

Review Paper: The Shape of Phylogenetic Treespace

KATHERINE ST. JOHN*

Department of Mathematics and Computer Science, Lehman College, NY 10034, USA

*Correspondence to be sent to: *Department of Mathematics and Computer Science, Lehman College, and Departments of Anthropology and Computer Science, The Graduate Center, City University of New York; and Division of Invertebrate Zoology, American Museum of Natural History; E-mail: stjohn@lehman.cuny.edu*

Received 20 November 2015; reviews returned 16 December 2015; accepted 22 March 2016

Associate Editor: David Bryant

Abstract.—Trees are a canonical structure for representing evolutionary histories. Many popular criteria used to infer optimal trees are computationally hard, and the number of possible tree shapes grows super-exponentially in the number of taxa. The underlying structure of the spaces of trees yields rich insights that can improve the search for optimal trees, both in accuracy and in running time, and the analysis and visualization of results. We review the past work on analyzing and comparing trees by their shape as well as recent work that incorporates trees with weighted branch lengths. [Maximum likelihood; maximum parsimony; tree metrics; treespace.]

Tree structures have long been used to represent the evolutionary histories of sets of species. For example, the tips of the trees represent extant species and the internal nodes represent speciation events. Despite its simplicity, the tree model captures much of the complexity of the underlying phenomena. However, the sheer number of possibilities forces many simply presented operations on trees to be computationally hard. For example, the maximum parsimony criteria (Farris 1970; Fitch 1971) that seeks the tree with the minimal number of changes across the edges is computationally hard to compute exactly (Foulds and Graham 1982). The addition of weights on the branches, to denote quantities such as amount of evolutionary change, the time, or the confidence on the existence of the branch, adds complexity to the model (Felsenstein 1973; 1978). A popular corresponding optimality criteria for weighted trees, the maximum likelihood criteria, is also computationally hard (Roch 2006).

The power of the tree model comes from the same property that adds the complexity: the vast number of trees to explain different possible evolutionary scenarios. This review focuses on organizing sets of trees, viewed through the lens of improving the efficiency of exact and heuristic algorithms that operate on trees. For each set of n leaves, the set of possible trees can be organized into a space with a distance that delineates neighbors. This rigorous mathematical concept has very practical uses: almost all software tools used to find optimal phylogenetic trees rely on some variant of a local search strategy, where the next tree in the search is chosen from the neighbors of a current tree. Choosing the appropriate metric for the neighbors can greatly simplify searches for optimal trees, turning unsuccessful searches into efficient ones by employing the appropriate metric (e.g. Charleston 1995; Maddison 1991; Urheim et al. 2016, Figure 1). While we touch on the underlying mathematical beauty in these structures and related algorithms, we have omitted many results of mathematical interest and focused on how understanding the underlying space can improve

the search for optimal trees and the analysis of sets of trees.

There are two distinct classes of tree spaces: those that correspond to tree rearrangement metrics and those that correspond to vector-based metrics (Figures 2 and 3). This article first addresses the tree rearrangement metrics. While ignoring edge weights, the tree rearrangement moves and associated metrics are extremely powerful for both weighted and unweighted trees and are included in many software tools (e.g. Goloboff et al. 2008; Guindon et al. 2010; Ronquist et al. 2012; Tamura et al. 2013; Stamatakis et al. 2007). These metrics yield discrete treespaces that can be modeled by undirected graphs. We then address the vector-based metrics that primarily yield continuous treespaces. Some of these metrics allow edge weights to be incorporated seamlessly into the analysis. The use of these latter spaces in phylogenetics is novel and many techniques are being refined, but the ability to compute meaningful statistics makes this a powerful tool for phylogenetic analysis.

DEFINITIONS

This section includes some basic definitions related to trees and metrics. For a more thorough introduction, see Semple and Steel (2003).

Trees

A simple and elegant way to represent the evolutionary relationships between species is with a tree $T = (V, E)$, that is, a connected graph with no cycles (Fig. 3a). The trees can be further decorated to include a root node, representing the hypothetical ancestor of the species under study. When rooted, the branches can be viewed as directed away from the root. Other than the root, all internal (non-leaf) nodes have degree three or higher. If all internal nodes are of degree three, then the tree is called *binary* or *fully resolved*. Nodes of degree

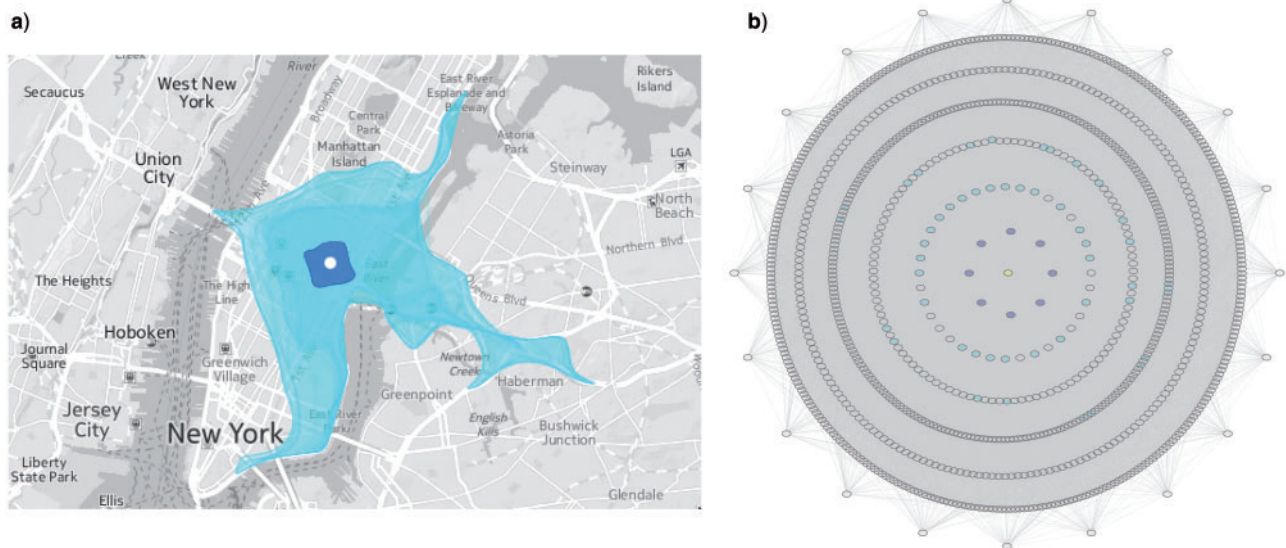


FIGURE 1. a) An analogy to organizing points via different metrics is the points reached in walking 10 minutes (dark shaded regions) versus the points reached by walking or transit in 10 minutes time (light shaded regions). Image generated with Isoscope (Gortana et al. 2014). b) Similarly, an NNI (dark shaded regions) and SPR (light shaded regions) neighborhood of the same point in the 7-leaf treespace.

