JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 6, 2008
© Mary Ann Liebert, Inc.
Pp. 577–591
DOI: 10.1089/cmb.2008.0068

An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees

ANNE KUPCZOK, ARNDT VON HAESELER, and STEFFEN KLAERE

ABSTRACT

The geometrical representation of the space of phylogenetic trees implies a metric on the space of weighted trees. This metric, the geodesic distance, is the length of the shortest path through that space. We present an exact algorithm to compute this metric. For biologically reasonable trees, the implementation allows fast computations of the geodesic distance, although the running time of the algorithm is worst-case exponential. The algorithm was applied to pairs of 118 gene trees of the metazoa. The results show that a special path in tree space, the cone path, which can be computed in linear time, is a good approximation of the geodesic distance. The program `GeoMeTree` is a python implementation of the geodesic distance, and it is approximations and is available from *www.cibiv.at/software/geometree*.

Key words: cone path, geodesic path, phylogeny, Robinson-Foulds distance, tree space.

1. INTRODUCTION

COMPARING PHYLOGENETIC TREES is a major task in phylogenetic research. Comparisons are necessary when trees derived from different genes are incongruent (Rokas and Carroll, 2005), when the outcomes of different reconstruction methods disagree (Dutilh et al., 2007), or when one compares the outcome of different tree reconstruction methods by simulation (Gadagkar et al., 2005).

A natural way to compare pairs of trees is to apply a distance measure. Most measures only take the topological information into account, for example, the Robinson-Foulds distance (Robinson and Foulds, 1981), the Nearest Neighbor Interchange distance (Waterman and Smith, 1978), the Subtree Prune and Regraft distance (Hein, 1990), or the quartet distance (Estabrook et al., 1985). On the other hand, there are a few measures that focus on the branch lengths of the trees, for example, the branch score distance (Kuhner and Felsenstein, 1994). The branch score distance between any two trees is the sum of the squares of differences between the branch lengths. Topological information is not incorporated explicitly in this measure, but it is considered implicitly by setting the branch lengths of non-existing splits to zero.

However, for a proper comparison of trees, it is desirable to combine both topological and branch lengths information into a single measure. One attempt for this is the weighted Robinson-Foulds distance

Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Veterinary University of Vienna, Austria.

(Robinson and Foulds, 1978), which is the sum of the absolute differences between the branch lengths of two trees. A further advantage of distance measures that consider branch length information is that they yield continuous values. This increases the distinguishability between different comparisons and allows for applications in the clustering and visualization of trees (Stockham et al., 2002; Hillis et al., 2005; Smythe et al., 2006).

Billera et al. (2001) define the tree space as the space of all weighted trees (the mathematical properties of this space have been further studied in Pachter and Sturmfels [2007] and Owen [2007]). Each tree topology is identified by a positive, real-valued hypercube, where branch lengths identify the exact position of a tree in such a hypercube. The tree space is the union of these hypercubes. Topologies are connected by less resolved topologies, therefore the tree space is connected. Every point on a path in tree space is a phylogenetic tree. It can be shown that the shortest path between any two trees exists and is unique (Billera et al., 2001). This shortest path is called geodesic path and its length is the geodesic distance. The question whether the problem of finding the geodesic path is NP-hard is still open (for opposite conjectures, see Epstein and Ingram [2003] and Owen [2007]); however, lower and upper bounds can be computed in linear time (Amenta et al., 2007).

The geodesic path can also be used to define a consensus tree method. For two trees, the midpoint of a path is a consensus tree with branch lengths. For sets of trees, different consensus methods can be defined using the tree space (Billera et al., 2001). Further issues of tree comparisons can be also be addressed in the space of trees, like finding the neighborhood of a tree or testing whether data sets are congruent (Holmes, 2005).

Here we present an exact algorithm to calculate the geodesic path and its length between two trees. We will define the mathematical framework of tree space and the steps necessary to compute a path in this space. The implemented algorithm was tested on a dataset of 118 gene trees from 21 species (Ewing et al., 2008; Ebersberger, 2007).

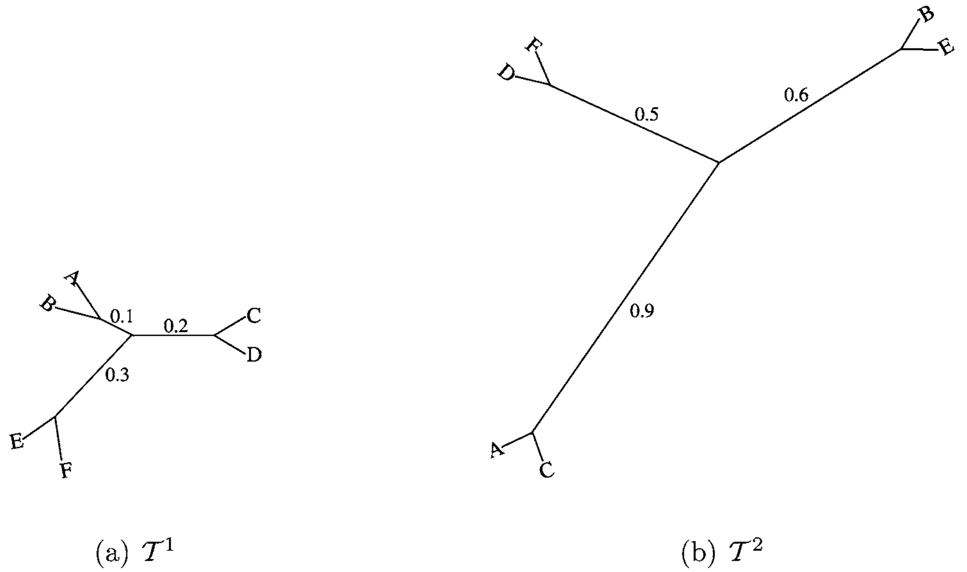
2. THE TREE SPACE

2.1. Topologies

Phylogenetic trees are leaf-labeled trees, where the leaves are called *taxa*. One distinguishes between rooted or unrooted phylogenetic trees. In the case of rooted trees, we treat the root as an additional taxon of an unrooted tree. The term *phylogenetic tree* can stand for a topology or a weighted tree. A *topology* is the branching pattern of the taxa, whereas a *weighted tree* adds branch lengths to such a topology. We will use these two terms if the discrimination is important or the term *(phylogenetic) tree* if the meaning is unambiguous in the context.

A *topology* $\mathcal{T}$ is identified by its taxon set X and its edge set, where *terminal* edges connect a leaf with an inner node and *interior* edges connect two inner nodes. If an edge of a phylogenetic tree is deleted, the tree decomposes into two connected components. Thus, the taxon set is partitioned into two sets (X_1 and X_2), one for each component. Such a bipartition is called a *split* and is identified by $X_1|X_2$ or the smaller set of X_1 and X_2 if the underlying taxon set $X = X_1 \cup X_2$ is clearly stated. A k-split refers to a partition into k and $n-k$ taxa, i.e., $k = \min(|X_1|, |X_2|)$. Since each edge in a topology corresponds to a split, we will identify a topology for taxon set X by the corresponding split set (Fig. 1).

