], relationships between evolution and disease [5], evolutionary dynamics of pathogens [6–9], and bacterial taxonomy [10].

However, there is increasing recognition of the necessity for developing novel computational phylogenetic methods to further accelerate the application of phylogenetic analyses in various researches [11–13].

Furthermore, recent advances in lineage-tracing methods based on single-cell and genome-editing technologies have led to elucidation of cellular lineages [14], and phylogenetic methods have played an essential role in reconstructing and analyzing these cellular lineages. In addition, phylogenetic methods have contributed to research in cancer genomics (evolution of

cancer) [15] and immunogenomics (evolution of antibody lineages) [16, 17]. In addition to the essential role of phylogenetic methods in evolutionary biology, these approaches are increasingly becoming relevant for other research areas; thus, the development of novel phylogenetic methods can contribute broadly to various fields in the life and biomedical sciences.

Hyperbolic geometry, as a non-Euclidean form of geometry, has attracted attention in artificial intelligence research, with several methods using hyperbolic geometry recently developed for various applications. Representation learning models have shown that the hyperbolic space could be used to represent latent hierarchical structures, exhibiting significant performance improvements over other approaches [18, 19]. This finding motivated several subsequent machine-learning studies using hyperbolic space, including the construction of hyperbolic neural networks [20]. In addition, a novel approach for hierarchical clustering that optimizes the coordinates in hyperbolic space was proposed [21]. Recently, a novel hyperbolic geometry-based approach for learning hierarchical structure, including phylogenetic data, was proposed, and it showed the potential of hyperbolic geometry for novel phylogenetic methods [22].

The coordinates in 2D hyperbolic geometry can be described with the Poincaré disk (or the Poincaré ball model for 3D or *n*-dimensional hyperbolic space). In contrast to the Euclidean space, where the shortest path between two points (i.e. a geodesic) is a straight line, a geodesic in the Poincaré disk is an arc (Fig. 1A). This characteristic of geodesics on the Poincaré disk effectively represents a hierarchical structure because the geodesic provides a better match to a tree structure than the path on Euclidean space represented as a straight line (Fig. 1B). In addition, the area of the Poincaré disk grows exponentially in accordance with the distance from the origin, which is an advantageous characteristic for representing the nodes of a tree which increase in number exponentially with branching of the tree (Fig. 1C). These characteristics of a Poincaré disk are effective for embedding a phylogenetic tree, and visualization tools based on such hyperbolic phylogenetic tree embeddings have been developed [23, 24]. Hyperbolic embeddings have also recently been used to reconstruct cell lineage trees from single-cell RNA-sequencing data [25, 26]. In addition, the hyperbolic space has been used for several other types of biological studies, such as for analyzing protein function based on protein interaction network embedding [27] and interpreting the mechanisms of olfactory space [28]. As the hyperbolic embeddings are effective for analyzing biological data with a hierarchical structure, they are expected to be useful for various types of phylogenetic methods. To extend the applications of hyperbolic space for phylogenetic methods, we here propose a novel metric for accurate phylogenetic tree embedding in hyperbolic space. The proposed method is expected to contribute to the development of novel phylogenetic methods using hyperbolic space.

For both Euclidean and hyperbolic embeddings, the input data are generally represented as a distance matrix among data points, and objective functions are designed to preserve the input distances with the geodesic distances on the embedded space. In the case of phylogenetic tree embeddings, the distance between two nodes on the tree corresponds to the evolutionary distance, which is the sum of branch lengths between the two nodes (additivity). For example, considering the two external nodes i and j and their most recent common ancestor (MRCA) node k, the evolutionary distance d_{ij} represents the sum of the respective distance of each node to the MRCA, $d_{ik} + d_{jk}$. Although hyperbolic phylogenetic tree embedding is expected to preserve distance information better than the Euclidean embedding, it still cannot perfectly preserve the evolutionary distances of the phylogenetic tree. This is because the additivity can only accurately reflect the geodesic distance when all nodes are located on a single geodesic line, and this is not the case for tree structures constructed with general embeddings. To overcome this limitation, we developed a novel metric that allows for embedding the evolutionary distance in a more precise manner by using the cosine rule in hyperbolic geometry (i.e. the hyperbolic law of cosines). We compared the performance of conventional Euclidean and hyperbolic embeddings with that of our proposed embeddings, which confirmed that our approach could precisely reconstruct the evolutionary distance in a variety of scenarios. The proposed embeddings also exhibit an advantageous property in that the angles among external nodes and their MRCA nodes are stable and can be used to predict MRCA. In addition, we investigated the ability of the three types of embeddings to predict the nearest node of an external node not included in the partial phylogenetic tree, demonstrating that our proposed approach had the best predictive performance. The proposed metric also offers meaningful way to integrate multiple trees, which is not achieved with naive distance averaging approach. This approach also presents a novel algorithm for integrating partial trees and imputing missing distances.

Recent advancements of the algorithms using hyperbolic geometry have shown the efficacy of the geometry for analyzing and reconstructing various tree structures [22, 25, 26]. In this work, we demonstrated the effectiveness of our metric that is designed for the phylogenetic tree. The previous researches and our results demonstrated the potential of adopting a geometric approach for phylogenetic analyses. The importance of novel phylogenetic methods has been increasingly recognized to handle the new and abundant data emerging from the current genomic era. In this regard, our approach has potential to contribute to several types of phylogenetic analyses by proposing a novel concept of "tree thinking" [29] based on geometric thinking.
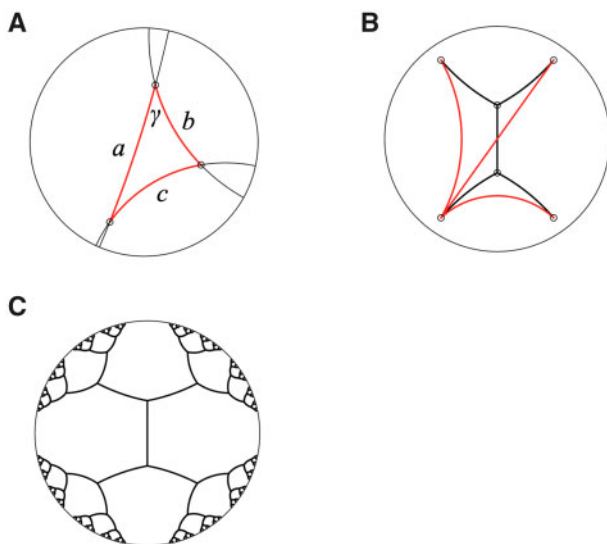

Figure 1: graphical schematic of the Poincaré disk. (**A**) Geodesics and triangle on the Poincaré disk. (**B**) Geodesics for the external nodes of a simple tree structure. (**C**) Tree embedding on the Poincaré disk. The geodesic distances based on the Poincaré disk for all branches are equal.