four or more in a tree are called *polytomies*. A tree with no internal edges is called the *star tree*. For each branch (or edge) of a tree, there is a corresponding *bipartition* or *split* on the set of leaves—namely, the two sets of leaves that would result from removing the edge. Trees can be augmented by assigning lengths to the edges (often representing the amount of evolutionary change across the edge or the confidence in that edge) and are called *weighted* or *continuous* trees (Fig. 3). Every tree induces a set of splits that are pairwise *compatible* (i.e., the splits $A_1|B_1$ and $A_2|B_2$ have at least one of the intersections $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, and $B_1 \cap B_2$ empty). Buneman (1971) showed that if a set of splits is pairwise compatible then it uniquely determines a tree. For each set of n leaves, there are $N = 2^{n-1} - 1$ possible splits of the leaves (all possible ways to partition n objects into 2 nonempty sets). Since trees are acyclic, only a limited number of splits can be present in any given tree (at most $2n - 3$ splits). Each tree can be written as a vector where coordinates correspond to the length of the edges in the tree (Fig. 4). For edges that do not occur in the tree, the corresponding coordinate is set to 0.

The number of possible trees is huge. When there are 4 leaves, {1,2,3,4}, the number of possible trees is the number of ways to group the species into sets of size 2: 12|34, 13|24, and 14|23. For 5 leaves, there are 15 different possible trees (Fig. 5(a)). If the number of leaves of the tree is n , the number of possible tree shapes or “topologies” grows super-exponentially in n . The number of leaf-labeled unrooted trees is $(2n - 5)!! = 1 \cdot 3 \cdot 5 \cdots (2n - 5)$ (Schröder 1870). Similarly, the number of rooted trees is $(2n - 3)!!$.

Given the tremendous number of possible trees as the number of leaves grows large, the organization of these

trees has profound effects on the success of the search for optimal trees and visualization (Hillis et al. 2005). Counting the number of moves needed to transform one tree into another in a search induces a measure of how similar or different trees are. Measures of similarity can also be based on the overlap of the edges (often listed as vectors of all possible edges of the space). Different metrics yield different neighbors and provide a way to adjust the range and depth of the search (Fig. 1). We consider the set of trees for a set S of n taxa, which with a distance metric, forms a *treespace*. A natural coordinate system for points (i.e., trees) in treespace is the splits on S (Fig. 4). When each tree has an optimality score, it is called a *landscape* (Bastert et al. 2002).

Complexity

We give a very brief overview of time complexity; for a thorough treatment, see Cormen et al. (2001). When working with large data sets, the amount of time it takes to compute the answer can often trump the correctness, since if it takes too long to compute the answer exactly, it cannot be used. Complexity refers to the amount of time (or space) needed to compute an answer, often parametrized by the number of inputs. For example, to find the longest branch length in a tree with n leaves, you can examine each edge in turn and store the branch length if is longer than the best seen thus far. This can be accomplished in time proportional to the largest number of edges possible in a tree on n leaves, $2n - 3$ edges. Since $2n - 3$ is a linear function in n , we say this algorithm would run in linear time (in n) or has a worst-case running time complexity of n .

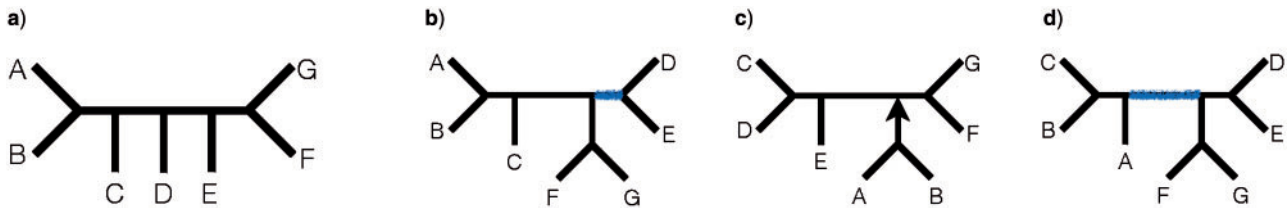


FIGURE 2. Tree rearrangements: a) The starting tree, b) the interchange of neighboring subtrees yields a tree one Nearest Neighbor Interchange (NNI) move away, c) A Subtree Prune and Regraft (SPR) move: the subtree (A, B) is pruned from the initial trees and reattached, and (d) a Tree Bisection and Reconnection (TBR) move: the edge separating ABC from $DEFG$ is bisected and reconnected by a new edge.

(written $O(n)$ and pronounced “big-Oh of n ”). Similarly, if you want to alphabetize the taxon names for n species, there are many algorithms that can accomplish this. A simple one, “bubbleSort,” can order a list of n items in time proportional to n^2 , and thus has worst-case time complexity of $O(n^2)$ (Cormen et al. 2001). If a problem has lower and upper bounds on its running time proportional to $f(n)$, we say it runs in $\theta(f(n))$ (often called “tight bounds” on the running time). All problems that have an algorithm with worst-case running time of $O(n^k)$ for some k are in the class P of problems with polynomial running time. Problems whose solutions can be checked in polynomial time are in the class NP of problems with nondeterministic polynomial time algorithms.

Finding the optimal tree, under the most popular criteria, is *NP-hard* (Foulds and Graham 1982; Roch 2006). That is, one can check quickly, when given a tree, if it has a score better than some bound. However, there is no known polynomial-time algorithm for finding such a tree. Although it is not known if *NP-hard* problems can be solved quickly in polynomial time (this open question has generated much interest and a million dollar prize (Clay Mathematics Institute 2000)). *NP-hardness* is viewed as difficult to compute effectively. *NP-hardness* is usually framed in terms of worst-case instance of complexity, or longest amount of time to solve any instance of the problem. Although it is practical to know the maximal amount of time a problem instance could take, this masks the differences between *NP-hard* problems. A way to capture easy instances of *NP-hard* problems is to identify a parameter that captures the difficulty of the problem. Roughly, the ability to efficiently calculate instances that are small with respect to some parameter is called *fixed parameter tractability* (FPT). For example, although the TBR and SPR tree distances are *NP-hard*, for a fixed distance k on n -leaf trees, they are tractable and can be calculated quickly in n , that is, in $n^c f(k)$ where c is a constant and $f(k)$ is a function that does not depend on n (Allen and Steel 2001; Bordewich and Semple 2004; Bonet and St. John 2010; Whidden et al. 2013).