For $n = |X|$ taxa, there are $m = 2^{n-1} - 1$ possible splits. We will denote the set of all splits for n taxa by $\mathbb{S}_n$. Analogous to the edges, we will distinguish between the n *terminal splits* and the $m - n$ *interior splits*. Two splits are called *compatible* if there is a phylogenetic tree, which contains both splits. This holds for two splits $X_1|X_2$ and $Y_1|Y_2$ if at least one of the following taxon sets is empty: $X_1 \cap Y_1$, $X_1 \cap Y_2$, $X_2 \cap Y_1$ or $X_2 \cap Y_2$. Note that terminal splits are compatible to any other split. The *compatibility graph* for a set of splits is a graph whose nodes represent the splits, and edges in the graph indicate compatibility between two splits. Figure 2 shows the compatibility graph for the interior splits for five and six taxa. For six taxa, only the induced subgraphs for $AB|CDEF$ (Fig. 2b) and $ABC|DEF$ (Fig. 2c) are shown. The *subgraph* of the compatibility graph *induced* by a split $\mathcal{S}$ consists of all splits compatible with $\mathcal{S}$. The observations for compatibility relationships on six taxa can be extended to compatibility graphs for an arbitrary number of taxa:

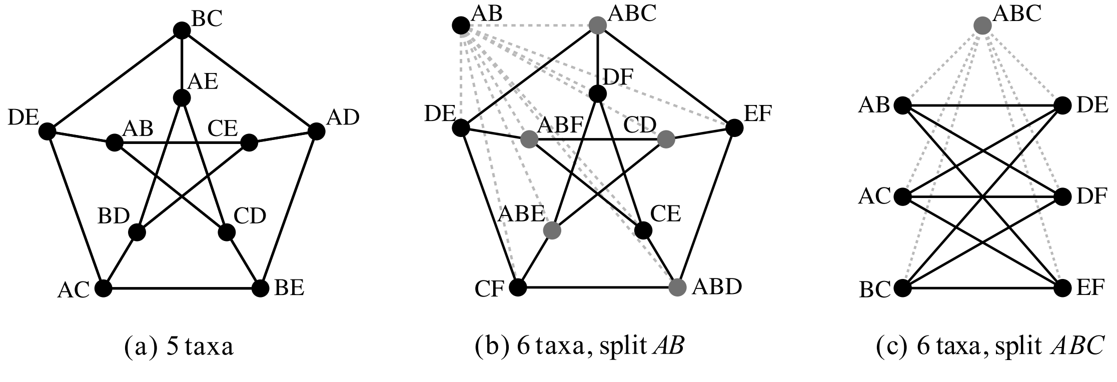


(a) $\mathcal{T}^1$

(b) $\mathcal{T}^2$

FIG. 1. Examples for phylogenetic trees on leaf set $X = \{A, B, C, D, E, F\}$: $\mathcal{T}^1 = \{A, B, C, D, E, F, AB, CD, EF\}$ and $\mathcal{T}^2 = \{A, B, C, D, E, F, AC, BE, DF\}$.

1. The induced compatibility graph for a 2-split is isomorphic to a complete compatibility graph for splits of $n - 1$ taxa. In Figure 2b, the compatibility graph of all splits compatible to the 2-split $AB|CDEF$ is isomorphic to a compatibility graph for five taxa (Fig. 2a).
2. The compatibility graph of a k-split ($k > 2$) consists of two types of nodes: Type-1-nodes correspond to splits for $k + 1$ taxa and are connected according to the complete compatibility graph for $k + 1$ taxa, and type-2-nodes correspond to $n - k + 1$ taxa and are connected accordingly. Further, all edges between the nodes of the two types exist, since all the splits in the independent subtrees are compatible. In Figure 2c, there are two compatibility graphs for four taxa, which are simply three disconnected nodes, and both classes of nodes are completely connected.



(a) 5 taxa

(b) 6 taxa, split AB

(c) 6 taxa, split ABC

FIG. 2. Compatibility graphs for five and six taxa. Vertices denote interior splits and edges indicate compatibility between the connected splits. **(a)** All interior splits of five taxa ($X = \{A, B, C, D, E\}$). Here, an edge also depicts a bifurcating topology identified by the two compatible interior splits. There are 15 edges and 15 topologies. The graph is the well-known Peterson graph. **(b)**,**(c)** Compatibility subgraphs for six taxa ($X = \{A, B, C, D, E, F\}$) induced by the splits AB, respectively ABC. The splits AB and ABC are representatives for all 2-splits and all 3-splits, since induced compatibility graphs for other splits are isomorphic to one of the graphs. The full graph for six taxa would consist of 25 vertices and 105 edges forming 105 3-cliques. Thus, there are 105 different bifurcating topologies for six taxa.

A phylogenetic tree of n taxa contains at most $n-3$ interior splits. If it contains exactly $n-3$ interior splits, all inner nodes have degree three, and the tree is called *bifurcating*, and *multifurcating* or *unresolved* otherwise. It is well-known, that $(2n-5)!! = 1 \times 3 \times \cdots \times (2n-5)$ distinct bifurcating topologies exist for $n \geq 3$ taxa (Felsenstein, 2004). In the compatibility graphs, the bifurcating trees are given as cliques of $n-3$ vertices, i.e., 2-cliques for five taxa and 3-cliques for six taxa (Fig. 2).

2.2. Weighted trees

The tree space $\mathbb{T}_n$ is the space of all weighted trees on n taxa. $\mathbb{T}_n$ is defined as follows (Billera et al., 2001). Each split is identified with a different orthogonal unit vector $\mathbf{e}^{\mathcal{S}}$ ($\mathcal{S} \in \mathbb{S}_n$) in the m-dimensional space; these are the axes of $\mathbb{T}_n$. Thus, $\mathbb{T}_n$ is a subspace of $\mathbb{R}^m$.

For each topology $\mathcal{T}$, the unit vectors associated with its splits span a $|\mathcal{T}|$-dimensional subspace. Recall that $n \leq |\mathcal{T}| \leq 2n-3$, because each topology consists at least of n terminal splits and at most $n-3$ pairwise compatible interior splits can be added.

A weighted tree $\mathbf{p}$ with topology $\mathcal{T}$ is a point in $\mathbb{T}_n$ given by

$$\mathbf{p} = \sum_{\mathcal{S} \in \mathcal{T}} p_{\mathcal{S}} \, \mathbf{e}^{\mathcal{S}}, \tag{1}$$

where $p_{\mathcal{S}}$ denotes the split weight of split $\mathcal{S}$ from topology $\mathcal{T}$. With this, every weighted tree $\mathbf{p}$ defines a split weight function $\lambda_{\mathbf{p}} : \mathbb{S}_n \to \mathbb{R}_+$ with $\lambda_{\mathbf{p}}(\mathcal{S}) = p_{\mathcal{S}}$ if $\mathcal{S} \in \mathcal{T}$ and 0 otherwise. In other words, $\lambda_{\mathbf{p}}$ assigns to each split $\mathcal{S} \in \mathcal{T}$ its weight for tree $\mathbf{p}$. Therefore, the weight function also identifies the tree (and implicitly also its topology), and we use the convention $\lambda_{\mathbf{p}}(\mathcal{T}) = \mathbf{p}$. We can apply $\lambda_{\mathbf{p}}$ to any collection of splits $\mathcal{A}$ and get a point in $\mathbb{T}_n$ with

$$\lambda_{\mathbf{p}}(\mathcal{A}) = \sum_{\mathcal{S} \in \mathcal{A}} \lambda_{\mathbf{p}}(\mathcal{S}) \, \mathbf{e}^{\mathcal{S}}.$$

In particular, $\lambda_{\mathbf{p}}$ assigns 0 to each split not in $\mathcal{T}$; thus, the point $\lambda_{\mathbf{p}}(\mathcal{A})$ lies on the subspace spanned by the splits in $\mathcal{T} \cap \mathcal{A}$. In the following, we are mainly concerned with either one or two trees and thus will use λ, respectively λ_i, $i = 1, 2$ to identify the trees.

The union of weighted trees (analogously, topologies and weight functions) forms $\mathbb{T}_n$. Unresolved topologies are also included in this space. In detail, an unresolved topology lies on the boundary of more resolved topologies. Thus, unresolved topologies connect the bifurcating topologies. An example is shown in Figure 3a, there the unresolved topology corresponds to the single axis $CD|ABE$ and connects the two bifurcating topologies.