Materials and methods

Novel metric for hyperbolic phylogenetic tree embeddings

For representing a phylogenetic tree with continuous coordinates, hyperbolic space is expected to preserve evolutionary distance better than Euclidean space [18, 19]. However, the general distance matrix embedding methods, including the general hyperbolic embedding, had the limitation for the precise representation of evolutionary distance. To overcome this limitation, we developed a novel metric that was based on the hyperbolic law of cosines to represent the evolutionary distance precisely in hyperbolic space. For a hyperbolic triangle with side lengths a, b, and c, and angle (in radians) γ (Fig. 1A), the hyperbolic law of cosines is satisfied, as follows:

$$\cosh(c) = \cosh(a)\cosh(b) - \sinh(a)\sinh(b)\cos(\gamma). \quad (1)$$

In the case of $\gamma = \pi/2$, the above equation becomes

$$\cosh(c) = \cosh(a)\cosh(b). \quad (2)$$

For the two nodes i and j in a phylogenetic tree and MRCA node k, we describe the evolutionary distances between any two nodes (i, j), (i, k), and (j, k) as d_{ij}, d_{ik}, and d_{jk}, respectively. We also describe the geodesic distances in the hyperbolic space as $d_{ij}^{(H)}$, $d_{ik}^{(H)}$, and $d_{jk}^{(H)}$. If we assume that the angle $\angle ikj$ is $\pi/2$ and the geodesic distance satisfies $d_{ij}^{(H)} = \mathrm{acosh}(\exp(d_{ij}))$, $d_{ik}^{(H)} = \mathrm{acosh}(\exp(d_{ik}))$, and $d_{jk}^{(H)} = \mathrm{acosh}(\exp(d_{jk}))$, we can derive the following from Equation (2):

$$\exp(d_{ij}) = \exp(d_{ik})\exp(d_{jk}) = \exp(d_{ik} + d_{jk}). \quad (3)$$

When the phylogenetic tree is given, the evolutionary distance d_{ij} over the phylogenetic tree is given by $d_{ik} + d_{jk}$ (additivity of evolutionary distance). Equation (3) shows that the additivity of evolutionary distances ($d_{ij} = d_{ik} + d_{jk}$) can be represented if the above assumptions ($\angle ikj = \pi/2$) are satisfied.

Thus, we translate the evolutionary distance between two nodes, i and j, as $X_{ij} = \mathrm{acosh}(\exp(d_{ij}))$ and embed the translated distance matrix X onto the hyperbolic space (Poincaré ball) instead of directly embedding the evolutionary distance matrix.

Phylogenetic tree embeddings

The evolutionary distance matrix is represented as D (with D_{ij} considered to be the evolutionary distance d_{ij}), and then D was embedded with both Euclidean and general hyperbolic embeddings, in addition to hyperbolic embeddings with the proposed metric (Fig. 2). The performance of each approach was then compared.

For Euclidean embeddings, we used Sammon mapping, which is a type of multidimensional scaling (MDS) [30]. Toward this end, we used the *sammon* function in the MASS package in R language, and embedded D for the Euclidean space with various dimensions (M). The performance of the embeddings was evaluated by calculating the mean squared error (MSE) between D_{ij} and the Euclidean distance $d_E(z_i, z_j)$, where z_i represents the coordinate of node i of the embedded space.

For general hyperbolic embeddings, we used the *hydraPlus* function in the hydra package in R [31], and embedded D with parameters "curvature" = 1 and "alpha" = 1 under various dimensions M. The performance was then evaluated by
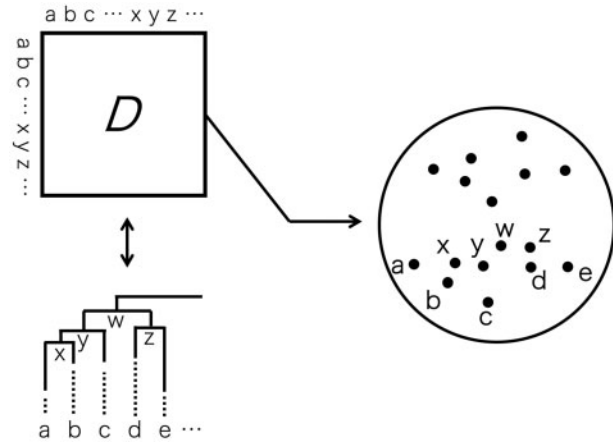

Figure 2: graphical representation of hyperbolic phylogenetic tree embeddings. The input evolutionary distance matrix D contains all nodes, including the internal nodes of the phylogenetic tree. In our proposed embeddings, the input matrix is the translated matrix X, instead of D.

calculating the MSE between D_{ij} and the geodesic distance on the Poincaré ball $d_P(z_i, z_j)$, where the distance is defined as follows:

$$d_P(z_i, z_j) = \mathrm{acosh}\left(1 + 2\frac{||z_i - z_j||^2}{(1 - ||z_i||^2)(1 - ||z_j||^2)}\right). \quad (4)$$

We also used the *hydraPlus* function for hyperbolic embeddings with the proposed metric. We embedded the translated matrix X, where $X_{ij} = \mathrm{acosh}(\exp(D_{ij}))$, and evaluated its performance by calculating the MSE between D_{ij} and the inverse transformed distance $\log(\cosh(d_P(z_i, z_j)))$.

Folding-in internal nodes

In the previous section, we considered the distance matrix D for all nodes in a phylogenetic tree as the input, including internal nodes, and embedded all nodes simultaneously. We further evaluated the performance of each type of embedding in the case of "folding-in" the internal nodes. In this case, the input was a distance matrix for external nodes, and the external nodes were first embedded, followed by optimization of the coordinates of the internal nodes for the embedding space (Fig. 3).

We first conducted these steps for complete phylogenetic tree. The external nodes were first embedded, and the coordinates of each internal node is optimized numerically and independently (Fig. 3A). The objective function used for optimizing the coordinates of an internal node was the minimization of the MSE between the true evolutionary distances and the reconstructed evolutionary distance based on the geodesic distance for the internal node and all external nodes.

We next embedded the internal nodes of a partial tree, which is the phylogenetic tree for only a portion of the external nodes (Fig. 3B). The objective function used here is the minimization of the MSE between the true evolutionary distances and the reconstructed evolutionary distance for the internal node and external nodes included in the partial tree.