Optimality criteria

We briefly outline the two most popular optimality criteria (see (Hillis et al. 1996) for a more thorough

treatment of the subject). Given character sequences for a set of species, our goal is to reconstruct the tree that best explains the data.

Maximum Parsimony.—Seeks the most parsimonious tree—the one with the smallest *tree length* or *parsimony score*, which is, roughly, the minimum amount of evolution across the edges of the tree, measured by the sum of character state changes (Farris 1970; Fitch 1971). Although computing a tree length is linear in the number of leaves, the overall problem of finding the most parsimonious tree is *NP-hard* (Foulds and Graham 1982).

Maximum likelihood.—Seeks the tree that is most consistent with the observed data. Given a model of evolution, trees are evaluated by the likelihood that they generated the observed sequences assigned to their leaves. The branch lengths (representing the evolutionary change expected) are used as parameters of the model (Felsenstein 1973; 1978). For a single tree, this calculation, along with estimating the parameters of the model, can be computationally expensive (linear in number of leaves, but with a large constant factor). The overall problem of finding the maximum likelihood tree is *NP-hard* (Roch 2006).

DISCRETE TREESPACES

The treespaces generated from tree rearrangement metrics are often called “discrete treespaces”: they can be modeled by graphs where the trees are vertices and the edges are single, discrete, moves (Fig. 5). As the name suggests, each of these moves rearranges or “edits” the original tree to create a new tree. Although these moves and metrics have the same goal: to compare and organize sets of trees, they do so in very different ways. An analogy is that the treespace is a map (Fig. 1). For the moves defined below, the NNI move is analogous to “walking”, where as the SPR and TBR moves are analogous to “transit and walking”. Starting at the same point, you can go more places if you are allowed to both walk and take transit, over just walking. In this analogy, a neighborhood of a point, under a mode of transportation, is all places you can reach in one time unit. Similarly, the diameter is the greatest distance (measured in unit steps) you can travel in the space,

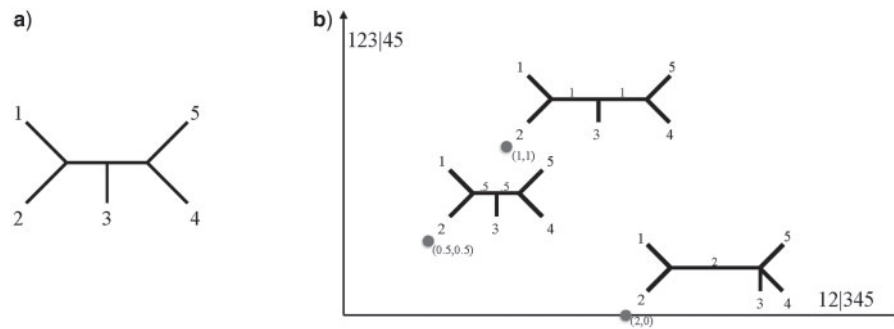


FIGURE 3. a) A tree on 5 leaves. Each edge induces a bipartition or split on the leaves, for example the internal edges induce the splits: 12|345 and 123|45. b) The same tree with branch lengths. In the orthant, the horizontal axis corresponds to the weight of 12|345, and the vertical axis to the weight of 123|45.

and varies under the different modes of transportation. To carry the analogy farther, generalized NNI (Sankoff et al. 1994) and p-ECR (Ganapathy et al. 2003) measures (described below) are similar to bicycling, since each covers similar regions to walking, but can cover more ground, without the “jump” to new NNI neighborhoods found in SPR or TBR. The matching move of Diaconis and Holmes (2002) (described below), which creates random walks that are rapidly mixing, is analogous to “flying”. Although these moves can be computed quickly and are used for searching for optimal trees, their corresponding distance metrics are computationally hard. These tree rearrangement moves are also used to traverse the space of trees with branch lengths.

Metrics and Neighbors

We outline metrics based on tree rearrangements (and will define those based on vectors in “Continuous treespaces” section). We begin with the most common— NNI, SPR, and TBR— and mention some of their variants: generalized NNI and e-PCR. The corresponding *distance* for a given tree rearrangement move is the minimal number of such moves to transform one tree to another. A *neighborhood* of a tree T is all trees within one move of T (or equivalently for tree rearrangement metrics: within distance 1 of T). The *diameter* of a space is the maximal distance between any two trees under the metric.

Nearest Neighbor Interchange.—A nearest neighbor interchange (NNI) swaps subtrees on opposite sides of an internal edge. The distance is the minimal number of moves needed to transform one tree into another (Figs. 2 and 6), and computing it is NP-hard (Li et al. 1996). Although used less for heuristic search, there has been renewed interest since it is embedded into the continuous BHV treespace (see “Continuous treespaces” section). The size of a NNI neighborhood is $2n-6$ (Robinson 1971), the distance is NP-hard to compute, and the diameter of the induced treespace has tight bounds dominated by $n \log_2 n$ (Li et al. 1996).

Subtree Prune and Regraft.—Due to its connection to recombination and hybridization, subtree-prune-and-reconnect (SPR) is used both to analyze phylogenies and in searches of treespace (Hillis et al. 1996). An SPR move between two unrooted trees breaks a subtree from the first tree and reattaches it to an edge of the second tree, contracting resulting vertices of degree two (Fig. 2). Because SPR can differ depending on whether the underlying trees are rooted or not, “rSPR” and “uSPR” are used to refer to SPR on rooted and unrooted trees, respectively. Calculating the rSPR and uSPR distances, the minimal number of moves to transform one tree to another, has been shown to be NP-hard and FPT (Bordewich and Semple 2004; Hickey et al. 2008; Bonet and St. John 2010). Further, there are approximation algorithms that give answers within provable bounds (Bonet et al. 2006; Bordewich et al. 2008; Whidden et al. 2013). Every NNI move is an SPR move. The size of an uSPR neighborhood is $2(n-3)(2n-7)$ (Allen and Steel 2001), where n is the number of leaves. The diameter of the uSPR space is $n - \theta(\sqrt{n})$ (Ding et al. 2011). Song (2003) showed explicit formulas for the size of the rSPR neighborhood (which depends on the shape of the tree) and showed the diameter of the rSPR space satisfies similar bounds to uSPR space.

Generalized NNI (Lazy SPR).—Developed by Sankoff et al. (1994) to traverse treespace more quickly, the move approximates an SPR move by a fixed number of NNI moves (also called “lazy SPR” and used in RAXML Stamatakis et al. 2007). That is, if the number of fixed moves is 5, all trees that can be reached within 5 NNI moves of the starting tree are considered one generalized NNI move from it. These moves have the advantage of the quickness of computing NNI moves but lack the ability of SPR to see more diverse trees quickly.