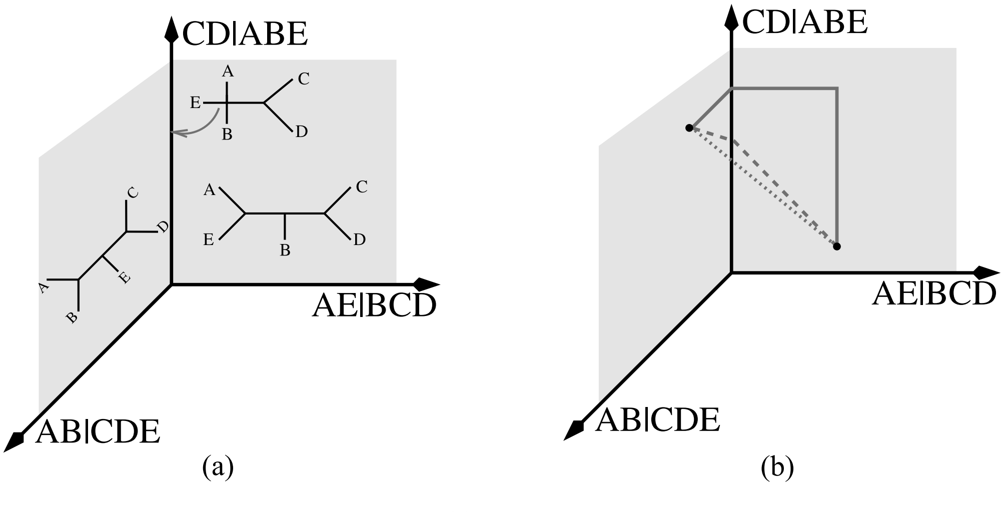
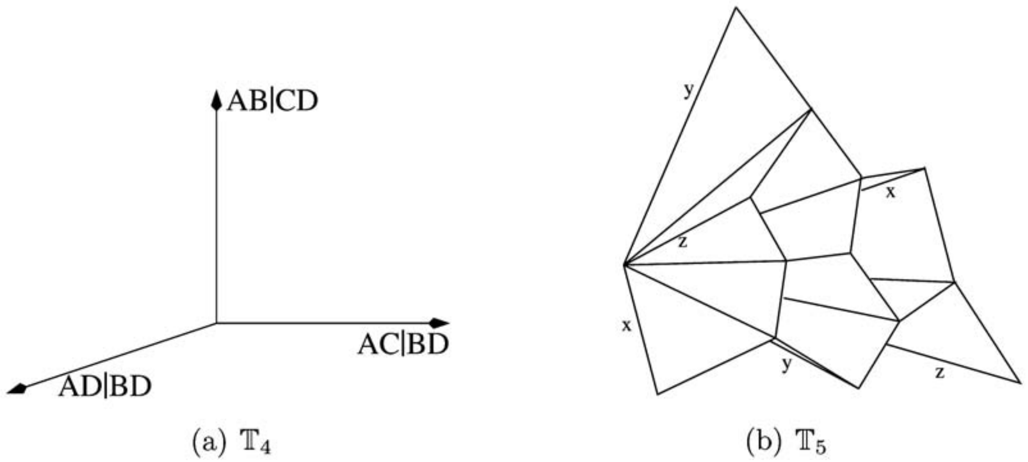


FIG. 3. Subspace of the tree space for five taxa. **(a)** Example of an unresolved topology (marked by the arrow) connecting two resolved topologies. **(b)** Example of paths in tree space: Manhattan path (line), Euclidean path (spotted) and geodesic path (dashed).



(a) $\mathbb{T}_4$

(b) $\mathbb{T}_5$

FIG. 4. Tree space for four and five taxa where only the internal splits are shown. **(a)** The split corresponding to each axis is given. Only points on the axes lie in $\mathbb{T}_4$. Billera et al. (2001) introduced **(b)**, which shows a two-dimensional description of the space spanned by the 10 nontrivial splits for five taxa. Here, each topology is a 2D-plane. Note that the picture is entangled as some splits (x, y, z) are shown twice at the boundary of the figure.

Figure 4 shows a visualization of $\mathbb{T}_4$ and $\mathbb{T}_5$ where only the interior splits are illustrated. The previous considerations about compatibilities of splits help to understand how to extend these pictures for higher dimensions:

1. When deleting all splits incompatible with a 2-split, one reduces the dimension of $\mathbb{T}_n$. In particular, one projects $\mathbb{T}_n$ on the subspace spanned by the compatible splits. This results in a space isomorphic to $\mathbb{T}_{n-1}$ with two extra dimensions, one for the 2-split and one for the additional terminal split.
2. When projecting $\mathbb{T}_n$ on the vector space spanned by the splits compatible to a k-split ($k > 2$), the resulting space has the following structure: $\mathbb{T}_{k+1} \times \mathbb{T}'_{n-k+1}$, where $\mathbb{T}'_{n-k+1}$ is $\mathbb{T}_{n-k+1}$ with one terminal split missing. An example is shown in Figure 5.

Further note that $\mathbb{T}_n$ is a true subspace of $\mathbb{R}^m_+$. Already in $\mathbb{T}_4$ we see that the tree space is sparse. Although the dimension of the space is $m = 3$, each tree has only one internal split. For $\mathbb{T}_5$, the dimension is $m = 15$, but each tree lies in a 2D-plane. This disparity increases for higher dimensions, since the number of splits for one tree increases linearly with n but the number of possible splits, m, increases exponentially with n.

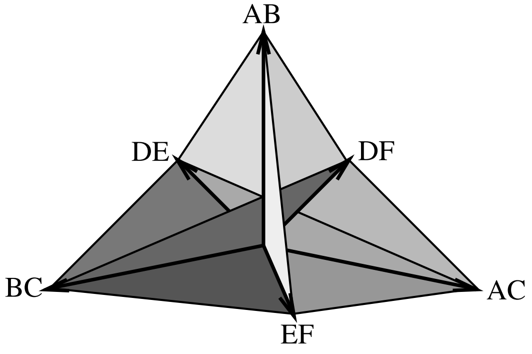


FIG. 5. Subspace of $\mathbb{T}_6$ showing only the internal splits compatible to the split ABC. This corresponds to a cross-product of the two $\mathbb{T}_4$-spaces with axes $\{AB, AC, BC\}$ and $\{DE, DF, EF\}$, respectively. For the compatibility relationships, see Figure 2c.

2.3. Paths and distances between trees

There are many different distance measures for the topological difference between two trees. The most-common one is the *Robinson-Foulds distance* (RF) (Robinson and Foulds, 1981), which counts the splits not in both topologies. In the set-theoretical sense, the RF distance between two topologies $\mathcal{T}^1$ and $\mathcal{T}^2$ is given by the size of the symmetric difference:

$$\mathrm{RF}(\mathcal{T}^1, \mathcal{T}^2) = |\mathcal{T}^1 \Delta \mathcal{T}^2| = |(\mathcal{T}^1 \cup \mathcal{T}^2) \setminus (\mathcal{T}^1 \cap \mathcal{T}^2)|.$$

The example topologies from Figure 1 have $\mathrm{RF}(\mathcal{T}^1, \mathcal{T}^2) = 6$ because each topology has three interior splits and no interior split in common.

One measure to compare two trees with respect to both topology and branch lengths is the *weighted Robinson-Foulds distance* (RF_w) (Robinson and Foulds, 1978). For two weighted trees λ_1 and λ_2 with topology $\mathcal{T}^1$ and $\mathcal{T}^2$, respectively, the RF_w distance is given as:

$$\mathrm{RF}_w(\lambda_1, \lambda_2) = \sum_{\mathcal{S} \in \mathbb{S}_n} |\lambda_1(\mathcal{S}) - \lambda_2(\mathcal{S})|.$$

This measure corresponds to the L^1 norm or the length of the Manhattan path in $\mathbb{T}_n$. An example of the Manhattan path in $\mathbb{T}_5$ is shown in Figure 3b. For the weighted trees in Figure 1, the RF_w distance is equal to 2.6.