Prediction of the MRCA

Based on the results of each method of embedding, we developed new indicators to predict the MRCA of two external nodes i and j. The indicators are based on the distance or angle information on
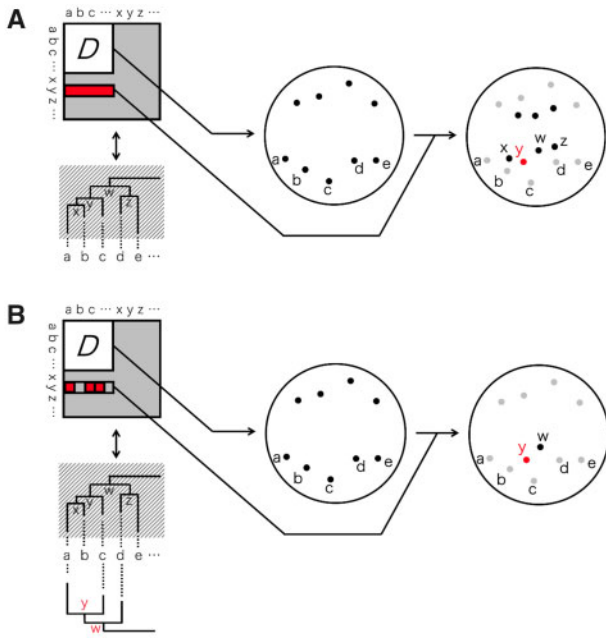
Figure 3: graphical representation of the hyperbolic embeddings of a distance matrix corresponding to the external nodes and folding-in of the internal nodes of the complete phylogenetic tree (**A**) and the partial tree (**B**).

the embedded space, which are used to determine whether the internal node k is the MRCA of i and j. When node k is the MRCA for i and j, the additivity of the evolutionary distance ($d_{ij} = d_{ik} + d_{jk}$) is satisfied. However, the above condition is not sufficient to determine the MRCA because all internal nodes l on the path from i to j satisfy the additivity $d_{ij} = d_{il} + d_{jl}$. Therefore, we also used the outgroup node o, which is a distant species from the target species and is used to determine the root of the tree. The additivity is also satisfied for the outgroup node o only for the MRCA node k ($d_{io} = d_{ik} + d_{ok}$ and $d_{jo} = d_{jk} + d_{ok}$). As the first indicator, we developed the "distance-score," which is the maximum of ($d_{ij} - d_{ik} - d_{jk}$)^2, ($d_{io} - d_{ik} - d_{ok}$)^2, and ($d_{jo} - d_{jk} - d_{ok}$)^2. We also developed another indicator based on the angle information. If the internal node k' is a more distant (i.e. not the most recent) ancestor, then $d_{ik'} + d_{jk'}$ becomes larger than the d_{ij}, and the angle on the embedded space will be $\angle ik'j < \angle ikj$, where k is the MRCA of i and j. We calculated the angles $\angle ikj$, $\angle iko$, and $\angle jko$ based on the embedded coordinates, and then used the minimum value of each angle as the "angle-score," which served as the basis for predicting whether or not k is the MRCA.

We evaluated the performance of distance-score and angle-score to predict the MRCA for each embedding type by randomly selecting the external nodes i and j 1000 times, and used i, j, and their MRCA k as the positive-control dataset. We also randomly selected the external node l, and used the combination i, l, and k as the negative-control dataset (if the MRCA of i and l was k, we resampled l randomly). We performed the above process for each random simulated tree and calculated the mean of the area under the receiver operating characteristic (ROC) curve (AUC) values to evaluate the prediction of the MRCA with the angle-score for each embedding method.

Prediction of the nearest-neighbor node in the partial tree

Numerous genomes have been determined in the genome era, and the importance of integrating multiple partial phylogenetic trees

[32, 33] or placing a novel species into an already-established phylogenetic tree (phylogenetic placement) [34] is increasing. Therefore, we evaluated the ability of our method to predict the nearest-neighbor node in the partial tree for some external nodes missing from the tree. First, we embedded the external nodes and folding-in the partial tree as described above (Fig. 3B). Second, we predicted the nearest-neighbor node in the partial tree for an external node that was not included the tree based on the geodesic distance for each embedding method. We evaluated the performance of each embedding by calculating the rank of the actual nearest-neighbor node calculated from the complete phylogenetic tree.

Integration and embeddings of multiple phylogenetic trees

As an application of our distance transformation, we here focus on the use of our geometric perspective for the alternative approach of the integration of multiple gene trees. The coordinates of the M-dimensional Poincaré ball model z_i can be transformed to the coordinates of the M-dimensional hyperbolic geometry $x_i = \{x_{i,k} | k = 1, \ldots, M+1\}$, where x_i satisfies $\sum_{k=1}^{M} x_{i,k}^2 - x_{i,M+1}^2 = -1$.

The geodesic distance based on the above coordinates [hyperbolic geometry version of Equation (4)] is then defined as follows:

$$d_H(x_i, x_j) = \text{acosh}(-\langle x_i, x_j \rangle_{M,1}), \tag{5}$$

$$\langle x_i, x_j \rangle_{M,1} = \sum_{k=1}^{M} x_{i,k} x_{j,k} - x_{i,M+1} x_{j,M+1}. \tag{6}$$

Using our novel metric for the conversion of evolutionary distance, $\text{acosh}(\exp(d_{ij}))$, the pseudo-inner product ($\langle x_i, x_j \rangle_{M,1}$) has the following relationship to evolutionary distance:

$$\langle x_i, x_j \rangle_{M,1} = -\exp(d_{ij}). \tag{7}$$

Here, we consider two genes, a and b, and their phylogenetic gene trees independently. The evolutionary distances for each tree are then $d_{ij}^{(a)}$ and $d_{ij}^{(b)}$, respectively. Based on Equation (7), the mean of the pseudo-inner product satisfies the following relationship:

$$\frac{1}{2} \left(\langle x_i^{(a)}, x_j^{(a)} \rangle_{1,M} + \langle x_i^{(b)}, x_j^{(b)} \rangle_{1,M} \right) = -\frac{1}{2} \left(\exp(d_{ij}^{(a)}) + \exp(d_{ij}^{(b)}) \right). \tag{8}$$

Then, we define a new vector, x_i', which is combined with $x_i^{(a)}$ and $x_i^{(b)}$ as follows:

$$x_i' = \frac{1}{\sqrt{2}} \left(x_{i,1}^{(a)}, x_{i,1}^{(b)}, \ldots, x_{i,M+1}^{(a)}, x_{i,M+1}^{(b)} \right). \tag{9}$$

Using the above vector, the mean of the pseudo-inner product can be represented with the following bilinear form:

$$\frac{1}{2} \left(\langle x_i^{(a)}, x_j^{(a)} \rangle_{1,M} + \langle x_i^{(b)}, x_j^{(b)} \rangle_{1,M} \right)$$
$$= \sum_{k=1}^{2M} x_{i,k}' x_{j,k}' - x_{i,2M+1}' x_{j,2M+1}' - x_{i,2M+2}' x_{j,2M+2}' \tag{10}$$
$$= \langle x_i', x_j' \rangle_{2M,2}.$$

This form is known as an inner product related to a pseudo-Euclidean space. Therefore, the pseudo-Euclidean space may be useful for representing the ensemble trees.

In this research, we compared the approaches for integrating tree distances: averaging distance matrix (H0), averaging inner product based on the general hyperbolic method (H1), and averaging inner product based on the proposed method (H2). Hereafter, we only used the external nodes of the trees and described the distance between the two external nodes i and j of two trees as $d_{ij}^{(a)}$ and $d_{ij}^{(b)}$, respectively. As the typical integration of trees, we defined the averaged distance as $d_H(x_i, x_j) = (d_{ij}^{(a)} + d_{ij}^{(b)})/2$. Because the hyperbolic distance is defined by Equation (5), the indefinite inner product is given by $-\cosh((d_{ij}^{(a)} + d_{ij}^{(b)})/2)$, and we defined the inner product matrix $H^{(0)}$ such that $H_{ij}^{(0)} = -\cosh((d_{ij}^{(a)} + d_{ij}^{(b)})/2)$. Then, we performed the eigendecomposition of the matrix. Because the inner product corresponds to the hyperbolic geometry, there will be at least one negative eigenvalue. The eigenvalue decomposition of such inner product matrix is connected with the hyperbolic embeddings and is used for initializing the hyperbolic embeddings in the hydra package [31].