Tree Bisection and Reconnection (TBR).—A tree bisection and reconnection (TBR) operation removes an edge from a tree and adds a new edge to reconnect the subtrees, contracting resulting vertices of degree two (Fig. 2). The TBR distance between two phylogenetic trees T_1 and T_2 is the minimum number of TBR operations required

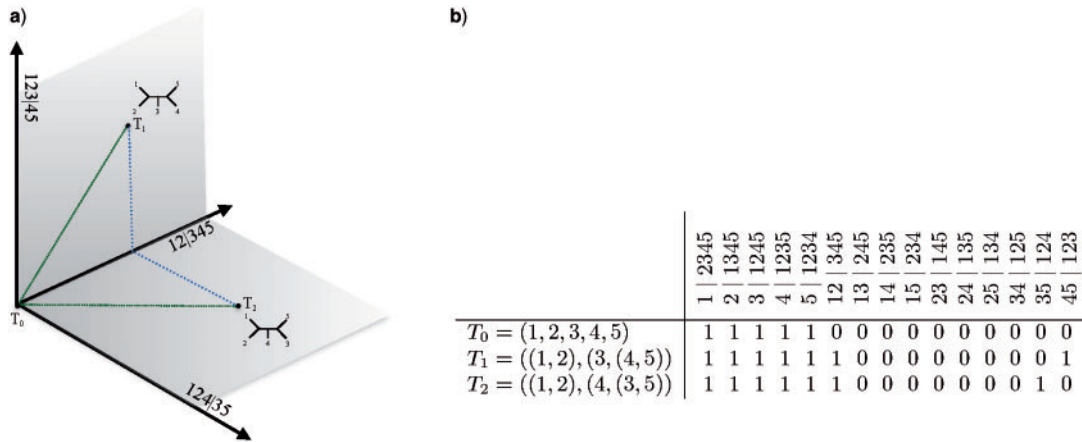


FIGURE 4. a) Representing trees, T_0 , T_1 , and T_2 , as vectors of splits. b) The Robinson–Foulds (Manhattan or L^1) distance is the sum of the pairwise differences which is 2 for all three pairs of these trees. The Branch Distance Score (Euclidean or L^2) is $\sqrt{2}$ for all three pairs of these trees. The BHV metric seeks the shortest path inside the space. For the pairs of trees T_0 and T_1 and T_0 and T_2 , the distance matches the Euclidean distance of $\sqrt{2}$. For the trees, T_1 and T_2 which lie on different orthants, the distance is 2.

to convert T_1 into T_2 . As with SPR, TBR is a popular and effective way to move through treespace when searching for heuristically useful solutions. Calculating TBR distance is NP-hard and fixed parameter tractable (Allen and Steel 2001). Every SPR move is a TBR move. The size of a TBR neighborhood is bounded by $(2n - 3)(n - 3)^2$ (Humphries and Wu 2013). The diameter of the space is $n - \theta(\sqrt{n})$ (Ding et al. 2011).

Combining Neighborhoods.—Several authors have focused on combining the best properties of several different types of neighborhoods. This includes the p-ECR neighborhoods of Ganapathy et al. (2003) which generalizes the NNI operation by allowing p edges to be contracted and then refined. (Goeffon et al. 2008) use “progressive neighborhoods” that evolve as the heuristic search progresses through the landscape.

Exploring Treespace

Each of the moves above can be used to explore the treespace, either as a basis of a random walk or as part of a heuristic search algorithm. The success of the search, both in terms of accuracy and efficiency, depends on the choice of the move since each organizes the search space in a different way.

Walks of Treespace.—Searches for optimal trees are often walks or paths of the space: sequences of trees where each tree in the sequence differs from the previous tree by a single move. For the NNI, SPR, and TBR spaces, there are walks of the treespace that visit every tree exactly once (often called Hamiltonian paths) (Gordon et al. 2013). Given the immense size of the spaces, visiting every node (even just once) requires too much calculation for all but small n . Instead, the space is sampled either by a random walk or by a local search (see below).

Diaconis and Holmes (2002) show a bijection between matchings and rooted, binary phylogenetic trees, and interchange pairs in the matchings to make steps in a random walk. Unlike the metrics above, these steps “mix up” the tree and can be used to explore new regions of the search space. The resulting Markov chains are “rapidly mixing” (roughly, after $\frac{1}{2}n \log n$ moves, the resulting tree is essentially random with respect to the uniform distribution (all trees occurring with equal probability)).

Heuristic Searches.—Many searches follow a local search strategy: start with a tree; at each step, choose a neighbor of the tree; and repeat. The simplest variation is called hill climbing where the best-scoring neighbor is chosen. This greedy approach continues until there are no neighbors that score better or time is exhausted. More sophisticated approaches include pruning of neighborhoods, using multiple starting points, choosing trees with nonoptimal scores with some probability, and dynamically changing parameters such as step length (see Wheeler (2012) for survey). These are not random walks of the space (and do not randomly sample all possibilities), and the size of the neighborhood can have large effects on the computational efficiency of the approaches. For example, although the NNI neighborhood is linear in n and can be computed quite quickly, it can get trapped at local optima. The SPR and TBR neighborhoods (whose size is respectively quadratic and cubic in n) are more difficult to enumerate for larger n but have generally fewer local optima to derail the search (Kirkup and Kim 2000; Money and Whelan 2012; Urheim et al. 2016).

Attraction basins under different metrics.—These treespaces differ in organization and by the distribution of optimal trees with respect to maximum parsimony