Another measure that respects branch lengths is the branch-score distance (BS) (Kuhner and Felsenstein, 1994). It corresponds to the Euclidean distance between the branch lengths of all splits in $\mathbb{R}^m$:

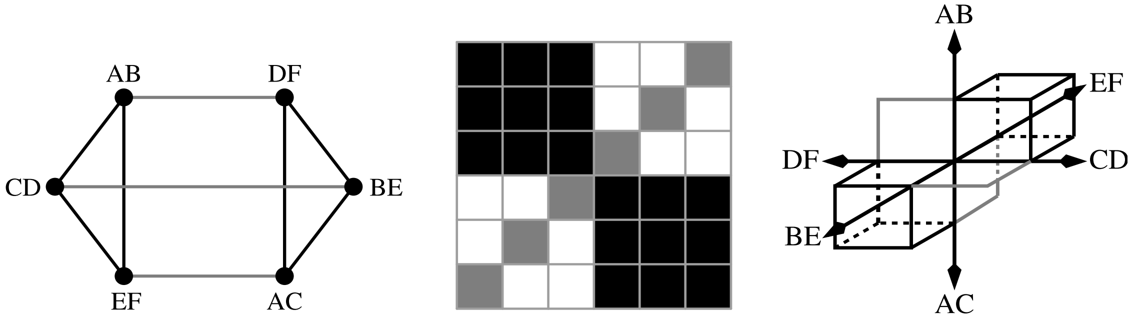
$$\mathrm{BS}(\lambda_1, \lambda_2) = \|\lambda_1(\mathcal{T}^1) - \lambda_2(\mathcal{T}^2)\| = \sqrt{\sum_{\mathcal{S} \in \mathbb{S}_n} (\lambda_1(\mathcal{S}) - \lambda_2(\mathcal{S}))^2}.$$

For two different topologies, the corresponding path in $\mathbb{R}^m_+$ (the Euclidean path) is not a path in $\mathbb{T}_n$ (Fig. 3b). This implies that the Euclidean distance does not correspond to the L^2-norm on tree space. But Billera et al. (2001) have shown that an L^2-norm in $\mathbb{T}_n$ exists by proving that $\mathbb{T}_n$ is a CAT(0)-space (Bridson and Haefliger, 1999). In CAT(0)-spaces, a unique shortest path exists between any two points. These paths are called *geodesics*, and their length, the *geodesic distance*, is a metric on $\mathbb{T}_n$ which corresponds to the L^2-norm.

In the following, we will provide an algorithm to determine the geodesic path between two trees of the same leaf set. As mentioned earlier, the dimension of the tree space, m, increases exponentially with n. But given two topologies $\mathcal{T}^1$ and $\mathcal{T}^2$, we only need to consider the splits in these two topologies since the geodesic path will never pass other splits (Vogtmann, 2003). Figure 6 depicts the implications of this statement for the trees in Figure 1.

For trees containing common internal splits, the problem of finding the geodesic path can be simplified further: for $\mathcal{S}_c \in \mathcal{T}^1 \cap \mathcal{T}^2$ with $\mathcal{S}_c = X_A|X_B$ every internal split in $\mathcal{T}^1$ further resolves either taxon set X_A (the splits $\mathcal{T}^1_A$) or X_B (the splits $\mathcal{T}^1_B$). These two split sets are then subtopologies on the taxon sets X_A and X_B (analogous for $\mathcal{T}^2$). Since all splits from subtopology $\mathcal{T}^i_A$ are compatible with all splits from subtopology $\mathcal{T}^j_B$, $i, j = 1, 2$, paths through these subtopologies are independent. Therefore, the geodesic for all splits can be found by looking separately at taxon set X_A (using the subtopologies $\mathcal{T}^1_A$ and $\mathcal{T}^2_A$) and taxon set X_B (using $\mathcal{T}^1_B$ and $\mathcal{T}^2_B$) and assembling the paths afterwards (Vogtmann, 2003). As a consequence, we will assume in the following that the topologies are fully decomposed and contain no common splits. This involves both reducing the topologies and setting the other split weights (including the terminal splits) to zero.

Another useful property of geodesic paths is that they are *piecewise linear*. In Figure 3b, the geodesic path between two trees with only one different split is shown. This path is linear between one tree and the intersection point with the axis. Therefore, the idea of the algorithm presented here is to enumerate all possible intersections efficiently, compute the length of a path given special intersections, and find the shortest among these paths.



(a) Compatibility graph (b) Adjacency matrix (c) Space visualization

FIG. 6. Three visualizations for the example topologies $\mathcal{T}^1$ and $\mathcal{T}^2$ from Figure 1. **(a)** Compatibility map of the six interior splits from the two topologies. Compatibilities between splits of the two topologies are highlighted in gray. We see that no bifurcating topology other then $\mathcal{T}^1$ and $\mathcal{T}^2$ can be formed from these splits. **(b)** Associated adjacency matrix with the same color code, where white entries denote incompatibility. **(c)** Fraction of tree space $\mathbb{T}_6$ spanned by the interior splits in $\mathcal{T}^1$ and $\mathcal{T}^2$ with the same color code; the three gray planes correspond to unresolved topologies spanned by splits from both trees.

3. FINDING THE SHORTEST PATH IN TREE SPACE

3.1. Introduction

In this section, we will introduce the prerequisites to compute the geodesic path between two weighted trees λ_1 and λ_2 with topology $\mathcal{T}^1$ and $\mathcal{T}^2$, respectively. As explained in Section 2.3, we already decomposed the topologies such that $\mathcal{T}^1$ and $\mathcal{T}^2$ contain no common splits. We will further assume that the respective topologies are bifurcating. Then both topologies contain the same number of splits, i.e., $d = |\mathcal{T}^1| = |\mathcal{T}^2|$. Thus, we reduced $\mathbb{T}_n$ to a $2d$-dimensional subspace by deleting all splits not in $\mathcal{T}^1 \cup \mathcal{T}^2$. The remaining split set is $\mathbb{S}'$ with $|\mathbb{S}'| = 2d$. Note that the splits in $\mathbb{S}'$ are the only ones that contribute to the Robinson-Foulds distance, and $2d$ corresponds to the RF distance for the reduced topologies.

Example. The trees in Figure 1 on leaf set $\mathrm{X} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$ are each composed of three interior splits. They do not have interior splits in common and therefore $d = 3$, $\mathbb{S}' = \{AB, CD, EF, AC, BE, DF\}$ and

$$\mathcal{T}^1 = \{AB, CD, EF\}, \qquad \mathcal{T}^2 = \{AC, BE, DF\}$$

with the following weights for the topologies:

$$\lambda_1(\mathcal{T}^1) = (0.1, 0.2, 0.3, 0, 0, 0), \qquad \lambda_1(\mathcal{T}^2) = (0, 0, 0, 0, 0, 0),$$

$$\lambda_2(\mathcal{T}^1) = (0, 0, 0, 0, 0, 0), \qquad \lambda_2(\mathcal{T}^2) = (0, 0, 0, 0.9, 0.6, 0.5).$$

3.2. Legal topologies

We construct *legal topologies* $\mathcal{A}$ that are formed by the splits in $\mathbb{S}'$ and fulfill the compatibility condition. Legal topologies are cliques in the compatibility graph with vertex set $\mathbb{S}'$.

Further, a legal topology $\mathcal{A}$ is required to be *maximal*, i.e., adding splits $\mathcal{S} \in \mathbb{S}'$ to $\mathcal{A}$ will violate the compatibility within $\mathcal{A}$. This corresponds to extracting the maximal cliques from the compatibility graph. Non-maximal topologies are by definition composed of fewer splits and are therefore subtopologies of a maximal topology. This implies that any path through legal subtopologies runs along the boundary of a legal topology. Thus, paths through subtopologies are already contained in the possible paths through other legal topologies by setting the corresponding split weight to 0.