We also developed the integration of trees based on Equation (8) and defined the inner product as $-(\exp(d_{ij}^{(a)}) + \exp(d_{ij}^{(b)}))/2$. Then, we defined the inner product matrix $H^{(2)}$ such that $H_{ij}^{(2)} = -(\exp(d_{ij}^{(a)}) + \exp(d_{ij}^{(b)}))/2$ and performed the eigendecomposition of the matrix. Based on Equation (10), there will be two negative eigenvalues (see online supplementary material for the discussion of the number of negative eigenvalues). We also defined the inner product matrix $H^{(1)}$ based on the general hyperbolic embeddings such that $H_{ij}^{(1)} = -(\cosh(d_{ij}^{(a)}) + \cosh(d_{ij}^{(b)}))/2$ and performed the eigendecomposition of the matrix.

Imputation of missing distances from partial trees

Next, we considered the case of integrating two partial trees, a and b, which have different sets of external nodes. We refer to the external nodes that appeared only in the tree a as $L_{a \setminus b}$, only in the tree b as $L_{b \setminus a}$, and in both of the trees as $L_{a \cap b}$, respectively. In such case, the distance d_{ij} for $i \in L_{a \setminus b}$ and $j \in L_{b \setminus a}$ is missing. Imputing such missing distances is important to estimate the comprehensive tree [35]. In this research, we proposed an alternative approach for imputing distances based on the eigenvalue of the indefinite inner product matrix $H^{(2)}$. We discussed that the first and second smallest eigenvalue of $H^{(2)}$ would be negative and positive when it is appropriated for hyperbolic representation (see online supplementary material). If we set inappropriate values to the missing distances, $H^{(2)}$ would not be represented by hyperbolic space and would have more than two negative eigenvalues. Therefore, we optimized the missing values by maximizing the second smallest eigenvalue of $H^{(2)}$. We also impute the missing values based on the inner matrix $H^{(1)}$ in the same manner. The detailed optimization procedures are described in online supplementary material.

Datasets

We randomly generated phylogenetic tree shapes comprising 100 external nodes using the *rtree* function [36]. We determined each branch length with $\alpha(1 - \log(u(e-1) + 1))$, where u is a uniform random value and α corresponds to the scaling factor, which was set to values of 0.25, 0.5, and 1 for this analysis. We added another external node as an outgroup for the phylogenetic tree, resulting in a tree with 101 external nodes. We

generated the trees 100 times independently for each α value. For analyses of the partial trees, we randomly selected 20 external nodes in addition to the outgroup node and extracted the partial tree corresponding to these external nodes set from the complete phylogenetic trees. Hereafter, we refer to the dataset with $\alpha = 0.25$, 0.5, and 1.0 as $\text{Data}_{0.25}$, $\text{Data}_{0.5}$, and Data_1, respectively.

We also used the phylogenetic tree of primates downloaded from TimeTree [37] for evaluation. We used the genus-level phylogenetic tree and added one outgroup node, resulting in a tree with 77 external nodes. First, we normalized the branch lengths such that the maximum branch length was 1.0, and then perturbed the branch lengths by adding $0.5(1 - \log(u(e-1) + 1))$. We generated the branch length-perturbed trees 100 times independently. The partial trees were generated in the same manner as implemented for the random phylogenetic trees. We refer to the dataset as Data_{pri}.

We also used the phylogenetic dataset of the carnivorous Caryophyllales downloaded from http://dx.doi.org/10.5061/dryad.vn730 [38]. This dataset contains 13 taxa, and we used the 1237 gene trees based on the CDS sequences that include all taxa. We normalized the branch lengths of each gene tree such that the maximum branch length was 1.0 and added one outgroup node. For analyses of the partial trees, we randomly selected five external nodes in addition to the outgroup node. We refer to the dataset as Data_{ca}.

For the analysis of integrating trees, we randomly generated tree shape comprising 50 external nodes using the *rtree* function and determined each branch length with $\alpha = 0.5$ and then normalized the branch length so that the maximum length is 1.0. We duplicated the tree and referred to trees 1 and 2, hereafter. Then, we randomly selected four external branches and set the two branch lengths of trees 1 and 2 to 0.5 and 0.01, respectively. We also set the other two branch lengths of trees 1 and 2 to 0.01 and 0.5, respectively. We regarded the four external nodes corresponding to the four external branches as the change nodes between two trees. We generated the two tree pair dataset 1000 times independently.

We used a tree of $\text{Data}_{0.5}$ for the analysis of integrating and imputing missing distances from partial trees. We randomly generated two partial trees from the tree so that both of the partial trees include $|L_{a \cap b}|$ common external nodes. In this research, we set $|L_{a \cap b}|$ to 80, 60, and 40 and imputed the missing distances for each $|L_{a \cap b}|$ condition.

Results

Full node embeddings

We embedded the evolutionary distance matrix D, including the internal nodes, with each embedding method, and evaluated the performance of each method according to the MSE calculated between the actual evolutionary distance and the reconstructed evolutionary distance based on the coordinate of the embedded space. The mean MSE values for the dataset $\text{Data}_{0.5}$ and Data_{ca} of MDS, general hyperbolic embeddings (H1), and the proposed hyperbolic embeddings (H2) for various embedding dimensions M are shown in Table 1. The results for the dataset $\text{Data}_{0.25}$, Data_1, and Data_{pri} are provided in online supplementary material.

The mean MSE values of MDS were larger than those of H1 and H2 for all cases, and the hyperbolic space offered a better representation of the phylogenetic trees. Although the mean MSE value of H1 for $\text{Data}_{0.5}$ was the smallest for $M = 5$, the

Table 1: Mean MSE values of full node embeddings for the datasets Data$_{0.5}$ (above panel) and Data$_{ca}$ (below panel) for each embedding method (columns): Euclidean embeddings (MDS), general hyperbolic embeddings (H1), and the proposed hyperbolic embeddings (H2). The rows represent the dimensions of the embeddings.

	$M=5$	$M=10$	$M=20$	$M=30$
MDS	1.4×10^{-1}	4.4×10^{-2}	2.3×10^{-2}	2.0×10^{-2}
H1	1.4×10^{-2}	4.3×10^{-3}	3.2×10^{-3}	3.2×10^{-3}
H2	2.3×10^{-2}	4.0×10^{-3}	5.6×10^{-4}	1.6×10^{-4}
	$M=4$	$M=6$	$M=8$	$M=10$
MDS	2.9×10^{-2}	1.7×10^{-2}	1.5×10^{-2}	1.5×10^{-2}
H1	1.3×10^{-2}	7.2×10^{-3}	6.4×10^{-3}	6.3×10^{-3}
H2	1.9×10^{-2}	5.6×10^{-3}	2.0×10^{-3}	8.5×10^{-4}

Table 2: Mean AUC values for the ability to identify the MRCA based on the distance-score for the datasets Data$_{0.5}$ (above panel) and Data$_{ca}$ (below panel) for each embedding method (columns). The rows represent the dimensions of the embeddings.