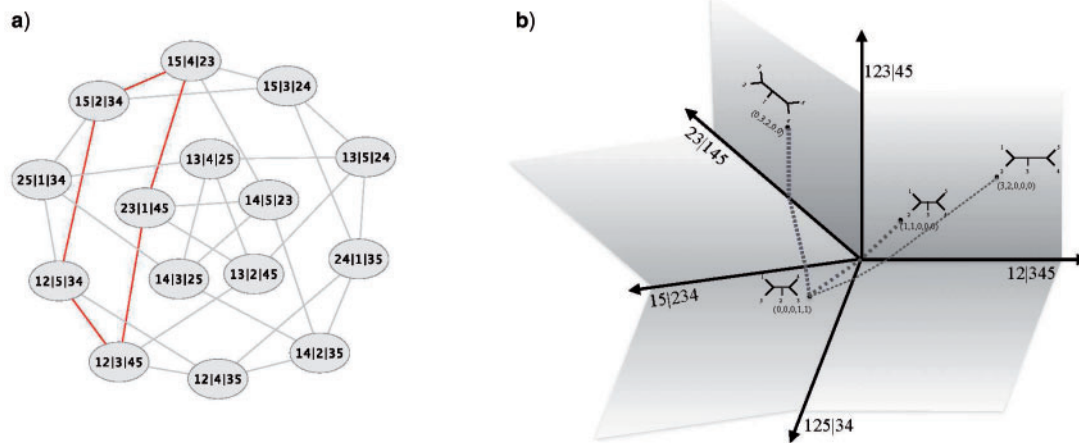


FIGURE 5. a) The NNI treespace of 5-leaf trees. Nodes are labeled using extended split notation: “12|3|45” refers to the tree with splits “12|3|45” and “123|45”. The highlighted circle corresponds to the orthants illustrated in the BHV space for unrooted 5-leaf trees b). The shortest path (geodesic) between trees depends both on the tree shape and the branch lengths. The dashed lines show geodesics that visit auxiliary orthants, whereas the dotted path passes through the origin.

optimality criteria. If we view the treespace as a two-dimensional map, the score can be viewed as the height above sea level. When searching for a maximum scoring point, these regions can be viewed as *phylogenetic islands* (Maddison 1991) that rise above some threshold and *terraces* that are regions where all trees contain a set of fixed subtrees and have the same score (Sanderson et al. 2011; 2015). When searching for a minimum scoring point, these islands are called *attraction basins*: for any optima, T_O , these are all the starting trees that will reach T_O using a greedy hill-climbing strategy. Kirkup and Kim (2000) showed empirically that the NNI treespace has many more attraction basins than the TBR treespace. Urheim et al. (2016) proved that if all the sequences are compatible, then maximum parsimony has a single attraction basin under SPR (and TBR since it extends it), but for NNI, there are terraces where the search will get stuck. Money and Whelan (2012) examined the yeast data set of Rokas et al. (2003) finding empirically similar distributions for maximum likelihood optimality.

While the maximum parsimony problem is NP-hard, there do exist instances where finding the exact answer is possible. Employing branch-and-bound techniques, Hendy and Penny (1982) and Holland et al. (2005) limit the search space by using the current best score to rule out regions. When all the character sequences are compatible, it is easy (i.e. takes linear time) to find this *perfect phylogeny* (Gusfield 1991). Ford et al. (2015) employed this for arbitrary character sequences by partitioning the inputted sequences into compatible subsequences and computing the perfect phylogeny for each. Because the maximum parsimony score is additive (the score for each character can be computed separately and then added together), they showed that the global optimum must exist within a fixed number of steps of these perfect phylogenies. This bounding of the search space works well empirically for data sets

with high consistency index for a tree but eventually devolves to the entire space as the consistency index decreases.

CONTINUOUS TREESPACES

Although many metrics are defined in terms of tree rearrangements, another class of metrics focuses on properties of trees that can be represented as vectors. The metrics are based on comparisons of these vectors. The most common vector representation has the tree’s branch lengths as coordinates. The computations are independent of the order of the coordinates, thus, any fixed order on the coordinates can be used. The resulting spaces are often called “continuous treespaces”. Although there are $N=2^{n-1}-1$ possible splits for n -leaf trees, at most $2n-3$ splits can occur in a tree. Similarly, for a tree vector, we set all coordinates that do not correspond to a split of the tree to 0. Thus, the vector for any tree can have at most $2n-3$ nonzero coordinates. If a tree vector has fewer than $2n-3$ nonzero coordinates, the corresponding tree is not fully resolved (i.e. it is nonbinary). A “star tree” refers to tree with only n branches (e.g. T_0 in Fig. 4).

The tree model becomes more complex when we allow branch lengths on the tree edges, but surprisingly, the metrics become computationally easier. We first review distances that depend solely on comparison of the vectors, and then restrict to spaces where all vectors correspond to a tree. For the latter, the distance between two trees is the shortest path between the two trees that does not leave the space (Fig. 5). (Billera et al. 2001) showed that the *geodesic* or shortest path exists and is unique. We will focus on their space since most statistical and computational tools have been developed for it.

Other vector-based treespaces have been proposed. Some, in particular those that use triples or quartets as

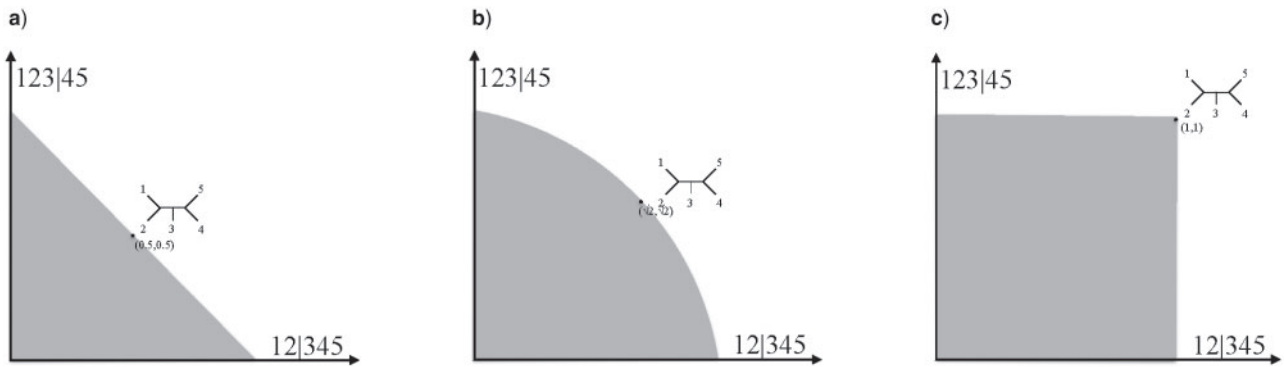


FIGURE 6. The shaded region contains all trees with the splits 12|345 and 123|45 that are within distance 1 of the star tree (origin) under a) L^1 (Robinson–Foulds), b) L^2 (Branch Score Distance), and c) L^∞ (maximum branch) distance.

their coordinates, can be computed quite quickly (Brodal et al. 2013; Sand et al. 2013) and are finding use, especially for comparing gene and species trees (DeGiorgio and Degnan 2010). Another class of intriguing spaces is parametrized by the paths between leaves. Much work is needed for these spaces— both theoretically (such as defining medians and averages when there are multiple shortest paths between points) as well as algorithmic tools (such as algorithms and software that can compute distances for more than 3-leaf trees). Given the huge complexity in computing even small examples and the topology of the underlying space (Moulton and Steel 2004; Gill et al. 2008; Engström et al. 2013), this is a daunting task. We briefly explain these spaces as well as their links to the phylogenetic orange space (defined below) that includes probabilistic models of evolution (Kim 2000).