Let $\mathfrak{A}$ be the set of all legal topologies. By this definition, $\mathcal{T}^1 \in \mathfrak{A}$ and $\mathcal{T}^2 \in \mathfrak{A}$.

Example. For the trees in Figure 1, the compatibility graph is shown in Figure 6a. The set $\mathfrak{A}$ of legal topologies are the maximal cliques of the graph; thus,

$$\mathfrak{A} = \{\mathcal{T}^1, \quad \mathcal{T}^2, \quad \{AB, DF\}, \quad \{CD, BE\}, \quad \{EF, AC\}\}$$

3.3. Legal paths

A *sequence of legal topologies* connects $\mathcal{T}^1$ with $\mathcal{T}^2$ by passing through legal topologies from $\mathfrak{A}$ in such a manner that in each step at least one split from $\mathcal{T}^1$ is replaced by at least one split from $\mathcal{T}^2$.

Accordingly, a sequence of legal topologies $(\mathcal{A}^j)_{j=0}^k$ with $k \leq d$ and $\mathcal{A}^0 = \mathcal{T}^1$, $\mathcal{A}^k = \mathcal{T}^2$ must fulfill the following two conditions for all $j = 0, \ldots, k-1$:

$$\mathcal{A}^{j+1} \cap \mathcal{T}^1 \subset \mathcal{A}^j \cap \mathcal{T}^1 \quad \text{and} \quad \mathcal{A}^j \cap \mathcal{T}^2 \subset \mathcal{A}^{j+1} \cap \mathcal{T}^2 \tag{2}$$

i.e., no split from $\mathcal{T}^1$ can reemerge in the sequence and no split from $\mathcal{T}^2$ can be lost.

From the adjacent topologies $\mathcal{A}^j$ and $\mathcal{A}^{j+1}$ we are interested in the splits that changed between these sets. These *transitions* I^j $(j = 0, \ldots, k-1)$ are given as the symmetric difference between these two sets, i.e., $I^j = \mathcal{A}^j \Delta \mathcal{A}^{j+1}$. Note that the series of transitions $(I^j)_{j=0}^{k-1}$ form a partition of the split set $\mathbb{S}'$.

From this sequence of topologies we generate a piecewise linear path. The path is linear while passing through a topology $\mathcal{A}$. Two adjacent topologies are connected by a *transition point* where all splits in I^j have weight 0.

A path is parameterized with constant speed by a piecewise linear function $g : [0,1] \to \mathbb{R}_+^{2d}$ with $g(0) = \lambda_1(\mathcal{T}^1)$ and $g(1) = \lambda_2(\mathcal{T}^2)$ (Vogtmann, 2003; Bridson and Haefliger, 1999). For each transition from topology $\mathcal{A}^j$ to $\mathcal{A}^{j+1}$, there exists a time t_j at which $g(t_j)_i$ is zero for the splits $i \in I^j$. In other words, t_j is the *transition time* in which the splits $i \in \mathcal{T}^1 \cap I^j$ are reduced to length zero and $1 - t_j$ is the transition time in which the splits $i \in \mathcal{T}^2 \cap I^j$ are expanded from zero to their weight in λ_2. Due to the constant speed condition of the path (Vogtmann, 2003), the transition times t_j are calculated from the transitions by

$$t_j = \frac{\|\lambda_1(I^j)\|}{\|\lambda_1(I^j)\| + \|\lambda_2(I^j)\|}$$

For a sequence of topologies to contain a *legal path*, the following condition must be satisfied: The points in which the sequence of topologies traverses from topology $\mathcal{A}^j$ to $\mathcal{A}^{j+1}$, $j = 0, \ldots, k-1$, can only be visited in the proposed order. This is ensured when the transition times $(t_j)_{j=0}^{k-1}$ are increasing. For a legal path g, the entries of g for an arbitrary time $t \in [0,1]$ are given by

$$g_i(t) = \left\{\begin{array}{ll} -\dfrac{\lambda_1(i)}{t_j}(t - t_j), & i \in \mathcal{T}^1 \cap I^j \text{ and } t < t_j, \\[2ex] \dfrac{\lambda_2(i)}{1 - t_j}(t - t_j), & i \in \mathcal{T}^2 \cap I^j \text{ and } t > t_j, \\[2ex] 0, & \text{otherwise} \end{array}\right\}.$$

With this, the function g describes the path between the two weighted trees, which changes direction at $(t_j)_{j=0}^{k-1}$ and its length is computed by:

$$\|g\| = \sum_{j=1}^{k} \|g(t_j) - g(t_{j-1})\|, \quad \text{with } t_0 = 0 \text{ and } t_k = 1.$$

There is always a legal path defined by a single transition at time t^* with $I^* = \mathbb{S}'$. This path simultaneously replaces all splits in $\mathcal{T}^1$ by all splits in $\mathcal{T}^2$ at time t^* and is called *cone path*, because it passes through the origin of the $2d$-dimensional subspace of $\mathbb{T}_n$.

Example. The example set $\mathfrak{A}$ from Figure 1 suggests four sequences of legal topologies with the following transitions and transition times:

Path 1: $\mathcal{T}^1 \xrightarrow[t_1=0.42]{\{CD,EF,AC\}} \{AB, DF\} \xrightarrow[t_2=0.09]{\{AB,AC,BE\}} \mathcal{T}^2$

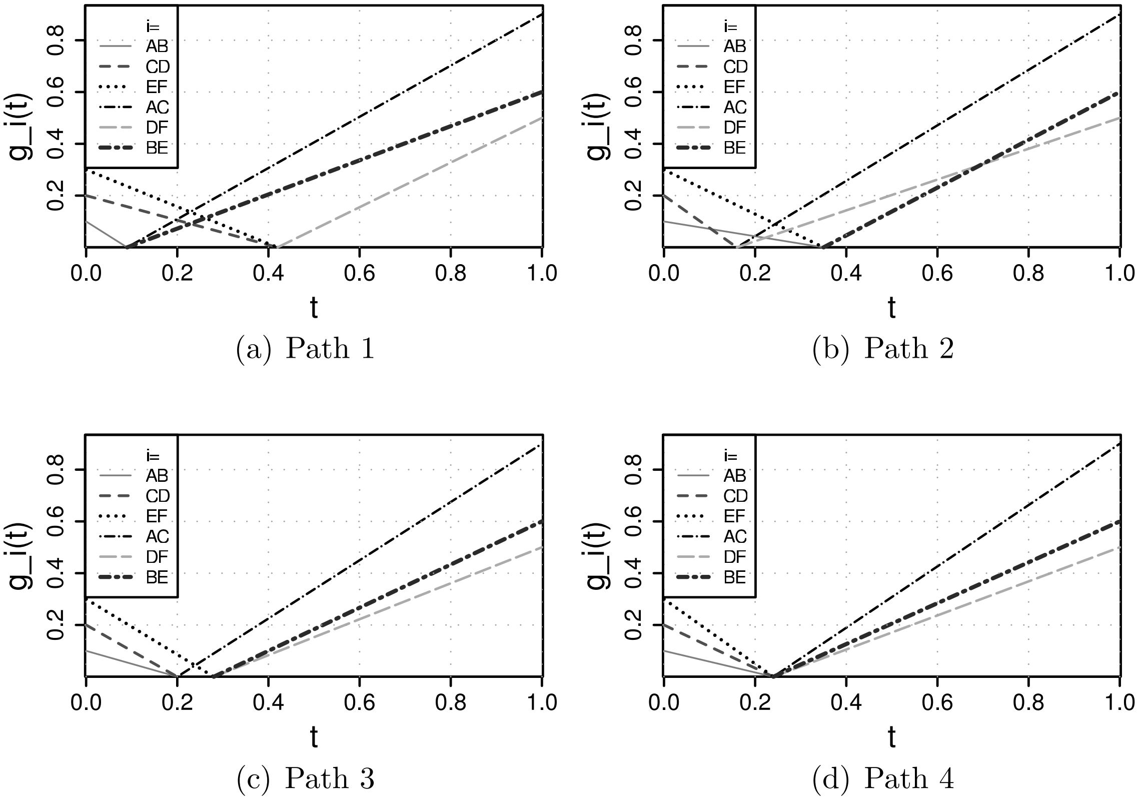
Path 2: $\mathcal{T}^1 \xrightarrow[t_1=0.35]{\{AB,EF,BE\}} \{CD, BE\} \xrightarrow[t_2=0.16]{\{CD,AC,DF\}} \mathcal{T}^2$