	$M=5$	$M=10$	$M=20$	$M=30$
MDS	0.75	0.82	0.84	0.84
H1	0.85	0.91	0.93	0.93
H2	**0.91**	**0.95**	**0.97**	**0.98**
	$M=4$	$M=6$	$M=8$	$M=10$
MDS	0.77	0.83	0.86	0.86
H1	0.88	0.92	**0.95**	0.95
H2	**0.94**	**0.94**	**0.95**	**0.97**

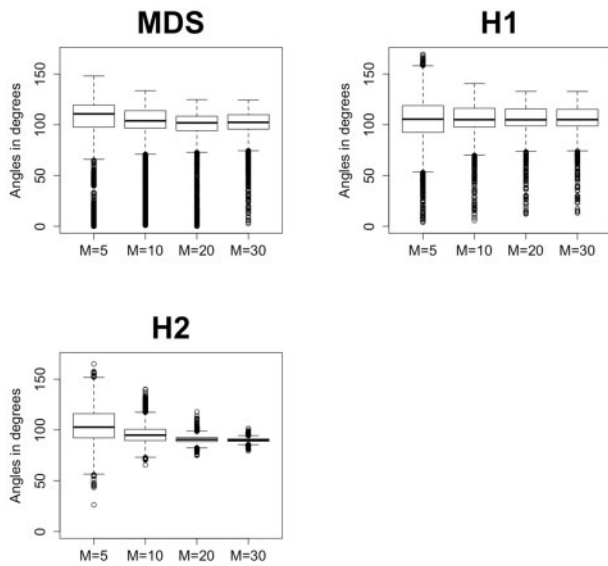The bold values represent the best results.

Figure 4: angles (degrees) of two external nodes and their MRCA for the dataset Data$_{0.5}$ for each embedding method and each dimension.

Table 3: Mean AUC values for the ability to identify the MRCA based on the angle-score for the datasets Data$_{0.5}$ (above panel) and Data$_{ca}$ (below panel) for each embedding method (columns). The rows represent the dimensions of the embeddings.

	$M=5$	$M=10$	$M=20$	$M=30$
MDS	0.62	0.71	0.78	0.79
H1	0.80	0.87	0.90	0.90
H2	**0.91**	**0.95**	**0.97**	**0.99**
	$M=4$	$M=6$	$M=8$	$M=10$
MDS	0.67	0.74	0.78	0.79
H1	0.82	0.88	0.92	0.92
H2	**0.94**	**0.94**	**0.95**	**0.98**

The bold values represent the best results.

Prediction of the MRCA

Based on the embeddings described in the subsection above, we next evaluated the ability of each method to predict the MRCA based on the distance-score and angle-score, respectively. We calculated the distance-score and angle-score for the positive and negative datasets (see Materials and methods section), and evaluated the performance according to the AUC values of ROC curves for each embedding method. The mean AUC values of 100 simulated phylogenetic trees with $\alpha = 0.5$ (Data$_{0.5}$) and the carnivorous Caryophyllales gene trees (Data$_{ca}$) with distance-score and angle-score are shown in Tables 2 and 3, respectively.

The AUC values of the proposed method (H2) were the highest compared with those of other embedding methods for all dimensions M. In particular, our method could predict the MRCA almost perfectly with $M=30$ and $M=10$ for Data$_{0.5}$ and Data$_{ca}$, respectively. Interestingly, the AUC value of H2 were significantly higher than that of H1, despite the fact that the MSE for standard hyperbolic embeddings were better than that of our embedding method with $M=30$ and $M=10$ for Data$_{0.5}$ and Data$_{ca}$, respectively. These results imply that the proposed embedding method maintains the integrity not only of distances but also of the angles even when the dimensions of the embeddings are small. This tendency was consistent for other datasets (see online supplementary material).

values of H2 were the smallest for $M \geq 10$. In particular, the mean MSE value of H2 for Data$_{0.5}$ was less than one-tenth the value of H1 for $M = 30$. The similar tendency that the mean MSE value of H1 was smallest for $M = 4$ and the values of H2 were the smallest for $M \geq 6$ were validated for Data$_{ca}$.

We also investigated the angles (degrees) of any external nodes i and j and their MRCA k ($\angle ikj$) for each embedding method under various M values (Fig. 4). The dispersion of angles of MDS and H1 were large, even when the dimension M was increased. In contrast, all of the angles of H2 merged to $\sim 90°$ in accordance with increasing M. The median angle of H2 for $M = 5$ was over $100°$, which had a gap to satisfy Equation (2), explaining the larger MSE value of H2 than that of H1 for $M = 5$.

Thus, the proposed hyperbolic embeddings can precisely represent the phylogenetic tree, especially with large embedding dimensions. In addition, our proposed embeddings offer an advantage with respect to the angle, which will be useful for analyses or machine learning using angle information.

Table 4: Mean MSE values of external node embeddings and folding-in internal nodes for the datasets $Data_{0.5}$ (above panel) and $Data_{ca}$ (below panel) for each embedding method (columns). The rows represent the dimensions of the embeddings.

	$M = 5$	$M = 10$	$M = 20$	$M = 30$
MDS	1.0×10^{-1}	4.2×10^{-2}	2.2×10^{-2}	2.0×10^{-2}
H1	1.4×10^{-2}	4.6×10^{-3}	3.6×10^{-3}	3.5×10^{-3}
H2	2.3×10^{-2}	4.0×10^{-3}	6.2×10^{-4}	2.3×10^{-4}
	$M = 4$	$M = 6$	$M = 8$	$M = 10$
MDS	3.9×10^{-2}	2.4×10^{-2}	2.1×10^{-2}	2.0×10^{-2}
H1	1.6×10^{-2}	9.2×10^{-3}	8.4×10^{-3}	8.3×10^{-3}
H2	2.0×10^{-2}	6.3×10^{-3}	2.8×10^{-3}	1.9×10^{-3}

Table 5: Mean MSE values of external node embeddings and folding-in internal nodes of the partial tree for the datasets $Data_{0.5}$ (above panel) and $Data_{ca}$ (below panel) for each embedding method (columns). The rows represent the dimensions of the embeddings.

	$M = 5$	$M = 10$	$M = 20$	$M = 30$
MDS	1.1×10^{-1}	8.7×10^{-2}	7.8×10^{-2}	8.0×10^{-2}
H1	2.0×10^{-2}	1.6×10^{-2}	1.6×10^{-2}	1.6×10^{-2}
H2	2.3×10^{-2}	8.3×10^{-3}	3.3×10^{-3}	3.0×10^{-3}
	$M = 4$	$M = 6$	$M = 8$	$M = 10$
MDS	1.4×10^{-1}	1.3×10^{-1}	1.3×10^{-1}	1.2×10^{-1}
H1	3.7×10^{-2}	3.7×10^{-2}	3.8×10^{-2}	3.9×10^{-2}
H2	5.2×10^{-2}	3.5×10^{-2}	2.6×10^{-2}	2.4×10^{-2}

External node embeddings and folding-in internal nodes

In the previous analysis, we embedded all nodes, including the internal nodes, of a phylogenetic tree simultaneously. Next, we embedded the external nodes and then appended the internal nodes onto the embedding space (Fig. 3A). The mean MSE values for the dataset $Data_{0.5}$ and $Data_{ca}$ for each embedding method are shown in Table 4.