Metrics and Neighbors

Representing trees as vectors opens up many ways to compare the trees. Much beautiful mathematics already exists on vector spaces, and we highlight here the concepts used for comparing phylogenetic trees (for a more detailed overview, see Rudin 1987). The length (or norm) of a vector \mathbf{v} is often written $\|\mathbf{v}\|$. Some of the metrics used in phylogenetics occur in this framework of norms, often called the p -norm or L^p -norm (named in honor of the mathematician Henri Lebesgue). Rooted triples and quartet metrics can also be represented in terms of vectors, but using underlying vectors that represent instead of the splits, the triples, and quartets, respectively. The Billera–Holmes–Vogtmann (BHV) space of Billera et al. (2001) is also defined as the set of all trees with branch lengths but uses the geodesic, or shortest path between two points that lies completely inside the space, as its metric. Although it can be approximated by metrics that compare vectors, its added requirement that the shortest path lie completely in the space complicates the computation. This requirement also yields midpoints between trees that are trees, allowing summary techniques not possible in other spaces. Unlike the tree rearrangement metrics, many

vector-based metrics used for comparing trees can be computed in polynomial time.

Robinson–Foulds.—The most commonly used distance, the Robinson–Foulds (RF) distance (Robinson 1971), is the sum of the positive difference of branch lengths of the edge set of the trees (often normalized by the number of edges). Although the RF distance was originally defined for tree topologies, it naturally extends to weighted trees (Fig. 4). It can be computed in linear time (Day 1985). It is equivalent to the L^1 or d_1 distances when the coordinates for missing edges are given the value 0 (Fig. 6a). It is often referred to as the *taxicab* or *Manhattan* distance since it would be the distance if you were required to traverse the streets (and not fly over buildings to cut corners). In terms of coordinates, for vectors $\mathbf{p} = (p_1, p_2, \dots, p_N)$ and $\mathbf{q} = (q_1, q_2, \dots, q_N)$, it is:

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^N |p_i - q_i| = |p_1 - q_1| + |p_2 - q_2| + \dots + |p_N - q_N|.$$

Branch score distance.—Kuhner and Felsenstein (1994) proposed a distance that summed the squared differences of branch lengths and then took the square root of this sum (Fig. 6b). When the coordinates for missing edges are given the coordinates 0, this can be viewed as the Euclidean distance or L^2 distance on tree vectors, \mathbf{p} and \mathbf{q} :

$$d_2(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2 = \left(\sum_{i=1}^N |p_i - q_i|^2 \right)^{1/2} = \sqrt{|p_1 - q_1|^2 + |p_2 - q_2|^2 + \dots + |p_N - q_N|^2}.$$

L^p and L^∞ distances.—This pattern can be continued, and an associated distance can be defined for any $p > 0$. L^p

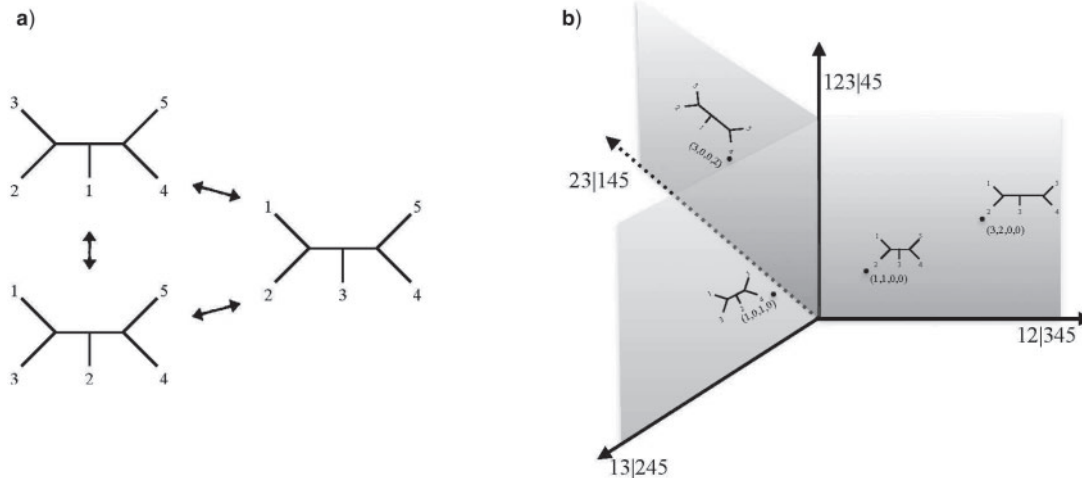


FIGURE 7. a) Three 5-leaf trees that differ by a single NNI moves (arrows). b) The same tree shapes represented in the continuous treespace. Each orthant contains all trees with the same underlying topology.

distance is a generalization:

$$d_p(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_p = \left(\sum_{i=1}^N |p_i - q_i|^p \right)^{1/p}$$

$$= (|p_1 - q_1|^p + |p_2 - q_2|^p + \dots + |p_N - q_N|^p)^{1/p}.$$

The L^∞ norm takes this concept to the limit to get:

$$d_\infty(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_\infty = \max_{i=1, \dots, N} |p_i - q_i|$$

That is, the L^∞ distance is the maximum difference between corresponding coordinates (Figure 6c).

BHV distance.—Billera et al. (2001) view weighted trees as vectors of their split weights yielding an $(2n-3)$ -dimensional space embedded inside the larger $(2^{n-1}-1)$ -dimensional space of all graphs. The distance between two trees is the geodesic, or shortest path, inside treespace (Fig. 5). This continuous treespace easily handles weighted edges, provides a rigorous environment to average trees (Billera et al. 2001), and its metric, the BHV distance, is polynomial-time (Owen and Provan 2011) and can be approximated in linear time (Amenta et al. 2007). The complexity of computing the metric was open almost a decade and is surprisingly $O(n^4)$ via a clever encoding as a network flow problem on bipartite graphs. Each tree shape corresponds to an *orthant*: a copy of \mathbb{R}_+^{2n-3} with each coordinate the length of the edge in the tree (Fig. 7).

Rooted triples.—The rooted triples distance counts the number of triples that occur in only one of the input trees (Critchlow et al. 1996). Like the RF distance, it can be viewed as a L^1 distance on vectors. Here, the vectors are all possible rooted triples on n leaves (Fig. 8a). It can be computed in $O(n \log n)$ (Brodal et al. 2013; Sand et al. 2013). This can be extended to include weighted

branches. The added structure of considering triples makes it useful for estimating species trees (DeGiorgio and Degnan 2010).