Path 3: $\mathcal{T}^1 \xrightarrow[t_1=0.2]{\{AB,CD,AC\}} \{EF, AC\} \xrightarrow[t_2=0.28]{\{EF,BE,DF\}} \mathcal{T}^2$

Path 4: $\mathcal{T}^1 \xrightarrow[t^*=0.24]{\{AB,CD,EF,AC,BE,DF\}} \mathcal{T}^2$

Because $t_1 > t_2$ for the first sequence, it does not yield a legal path. This is visualized in Figure 7a: The topologies suggested by the transition times are $\mathcal{T}^1 \xrightarrow[t_2=0.09]{\{AB,AC,BE\}} \{CD, EF, AC, BE\} \xrightarrow[t_1=0.42]{\{CD,EF,AC\}} \mathcal{T}^2$. So four splits coexist between time 0.09 and 0.42 where the splits from the first tree are incompatible to the splits from the second tree. Thus, the condition of a legal path can be checked by testing whether the times in a sequence of topologies are increasing.

Since $t_1 > t_2$ also holds for the second sequence (Fig. 7b), only path 3 (Fig. 7c; length 1.5592) and path 4 (Fig. 7d; the cone path, length 1.5658) correspond to legal paths. Comparing their lengths shows that path 3 is the geodesic path between the two trees.



(a) Path 1

(b) Path 2

(c) Path 3

(d) Path 4

FIG. 7. Parameterizations for the four possible paths of the example trees in Figure 1.

3.4. Multifurcating trees

Only one extension of our findings is necessary to include trees with less than $n-3$ interior splits (*multifurcating trees*). If $\mathcal{T}^1$ is multifurcating, there may be splits in $\mathcal{T}^2$ that are compatible to every split in $\mathcal{T}^1$ and vice versa. The length of a split in $\mathcal{T}^2$ with this property will be extended immediately from the beginning, while the length of a split in $\mathcal{T}^1$ with this property will be reduced until the end. Thus, only the remaining topologies $\hat{\mathcal{T}}^1$ and $\hat{\mathcal{T}}^2$ without these splits are relevant for the calculation of the transition times and contribute to the topologies in $\mathfrak{A}$. Note that $\hat{\mathcal{T}}^1$ and $\hat{\mathcal{T}}^2$ do not necessarily contain the same number of splits.

3.5. Implementation details

The presented algorithm for the geodesic path comprises several steps: decomposing the topologies to the sets $\mathcal{T}^1$ and $\mathcal{T}^2$; building the legal topologies $\mathfrak{A}$; arranging them in legal sequences; and extracting the legal paths, which are the legal sequences where the transition times are in the correct order. The geodesic path is then the shortest of these legal paths. Our implementation does not follow these steps, but computes the legal topologies together with the transitions and their respective times. The computation starts with $\mathcal{T}^1$ and generates all possible transitions I, which lead to a maximal legal topology. A directed acyclic graph (DAG) is thereby generated whose node set is $\mathfrak{A}$, and an edge is inserted for every generated transition and labeled with its time. A legal sequence is a directed path in the DAG which connects $\mathcal{T}^1$ and $\mathcal{T}^2$, and a legal path is a legal sequence with increasing edge weights on the path through the DAG. Not all sequences have to be enumerated until the end. The transition times are computed co-instantaneously and tested for an ascending sequence. Illegal paths are identified and terminated before reaching $\mathcal{T}^2$.

The time-limiting step is the generation of all transitions I, which is done for each topology in $\mathfrak{A}$. First, for topology $\mathcal{T}^1$, all possible I leading to a maximal legal topology are generated. There are not more than 2^d. Then the complete set $\mathfrak{A}$ is already known, and for each element, all possible index vectors are generated again. This yields

$$\underbrace{2^d}_{\text{Generation of } \mathfrak{A} \text{ from } \mathcal{T}^1} \quad + \quad \underbrace{2^d \times 2^d}_{\text{Generation of all } I \text{ for the other topologies}} \quad = \mathcal{O}(2^{2d})$$

for building the graph. But note that, because of incompatibilities, the size of $\mathfrak{A}$ is much smaller than 2^d. Further, the algorithm is exponential in d, which is small for topologically similar trees and can be decreased by decomposing the trees (Section 2.3).

The algorithm is implemented in a python program called `GeoMeTree`, which is available from *www.cibiv.at/software/geometree*. The program has been used for the calculations in the next section.

4. RESULTS

4.1. Data

A dataset was generated from 216 alignments of 20 metazoa species and yeast as an outgroup (Ewing et al., 2008; Ebersberger, 2007). The orthologs were extracted from the Inparanoid database (O'Brien et al., 2005), where pairwise orthologs of eukaryotes are stored. Orthology is expanded to cover all 21 species by taking an arbitrary order of the 21 species and determining the orthologous pairs between neighboring species. If a chain occurs where the protein in the first and last species are also orthologs, this protein is added to the data set. This resulted in 216 alignments for 21 species, where each alignment consists of putatively orthologs. For these, alignments were produced with T-coffee (Notredame et al., 2000). Maximum likelihood gene trees were reconstructed for each of the gene alignments with phyML (Guindon and Gascuel, 2003). A widely accepted species tree had been found for these gene trees with different methods (Ewing et al., 2008).

From the resulting set of 216 trees, we used the 118 strictly bifurcating trees (6903 pairs) for the distance computations. The resulting weights for each tree (see Equation [1]) are normalized (i.e., they have a Euclidean norm of 1). Otherwise, the branch lengths are expected to dominate the distance between two trees, while differences between their topologies have less influence on the measure.

4.2. Dimension and number of paths

The computations were first done without decomposition (Section 2.3). As stated earlier, the dimension d is the number of splits in one topology but not in the other. For pairs of bifurcating trees, d corresponds to half of the Robinson-Foulds distance. For 21 taxa, the maximal d is 18, but the observed dimension d in the data ranged from 0 to 12 (Fig. 8a). The few pairs with a high dimension d are mainly caused by a few gene trees with many incongruencies to the species tree.

Without decomposition, the mean number of legal paths in tree space increases exponentially with the dimension (Fig. 8b). This is due to the fact that the number of legal topologies increases exponentially with d, and more legal paths are expected for a larger set of legal topologies. A substantially smaller number of paths is explored when the topologies are decomposed. The number with decomposition in Figure 8b refer to the product of the number of paths through the independent decompositions, which is smaller than the complete number of paths.

4.3. Computing time

Although the algorithm is exponential in d, the program is reasonably fast. The mean runtime without decomposition was 0.4 sec. However, the time to compute the distance for one pair of trees strongly depends on the number of paths evaluated. This is reflected by the longer runtime for high dimensions and without decomposition (Table 1, left part). The runtime is highly improved when decomposition of the trees is applied (Table 1, right part). For the two runs which then show a runtime of >30 sec, the trees could not be decomposed.