Similar to the previous results, the mean MSE values of the proposed embedding method were better than those of other embeddings with folding-in of internal nodes, especially for $M = 30$ for $Data_{0.5}$. These results indicate that the proposed embeddings can be trained to learn the appropriate space even if the only information available is the evolutionary distance of external nodes.

External node embeddings and folding-in internal nodes of the partial tree

We further evaluated the performance of each embedding method when the evolutionary distances of some nodes are only partially known. As an example, we first embedded the external nodes and then appended the internal nodes of the partial tree according to the evolutionary distances for the external nodes included in the tree (Fig. 3B). Finally, we calculated the MSE values for the evolutionary distance between the external nodes that were not included in the tree and the internal nodes. The mean MSE values for the dataset $Data_{0.5}$ and $Data_{ca}$ for each embedding method are shown in Table 5. Similar to the previous results, the mean MSE values of the proposed embeddings were better than those of other embeddings, especially for $M = 30$ for $Data_{0.5}$. Thus, the proposed method can also effectively embed nodes when there is missing data on the evolutionary distances for some nodes. This approach can be useful for analyses of phylogenetic trees with only partial external nodes available.

Prediction of nearest-neighbor nodes in the partial tree

Based on the embeddings for the partial tree, we investigated the ability to predict the nearest-neighbor node for an external node that is not included in the partial tree. The prediction performances of the different embedding methods were compared by calculating the rank of the actual nearest-neighbor node for the external nodes based on the respective geodesic distances. The actual nearest-neighbor node can be determined from the

Table 6: Mean rank of predicting the true nearest-neighbor nodes in the partial tree for external nodes not included the tree for the datasets $Data_{0.5}$ (above panel) and $Data_{ca}$ (below panel). The rank is based on the geodesic distances for each embedding method.

	$M = 5$	$M = 10$	$M = 20$	$M = 30$
MDS	3.07	2.47	2.13	2.04
H1	1.75	1.52	1.47	1.46
H2	**1.43**	**1.24**	**1.18**	**1.14**
	$M = 4$	$M = 6$	$M = 8$	$M = 10$
MDS	2.00	1.77	1.67	1.64
H1	1.40	1.27	1.22	1.21
H2	**1.34**	**1.20**	**1.08**	**1.05**

The bold values represent the best results.

complete phylogenetic tree. The mean ranks of each method for the dataset $Data_{0.5}$ and $Data_{ca}$ are shown in Table 6, and the results for other datasets are shown in online supplementary material.

The mean ranks using embeddings based on hyperbolic space (H1 and H2) were superior to those obtained using Euclidean embeddings. Moreover, the mean ranks of the proposed method demonstrated the best performance overall under all conditions. In particular, the mean rank of our method was close to 1 with $M = 30$ for $Data_{0.5}$ and with $M = 10$ for $Data_{ca}$. Therefore, our approach would be effective for analyzing a partial tree and assigning missing nodes.

Integration and embeddings of trees

By averaging distance, averaging the inner product based on the general hyperbolic embeddings, and averaging the inner product based on the proposed method, we integrated the distance matrix of two trees and performed the eigendecomposition of each inner product matrix. The distribution of the first, second, and third smallest eigenvalues for the distance averaging approach (H0), general hyperbolic inner product integrating approach (H1), and the proposed distance based integrating approach (H2) are shown in Fig. 5A and B. The distance averaging approach (H0) showed one large negative eigenvalue. In contrast, the approaches that calculate the mean of the inner product (H1 and H2) showed two large negative eigenvalues in most cases. Although the third smallest eigenvalues of H1 were near 0, these
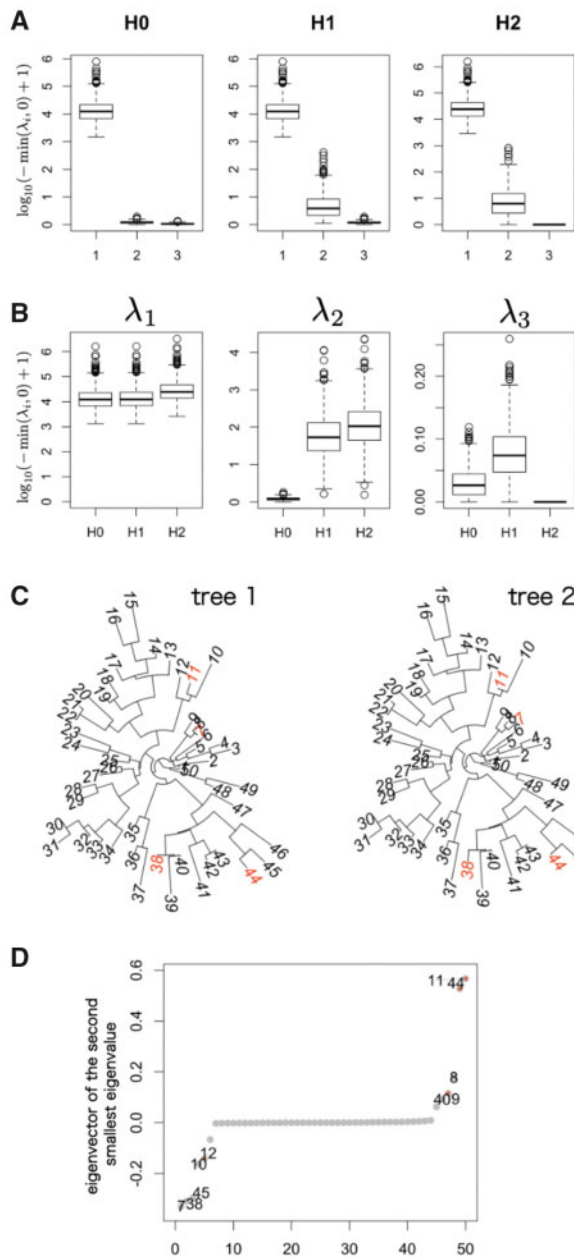
Figure 5: (A) the distributions of first, second, and third smallest eigenvalues with general distance averaging approach (H0), the general hyperbolic based approach (H1), and the proposed approach (H2). The y-axis represents the value of $\log_{10}(-\min(\lambda_i, 0) + 1)$, where λ_i is the i-th smallest eigenvalue. **(B)** The same distributions of **(A)** are visualized for first, second, and third smallest eigenvalues for visibility. **(C)** An example of a tree pair that have different branch lengths of four external branches. In this example, the branches of the nodes 11 and 44 are longer for tree 1, and the branches of the nodes 7 and 38 are longer for tree 2. **(D)** The eigenvector of the second smallest eigenvalue for the example tree pair. The eigenvector is shown in ascending order.

values were negative in most cases (993/1000). In contrast, the number of negative eigenvalues of H2 was two or less in all cases. This result implies that the proposed approach will be more effective for representing in the pseudo-Euclidean space defined by $\langle x_i, x_j \rangle_{M,K}$ (see online supplementary material for the detailed discussion of the number of negative eigenvalues).