Quartet distance.—The quartet distance counts the number of quartets that occur in only one of the input trees. Like the Robinson-Foulds distance, it can be viewed as a L^1 distance on vectors. Here, the vectors are all possible quartets on n leaves (Fig. 8b). It can be computed in $O(n \log n)$ time (Brodal et al. 2004; Sand et al. 2013). This can be extended to include weighted branches.

Path-Distance Spaces.—Although the above distances look at sets of leaves (e.g., bipartitions of all the leaves, triplets of leaves, and quartets of leaves), we can also use the distance between leaves, induced by each tree. Given a tree, T , the corresponding *dissimilarity matrix* or *tree metric*, $d_T: \mathcal{L} \rightarrow \mathbb{R}_+$, is defined, for any two leaves $x, y \in \mathcal{L}$:

$$d_T(x, y) = \sum_{e \in P_T(x, y)} w(e),$$

where $P_T(x, y)$ is the path of edges between the leaves x and y in the tree, T , and $w(e)$ is the weight of edge e (see Hillis et al. (1996)). Buneman (1971) gave a simple and elegant condition (the “4-point condition”) to test when a dissimilarity matrix corresponds to a tree. Distance-based methods such as neighbor joining (Saitou and Nei 1987) take these matrices and estimate a tree that matches the observed distances. The set of dissimilarity matrices that correspond to a tree form a space, with the distance defined as the shortest path in the space (Bandelt and Dress 1986; Moulton and Steel 1999; 2004). Given a weighted tree, T , its (additive) path distance can be represented as a vector of distances between any pair of leaves. For example, an unrooted 5-leaf tree has $(5 \cdot 4)/2 = 10$ coordinates. The tree T_1 from Figure 4 has coordinates (2, 3, 4, 4, 3, 4, 4, 3, 3, 2) under the additive-path distance.

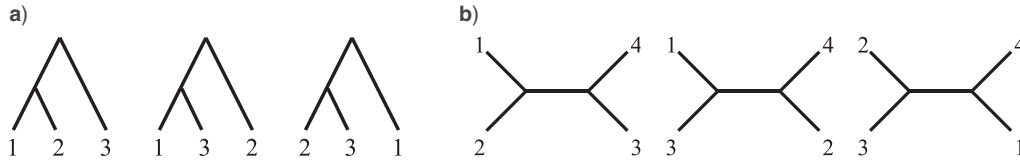


FIGURE 8. a) The three possible rooted triples on leaves $\{1,2,3\}$ and b) the three possible quartets on $\{1,2,3,4\}$.

The space of all matrices (including those that do not correspond to trees) for n taxa is the $\frac{n(n-1)}{2}$ -dimensional space, $\mathbb{R}_+^{n(n-1)/2}$. This is the space of inputs to distance-based reconstruction methods such as Neighbor Joining (Sautou and Nei 1987). Restricting to matrices for which the 4-point conjecture holds yields a smaller subspace where each point corresponds to a weighted tree (Bandelt and Dress 1986; Moulton and Steel 1999). The points with the same underlying tree space are called “cones.” Due to the restriction to tree metrics, there is a natural correspondence between the orthants in BHV space and the cones in the space of dissimilarity matrices (just as there is a natural correspondence between the points that represent trees in the discrete NNI, SPR, and TBR treespaces), but the details of how the distances correspond between the spaces have not been determined. Despite its complex construction, the BHV space has unique shortest paths (geodesics) between points (Billera et al. 2001). For the path distance spaces, there can be multiple shortest paths.

A related treespace can be created by multiplying edge weights, instead of adding them, to yield a $\left(\frac{n(n-1)}{2}\right)$ -vector for each n -leaf tree. These edge-product vectors have as coordinates the products of the exponential of the negative of the weights of the edges of the path. That is, for leaves x and y in a tree T :

$$d_{Tp}(x,y) = \prod_{e \in P(x,y)} e^{-w(e)} = e^{-\sum_{e \in P(x,y)} w(e)}.$$

The edge product vector can be easily computed from the additive path vector. For example, for T_1 , we have the vector $(e^{-2}, e^{-3}, e^{-4}, e^{-4}, e^{-3}, e^{-4}, e^{-4}, e^{-3}, e^{-3}, e^{-2})$. The latter are the points of edge-product space of Moulton and Steel (2004). The space has nice mathematical properties but lacks unique geodesics and is difficult to visualize for even small trees (Figure 1 of Engström et al. (2013) illustrates the curved subspace corresponding to a rooted 3-leaf tree). As Moulton and Steel (2004) and Gill et al. (2008) note, this space is related to the “phylogenetic orange” space of Kim (2000). In the orange space, the points are probability distributions on the possible leaf labelings or site patterns. That is, for a fixed number of leaves n and r number of possible character states, there are r^n possible labelings of the leaves. These are used to form the coordinates of the vectors with the restriction that the coordinate values of these labels sum to 1. For example, a possible leaf labeling of T_2 is leaf 1 is A, leaf 2 is A, leaf 3 is C, leaf 4 is G and leaf 5 is T, or “AACGT.” The sequence of leaf labelings,

$\{AACGT, CACGT, AACGT, TACGA\}$ would correspond to the vector $(0.5, 0.25, 0.25, 0, 0, 0, \dots, 0)$ assuming that AACGT, CACGT, and TACGA are the first three coordinates. Note that AACGT occurs twice, whereas the other two labelings occur once, and the coordinate values represent the fraction of time each occurs and thus sum to 1. Given a weighted tree (T,w) , we can define transition rates for each edge to be $\lambda(e) = e^{-w(e)}$. The Markov process parametrized by the pair (T,λ) induces a joint probability distribution on the leaf labelings, giving a correspondence between points in the edge product space and points in phylogenetic orange space. In the orange space, two trees are assigned the same point if they generate the leaf labelings with the same probabilities (since the probabilities are exactly the coordinates of the points). More theoretical and algorithm advances are needed to compute distances and simple statistics such as averages.

Summary and Consensus Methods

Due to their newness, most continuous spaces lack the theoretical and algorithmic tools to compute distances efficiently, making it not yet possible to compute summaries and consensus methods. As such, this section focuses on the spaces that employ edge weights as their coordinates.