4.4. Approximations of the Geodesic distance

A lower and an upper bound is known for the geodesic distance (Amenta et al., 2007). The lower bound is the branch score distance (Kuhner and Felsenstein, 1994; Felsenstein, 2005), which is the Euclidean distance between the branch lengths vectors of the two trees. The upper bound is the length of the cone path. Amenta et al. (2007) showed that these two lengths differ at most by a factor of $\sqrt{2}$.

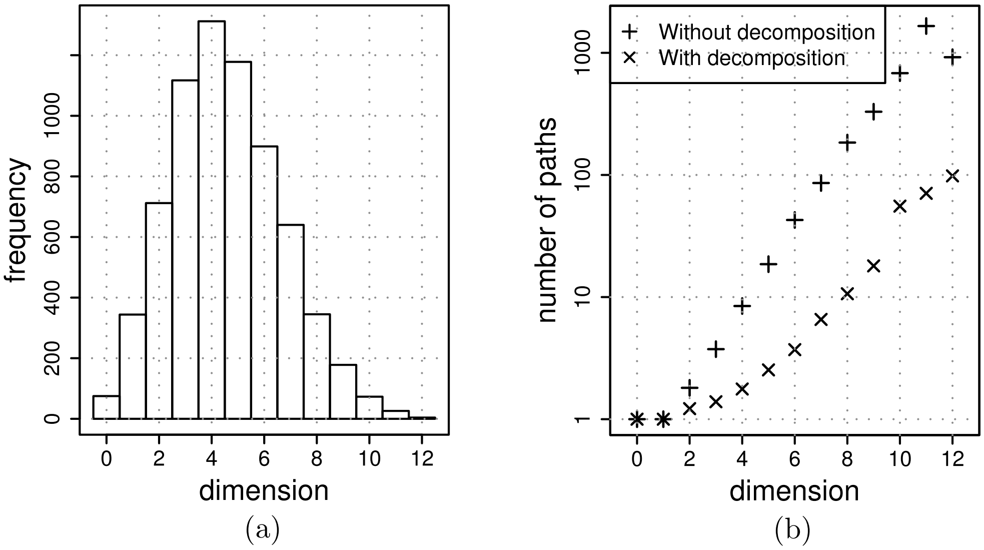
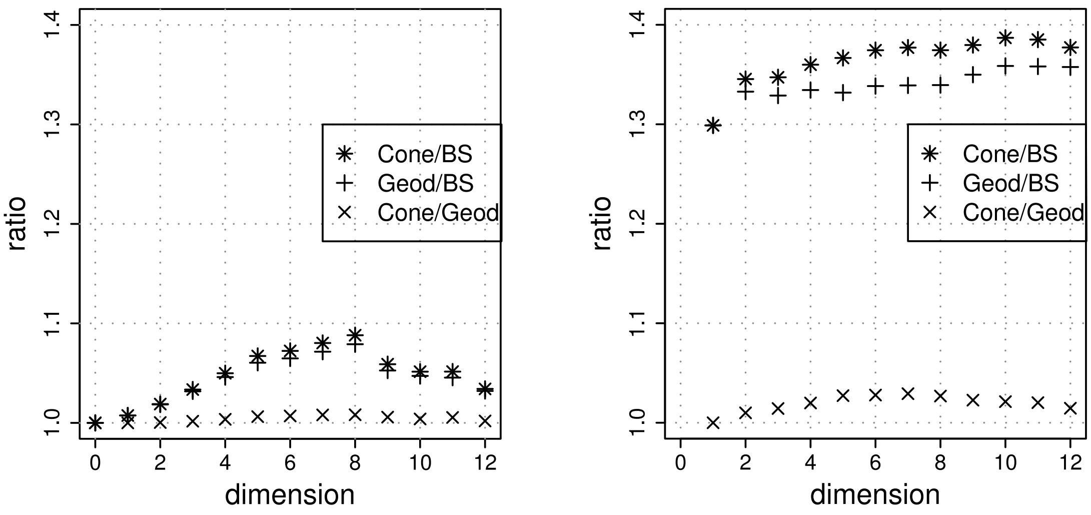


FIG. 8. Frequency of observed dimensions and mean number of paths. **(a)** Frequency (number of pairs of trees) of the observed dimension d. **(b)** Mean number of paths (log-scale) for the computation without decomposition and the product of the paths through independent decompositions].

Table 1. *Time*: Mean and Maximal Time Consumption for the Different Dimensions *d* Without and With Decomposition. *Dec.*: Mean and Maximal Number of Decompositions for a Dimension *d*

	Without decomposition		*With decomposition*			
	Time				*Dec.*	
d	*Mean*	*Max*	*Mean*	*Max*	*Mean*	*Max*
0	0 sec	0 sec	0 sec	0 sec	1	1
1	0 sec	0 sec	0 sec	0 sec	1	1
2	0 sec	0 sec	0 sec	0 sec	1.6	2
3	0 sec	0 sec	0 sec	0 sec	2.2	3
4	0 sec	0.1 sec	0 sec	0.1 sec	2.6	4
5	0.2 sec	1 sec	0 sec	0.2 sec	2.9	5
6	0.9 sec	9.9 sec	0 sec	0.7 sec	3	5
7	3.7 sec	102 sec	0 sec	0.8 sec	2.7	6
8	15.8 sec	11.4 m	0.1 sec	7.7 sec	2.7	6
9	56.1 sec	33.7 m	0.2 sec	22.6 sec	2.4	5
10	8.6 m	9.9 h	2.1 sec	46 sec	2.1	4
11	43.7 m	17.9 h	2.7 sec	44.3 sec	2	4
12	2.3 m	5.4 m	3.9 sec	13.9 sec	1.5	2

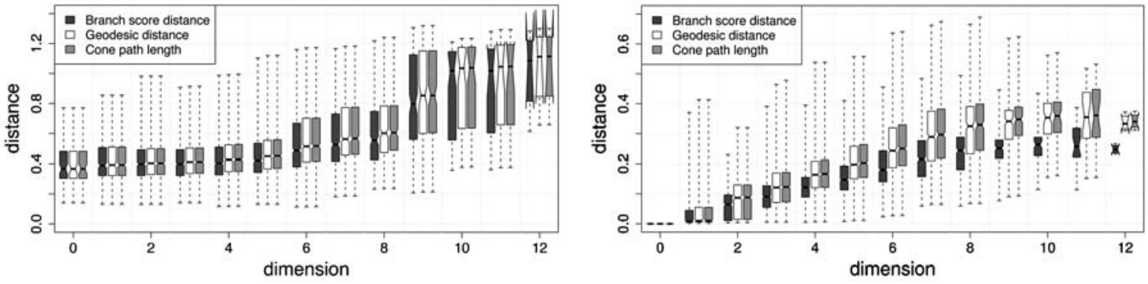
We observed that the mean ratio of the cone path to the branch score distance is close to 1.4, when the distance is computed only from the differing splits (Fig. 9). In contrast, the ratio is smaller than 1.1 when all splits are considered. Thus, both distances give a tight interval for the geodesic distance. Figure 9 also shows that the geodesic distance is better approximated by the cone path than by the branch score distance. This is expected, since the cone path is already a path in tree space. In contrast, the branch score distance measures the length of the Euclidean path between two trees, which is not a path in tree space for trees with at least one different split.



(a) All splits

(b) Different splits

FIG. 9. Means of the ratios of the distance measures. Cone, cone path length; BS, branch score distance; Geod, geodesic distance.



(a) All splits

(b) Different splits

FIG. 10. Relation between Robinson-Foulds distance and the three distances: Robinson-Foulds distance is two times the dimension d of the pair. For every category of d, a boxplot of the distribution of each of the three distances is drawn.

4.5. Relationship to the Robinson-Foulds distance

The Robinson-Foulds distance (Robinson and Foulds, 1981) is a pure topological and discrete distance measure, which counts the number of different splits between two trees and thus corresponds to $2d$ in our notation. Since it is the prevalent distance measure for phylogenetic trees, it would be preferable if continuous distance measures also display this topological information and extend it with the branch lengths information.