We also investigated the eigenvector of the second smallest eigenvalue of the proposed method for the example tree pair

Figure 6: the gene trees named "cluster3190" **(A)** and "cluster4222" **(B)** in the original data, respectively. **(C)** The eigenvector of the second smallest eigenvalue of the above trees.

(Fig. 5C). The tree 1 has longer external branches corresponding to the nodes 11 and 44, and the tree 2 has longer branches corresponding to the nodes 7 and 38. The eigenvector of the second smallest negative eigenvalue is shown in Fig. 5D. The result implies that the absolute values of the eigenvector of the change nodes are larger than the values of the remaining nodes. We also applied our method for 1237 gene trees of the carnivorous Caryophyllales and analyzed all gene trees' pairs. The distance matrices of these trees are basically similar (the mean of the correlation coefficient is about 0.93), and the second small eigenvalues of the pairs of trees were positive in most cases. We show an example pair of trees (Fig. 6A and B) and the eigenvector of the second small eigenvalue (Fig. 6C), of which the second small eigenvalue is negatively large. The external branches of Beta, MJM1652, and WPYJ are longer in the tree 1 (Fig. 6A), and the branches of Spol, DrobinSFB, and Retr are longer in the tree

2 (Fig. 6B) that is consistent with the tendency of the eigenvector.

Then, we investigated whether we can predict the change point between trees by using the absolute values of the eigenvector of the second smallest eigenvalue. We calculated the AUC values for the simulation dataset of 1000 gene trees' pairs. The mean AUC value of H1 is about 0.981 and that of H2 is about 0.987 in comparison that the mean AUC value of H0 is about 0.574. This result demonstrated that the eigenvector of H1 and H2 represents the difference between the tree pair. We also investigated the eigenvalues when we integrated K phylogenetic trees with some change points. We confirmed that our approach showed K negative eigenvalues in most cases (see online supplementary material). Thus, our approach has the potential to represent both the common feature of multiple trees and the difference of trees. Our results also suggest the importance of appropriate integration, that is, averaging the inner product, in comparison to just averaging the distance matrix for embeddings.

Imputation of missing distances from partial trees

We integrated the distance matrix of the partial trees and imputed the missing distances by maximizing the second smallest eigenvalue of $H^{(1)}$ and $H^{(2)}$, respectively. Figure 7 shows the comparison of the true evolutionary distances and the imputed distances for each dataset ($|L_{a \cap b}| = 80$, 60, and 40) with $H^{(1)}$ and $H^{(2)}$ based imputation. Compared with the MSE values before optimization (0.071, 0.104, and 0.379 for $|L_{a \cap b}| = 80$, 60, and 40 datasets), the MSE values of $H^{(2)}$ based imputation are reduced significantly (0.002, 0.002, and 0.004). Therefore, our approach can recover the missing distances precisely, especially for a large distance. Those of $H^{(2)}$ are slightly smaller than those of $H^{(1)}$ (0.007, 0.006, and 0.010), and $H^{(2)}$ based method can estimate more precisely, especially for a small distance.

Discussion

Although numerous algorithms for phylogenetic tree reconstruction and phylogenetic analyses have been developed with superior performance, there has been limited discussion of these algorithms from the perspective of geometry. Our results demonstrate that a geometric view has potential to provide novel knowledge for rethinking and improving these algorithms. Furthermore, our approach suggests that the arithmetic mean of the exponential of the evolutionary distance is useful for creating an ensemble of multiple gene trees. Recently, novel kernels that can deal with indefinite inner products have been developed in representation learning studies [39, 40]. The inner product of Equation (10) is associated with these kernels, and we plan to extend our phylogenetic approach by adopting these techniques in future work.

The phylogenetic methods have contributed to various researches, including differentiation, immunogenomics, and cancer evolution. In the analysis of single-cell RNA-sequencing data, the k-nearest neighbor (k-NN) graph is usually used to reconstruct cell lineages. Our method assumes the additivity of the distance and cannot be applied for such a k-NN graph directly. Recently, a novel hyperbolic-based approach that learns a tree structure first has been proposed [22], and our idea might be useful for such approaches. Moreover, our research proposed the importance of thinking of data-specific suitable distances for representation learning and gave the chance to investigate the reasonable distances for several biological data.

In conclusion, we here propose and validate a novel metric for hyperbolic phylogenetic tree embeddings, which could precisely reconstruct evolutionary distance. Our method had a particular advantage over other embedding methods in that the angles were consistent for any two external nodes and the node of their MRCA. Moreover, our approach was shown to be useful for predicting the MRCA and the nearest-neighbor node in a
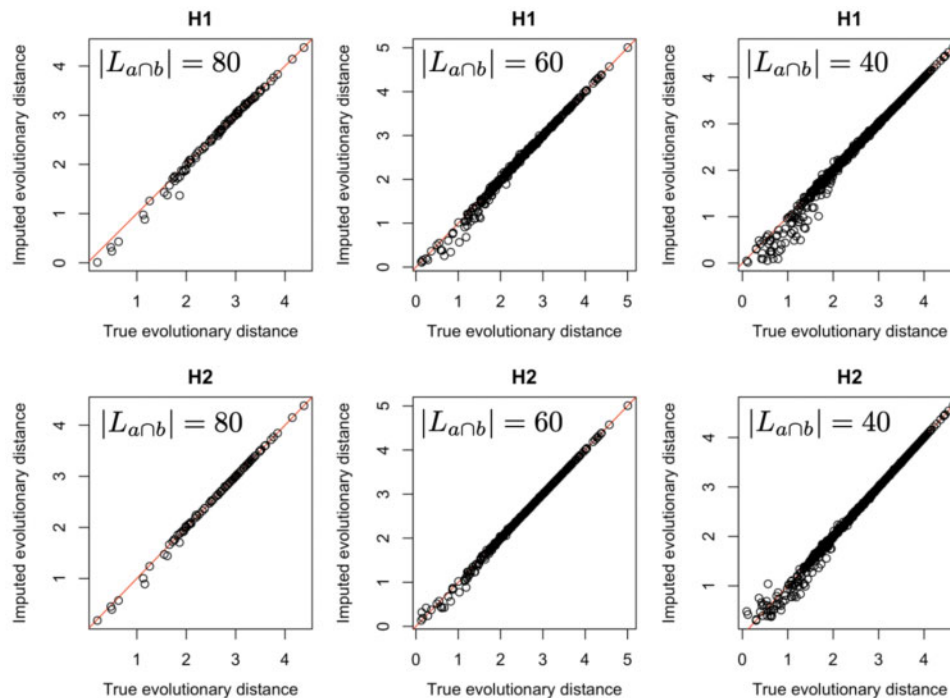