As Figure 10a indicates, comparing pairs of trees over all their splits results in a distance range that appears to have little correlation with the Robinson-Foulds distance, especially for small values of d. The notches of the boxplots of different dimensions are overlapping for the geodesic distance (and also for its approximations). This indicates that the medians do not differ significantly. Under these circumstances, the lengths of the branches have a much higher influence on the distance than the topological features. If one intends to make the comparison more sensitive to topological differences, we suggest reducing the study to the $2d$ different splits (Fig. 10b). Here, the notches do not overlap until a dimension of nine, indicating that the median geodesic distance is increasing with increasing dimension. However, the broad distributions show that the branch lengths do still have a substantial impact.

5. DISCUSSION

We presented an exact algorithm for computing the geodesic distance and showed its applicability for phylogenetic trees. For a pair of trees, the algorithm constructs legal topologies formed from splits of the input trees. From these topologies, it enumerates the legal paths leading from one tree through a sequence of legal topologies to the other tree. We employed computational techniques to reduce the number of paths enumerated. This facilitated the calculation of the geodesic path between two trees in reasonable time, although the algorithm is still exponential in the number of different splits. Currently, no other implementation of the geodesic distance is published, but another approach is in preparation (Owen, 2007).

The availability of a distance metric in tree space allows us to address further issues. These include clustering or visualizing trees (Hillis et al., 2005) or finding the center of a set of trees (Billera et al., 2001), which can be interpreted as a consensus method. One possibility to define a consensus method for a given distance metric are median trees. With the Robinson-Foulds distance, the median tree corresponds to the majority-rule consensus tree (Barthélemy and McMorris, 1986), which is one of the prevalent consensus methods (Bryant, 2003). For the weighted Robinson-Foulds distance, the median tree is given by the majority rule consensus tree in which each branch length is the minimum of the lengths of the respective split in the data (Pattengale, 2005). The median tree for a given set of trees under the geodesic distance corresponds to the tree with the smallest total distance to all trees in the regarded set. No closed formula to determine the median tree under the geodesic distance is known, so searching over the whole tree space would be necessary. For simplification, one could assume that the median tree is among the observed trees and thus determine the tree with the smallest total distance from all pairwise distances.

We showed the applicability of the presented algorithm on a metazoa dataset which was generated from 118 alignments of 21 species (Ewing et al., 2008; Ebersberger, 2007). In this example, the contribution of the branch lengths overwhelms the influence of the topologies (Fig. 10a) on the distance. To incorporate the topological signals, we suggest as an appropriate distance measure the length of the geodesic path through those splits exclusive to one of the trees considered (Fig. 10b).

Another notable observation is the small factor by which the cone path and the geodesic distance differ (Fig. 9). This was also observed for pairs of trees simulated under a Yule (1924) process (results not shown). Thus, the length of the cone path is a useful approximation of the geodesic distance, especially since it incorporates topological differences in a similar way. The previous finding of using only splits exclusive to one of the trees as a distance can also be applied to the cone distance (Fig. 10b). The approximation can be further improved if the trees are decomposed (results not shown). While the cone distance passes through the consensus of two trees, the *decomposed cone distance* passes through the consensus of every decomposition. This allows for more than one transition point. The resulting distance is an easily computable continuous distance measure on phylogenetic trees.

ACKNOWLEDGMENTS

We thank Karen Vogtmann and Megan Owen for the kind provision of unpublished material, Gregory Ewing and Michael Kopp for helpful comments on the manuscript, and Ingo Ebersberger for providing the metazoa data and discussing the results. Financial support from the Wiener Wissenschafts-, Forschungs- and Technologiefonds (WWTF) is greatly appreciated.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Amenta, N., Godwin, M., Postarnakevich, N., et al. 2007. Approximating geodesic tree distance. *Inf. Process. Lett.* 103, 61–65.

Barthélemy, J.-P., and McMorris, F.R. 1986. The median procedure for n-trees. *J. Classif.* 3, 329–334.

Billera, L.J., Holmes, S.P., and Vogtmann, K. 2001. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27, 733–767.

Bridson, M.R., and Haefliger, A. 1999. *Metric Spaces of Non-Positive Curvature*. Springer, Berlin.

Bryant, D. 2003. A classification of consensus methods for phylogenetics. In: *Bioconsensus*, Janowitz, N.F., Lapointe, F.-J., McMorris, F.R., and Mirkin, B. (eds.). ANS/DIMACS, 2003, 242 pgs.

Dutilh, B.E., van Noort, V., van der Heijden, R.T.J.M., et al. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics* 23, 815–824.

Ebersberger, I. 2007. Human genetic ancestry. Available at: *http://www.newton.can.ac.uk/webseminars/pg+ws/2007/plg/plgw01/0907/ebersberger/*. Accessed May 27, 2008.

Epstein, D., and Ingram, J. 2003. Computing the BHV metric. Available at: *http://www.aimath.org/WWN/geombio/epsteinbhv.pdf*. Accessed May 27, 2008.

Estabrook, G.F., McMorris, F.R., and Meacham, C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst. Zool.* 34, 193–200.

Ewing, G.B., Ebersberger, I., Schmidt, H.A., et al. 2008. Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* 8, 118.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Felsenstein, J. 2005. *PHYLIP (Phylogeny Inference Package), version 3.6*. Department of Genome Sciences, University of Washington, Seattle.

Gadagkar, S.R., Rosenberg, M.S., and Kumar, S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B Mol. Dev. Evol.* 304, 64–74.

Guindon, S., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.

Hein, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* 98, 185–200.

Hillis, D.M., Heath, T.A., and St. John, K. 2005. Analysis and visualization of tree space. *Syst. Biol.* 54, 471–482.

Holmes, S. 2005. Statistical approach to tests involving phylogenies, 91–120. In: Gascuel, O., ed. *Mathematics of Evolution and Phylogeny*. Oxford University Press, New York.

Kuhner, M.K., and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468.

Notredame, C., Higgins, D., and Heringa, J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217.

O'Brien, K.P., Remm, M., and Sonnhammer, E.L.L. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33, D476–D480.

Owen, M. 2007. Computing distances in the space of phylogenetic trees. Preprint. (Cornell University, Ithaca, NY)

Pachter, L., and Sturmfels, B. 2007. The mathematics of phylogenomics. *SIAM Rev.* 49, 3–31.

Pattengale, N.D. 2005. Tools for phylogenetic postprocessing [M.A. thesis]. University of New Mexico.

Robinson, D.F., and Foulds, L.R. 1978. Comparison of weighted labelled trees. *Lect. Notes Math.* 748, 119–126.

Robinson, D.F., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Rokas, A., and Carroll, S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* 22, 1337–1344.

Smythe, A.B., Sanderson, M.J., and Nadler, S.A. 2006. Nematode small subunit phylogeny correlates with alignment parameters. *Syst. Biol.* 55, 972–992.

Stockham, C., Wang, L.-S., and Warnow, T. 2002. Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics* 18, S285–S293.

Vogtmann, K. 2003. Geodesics in the space of trees [Technical report]. Cornell University, Ithaca, NY. Available at: *http://www.math.cornell.edu/~vogtmann/papers/TreeGeodesicss/geodesics07.pdf*. Accessed May 27, 2008.

Waterman, M.S., and Smith, T.F. 1978. On the similarity of dendrograms. *J. Theor. Biol.* 73, 789–800.

Yule, G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. Lond. B* 213, 21–87.

Address reprint requests to:
Dr. Arndt von Haeseler
Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories
Dr. Bohr-Gasse 9/6
A-1030 Vienna, Austria

E-mail: arndt.von.haeseler@univie.ac.at