Figure 7: comparison of the true evolutionary distances and the imputed distances for $|L_{a \cap b}| = 80$, 60, and 40 datasets with $H^{(1)}$ based method (above) and $H^{(2)}$ based method (below).

partial tree with missing nodes. Our approach was also shown to be useful for integrating distance matrices of multiple trees. Our results imply that our integration approach and the eigen-decomposition of the integrated inner product matrix is effective for embedding integrated tree. In the hyperbolic geometry, the negative eigenvalue and corresponding eigenvector will represent the temporal axis. In this case of integrating trees, the smallest negative eigenvalue and corresponding eigenvector will represent the common temporal axis of the trees, and the second smallest negative eigenvalue and corresponding eigenvector will describe the time difference between trees. Besides, we proposed an approach to impute missing evolutionary distances from partial trees by maximizing the second smallest eigenvalue of $H^{(2)}$. Thus, our "tree node space" based on the evolutionary distance transformation and hyperbolic (or pseudo-Euclidean) representation will be effectual for various phylogenetic methods.

There are other geometric methods for phylogenetics that represent multiple trees ("tree space") [41], and such an approach is useful to analyze various properties of trees [42]. Our results and these studies highlight the possibility of applying "geometric-thinking" as an effective approach to the novel "tree-thinking" approach. This is particularly relevant as evolutionary analyses are expanding to different research fields beyond evolutionary inferences themselves, including differentiation, immunogenomics, and cancer evolution, requiring a novel phylogenetic strategy, and we will extend our approach to these analyses in future work.

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Data availability

The demo code is attached as a supplementary material in a compiled jupyter notebook. The code used for analyses is available on GitHub at https://github.com/hmatsu1226/HyPhyTree.

References

1. Parks DH, Rinke C, Chuvochina M *et al*. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;**2**:1533–42.
2. Hug LA, Baker BJ, Anantharaman K *et al*. A new view of the tree of life. *Nat Microbiol* 2016;**1**:1–6.
3. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* 2012;**13**:303–14.
4. Kensche PR, van Noort V, Dutilh BE, Huynen MA. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J R Soc Interface* 2008;**5**: 151–70.
5. Alföldi J, Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. *Genome Res* 2013;**23**: 1063–8.
6. Grenfell BT, Pybus OG, Gog JR *et al*. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004; **303**:327–32.
7. Knowles LL. Statistical phylogeography. *Annu Rev Ecol Evol Syst* 2009;**40**:593–612.
8. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol* 2009;**5**: e1000520.
9. Rockett R J, Arnott A, Lam C *et al .*. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat Med* 2020;**26**:1398–404.
10. Parks DH, Chuvochina M, Waite DW *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004.
11. Burki F, Roger AJ, Brown MW, Simpson AG. The new tree of eukaryotes. *Trends Ecol Evol* 2020;**35**:43–55.
12. Nagy LG, Merényi Z, Hegedüs B, Bálint B. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Res* 2020; **48**:2209–19.
13. Smith M L, Hahn M W. New Approaches for Inferring Phylogenies in the Presence of Paralogs. Trends in Genetics 2021;**37**:174–87.
14. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat Rev Genet* 2020;**21**: 410–427.
15. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet* 2017;**18**: 213–29.
16. Miho E, Yermanos A, Weber CR *et al*. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front Immunol* 2018;**9**:224.
17. Yermanos AD, Dounas AK, Stadler T *et al*. Tracing antibody repertoire evolution by systems phylogeny. *Front Immunol* 2018;**9**:2149.
18. Maximillian N, Douwe K. Poincaré embeddings for learning hierarchical representations. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc. pp. 6338–47, 2017.
19. De Sa C, Gu A, Ré C, Sala F. Representation tradeoffs for hyperbolic embeddings. *Proc Mach Learn Res* 2018;**80**:4460–9.
20. Octavian G, Gary B, Thomas H. Hyperbolic neural networks. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc. pp. 5345–55, 2018.
21. Monath N, Zaheer M, Silva D. *et al*. Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 714–22, 2019.
22. Rishi S, Gilbert AC. Tree! i am no tree! i am a low dimensional hyperbolic embedding. In: *Advances in Neural Information Processing Systems*, pp. 845–856, 2020.
23. Bingham J, Sudarsanam S. Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics* 2000;**16**:660–1.

24. Hughes T, Hyun Y, Liberles DA. Visualising very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* 2004;**5**:48.

25. Klimovskaia A, Lopez-Paz D, Bottou L, Nickel M. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nat Commun* 2020;**11**:1–9.

26. Jiarui D, Aviv R. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *BioRxiv* 2019;853457.

27. Alanis-Lobato G, Mier P, Andrade-Navarro M. The latent geometry of the human protein interaction network. *Bioinformatics* 2018;**34**:2826–34.

28. Zhou Y, Smith BH, Sharpee TO. Hyperbolic geometry of the olfactory space. *Sci Adv* 2018;**4**:eaaq1458.

29. Baum DA, Smith SD, Donovan SSS. The tree-thinking challenge. *Science* 2005;**310**:979–80.

30. Sammon JW. A nonlinear mapping for data structure analysis. *IEEE Trans Comput* 1969;**C-18**:401–9.

31. Martin K-R, Stephanie N. Hydra: a method for strain-minimizing hyperbolic embedding of network-and distance-based data. *J Complex Netw* 2020;**8**:cnaa002.

32. Molloy EK, Warnow T. Statistically consistent divide-and-conquer pipelines for phylogeny estimation using njmerge. *Algorithms Mol Biol* 2019;**14**:14.

33. Molloy EK, Warnow T. Treemerge: a new method for improving the scalability of species tree estimation methods. *Bioinformatics* 2019;**35**:i417–i426.

34. Balaban M, Sarmashghi S, Mirarab S. Apples: scalable distance-based phylogenetic placement with or without alignments. *Syst Biol* 2020;**69**:566–78.

35. Bhattacharjee A, Bayzid MS. Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. *BMC Genomics* 2020;**21**:1–14.

36. Emmanuel P, Klaus S. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics* 2019;**35**:526–8.

37. Kumar S, Stecher G, Suleski M, Hedges SB. Timetree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 2017;**34**:1812–9.

38. Walker JF, Yang Y, Moore MJ *et al.* Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous caryophyllales. *Am J Bot* 2017;**104**:858–67.

39. Akifumi O, Geewook K, Hidetoshi S. Graph embedding with shifted inner product similarity and its improved approximation capability. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 644–53.

40. Geewook K, Akifumi O, Kazuki F, Hidetoshi S. Representation learning with weighted inner product for universal approximation of general similarities. *arXiv Preprint arXiv* 2019;**1902**:10409.

41. Billera LJ, Holmes SP, Vogtmann K. Geometry of the space of phylogenetic trees. *Adv Appl Math* 2001;**27**:733–67.

42. Kim J, Rosenberg NA, Palacios JA. Distance metrics for ranked evolutionary trees. *Proc Natl Acad Sci USA* 2020;**117**:28876–86.