The strength of the BHV continuous space is the consensus and summary statistics that yield resolved weighted trees. Many analyses use sets of trees, and methods that can capture the important features succinctly are valuable. Commonly used methods like strict consensus and majority rule consensus are fast to compute but ignore branch lengths and often return unresolved trees (Schuh and Polhemus 1980; Margush and McMorris 1981; McMorris et al. 1983; Amenta et al. 2003). This leads to situations where the summary contains no edges (particularly troublesome for the strict consensus tree that only contains an edge if it occur in all of the input trees). Interestingly, the weighted version of the majority rule consensus (which by construction is always a tree, albeit often unresolved) is the median under the L^1 (weighted RF) metric. The traditional Euclidean mean, when applied to tree vectors, can yield vectors that do not correspond to trees. The BHV space with its requirement that the distance be the shortest path in the space gives a promising way to “average” sets of trees that captures the contributions of all the inputted trees.

In this framework, the majority rule tree, is the mode, and the Fréchet mean plays the role of the average. The Fréchet mean is the tree that minimizes the sum of the squared BHV distances to the set of trees, $\mathcal{T} = \{T_1, \dots, T_m\}$:

$$T_\mu = \min_T \sum_{i=1}^m d_{BHV}(T_i, T)^2.$$

The property that guarantees that the geodesics are unique (that the space is nonpositively curved (Billera et al. 2001)) gives also that the Fréchet mean is unique. Further, an analog to the Law of Large Numbers holds (Sturm 2003) yielding an iterative algorithm for approximating the mean. The Fréchet mean exhibits unexpected non-Euclidean behavior: Miller et al. (2015) showed that the mean is “sticky”: perturbing a tree does not always change the mean, unlike in Euclidean space. This often occurs when the mean is on a lower dimensional orthant (i.e., the mean tree contains polytomies), and may explain why other summary methods, such as the majority-rule consensus tree (McMorris et al. 1983), often give degenerate trees. Independently, Bačák (2012) and Benner et al. (2014) gave algorithms for computing the median of a set of trees, which is the tree minimizing the sum of distances to those trees (opposed to squared distances for the Fréchet mean). Being a robust estimator, the median is even more sticky than the mean in tree space.

More statistical tools are under development. These include best-fit geodesics or one-dimensional approximations of the data can be computed using stochastic optimization (Nye 2011; 2014). Approximate mean hypothesis testing and approximate linear discriminant analysis have also been developed (Feragen et al. 2013). Additionally, there are a variety of methods for statistical analysis based on the BHV distance, including measures of the intrinsic curvature of the data (Chakerian and Holmes 2012; Cleary et al. 2014a, 2014b). Recently, Nye (2015) showed that Brownian motion on treespace can be approximated by a random walk, giving a promising way to sample the space, since computing a distribution under Brownian motion directly for $n > 5$ is extremely challenging.

Optimality Criteria on Continuous Treespace

Although the interplay of metrics and optimality criteria has been explored for the discrete treespace, less is known for the continuous treespace. Because branch lengths are part of the maximum likelihood paradigm, it makes sense to compare and analyze trees including this information, in addition to using the topology. This is also motivated by the fact there can be multiple local optima for a fixed tree topology (Steel 1994). In terms of computing maximum likelihood scores, this implies that the continuous counterpart of hill-climbing, gradient descent, does not work for computing maximum likelihood scores, even in a single orthant.

Chor et al. (2000) extended the Steel example to give ‘level curves’ of branch lengths that are local optima. There has been initial work on visualizing the search paths of continuous trees (e.g. Hillis et al. 2005; Park et al. 2010; Whidden and Matsen 2015) but these visualizations use mapping of the discrete space even for weighted trees.

CONCLUSION

As we seek optimal trees for biological data and ways to understand the results, the underlying treespaces chosen for these searches and analysis are an important aspect of their success. When searching spaces using discrete moves, the SPR moves seems most effective from both theoretical and empirical results (Kirkup and Kim 2000; Urheim et al. 2016). The interplay of metrics with optimality criteria has a large effect on the difficulty of the search. Often searches are done without first carefully examining the data. As the number of taxa grows, the number of tree shapes grows super-exponentially and even simple “pre-processing” of the character sequences can have large effects on the size of the search space, time and accuracy (Charleston 1995; Holland et al. 2005; Money and Whelan 2012; Ford et al. 2015). When branch lengths are used, the BHV treespace with efficient metrics and well-defined statistical methods seems the most effective at analyzing search results. For both discrete and continuous treespaces, better understanding of the underlying structure can improve the search for optima and the analysis of output.

ACKNOWLEDGMENTS

The author would like to thank the organizing committee of the Mathematical and Computational Evolutionary Biology 2015 conference, to Olivier Gascuel and David Bryant for inviting this article and editing this volume, and to editor Frank Anderson for his suggestions. She would also like to thank Sean Cleary, Eric Ford, Megan Owen, Ella Pavlechko, and Ward Wheeler for insightful conversations and helpful comments.

FUNDING

Partial funding for this work was provided by the Simons Foundation (316124).

REFERENCES

- Allen B., Steel M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Combinatorics* 5:1–13.
- Amenta N., Clarke F., St. John K. 2003. A linear-time majority tree. In: *Lecture Notes in Bioinformatics (subseries of Lecture Notes in Computer Science) Third International Workshop, WABI 2003 (Workshop on Algorithms in Biology)*, Budapest, Hungary, volume 2812, p. 216–227.

- Amenta N., Godwin M., Postarnakevich N., St. John, K. 2007. Approximating geodesic tree distance. *Informat. Process. Letters* 103(2):61–65.
- Bačák M. 2012. A novel algorithm for computing the Fréchet mean in Hadamard spaces. arXiv 1210.2145v1.
- Bandelt H.-J., Dress A. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.* 7(3):309–343.
- Bastert O., Rockmore D., Stadler P.F., Tinhofer G. 2002. Landscapes on spaces of trees. *Appl. Math. Comput.* 131(2-3):439–459.
- Benner P., Bačák M., Bourguignon P.-Y. 2014. Point estimates in phylogenetic reconstructions. *Bioinformatics* 30(17):i534–i540.
- Billera L., Holmes S., Vogtmann K. 2001. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27:733–767.
- Bonet M.L., St. John K. 2010. On the complexity of uSPR distance. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 7(3):572–576.
- Bonet M.L., St. John K., Mahindru R., Amenta N. 2006. Approximating subtree distances between phylogenies. *J. Comput. Biol.* 13(8):1419–1434.
- Bordewich M., McCartin C., Semple C. 2008. A 3-approximation algorithm for the subtree distance between phylogenies. *J. Discrete Algorithms* 6(3):458–471.
- Bordewich M., Semple C. 2004. On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combintorics* 8:409–423.
- Brodal G.S., Fagerberg R., Mailund T., Pedersen C.N., Sand A. 2013. Efficient algorithms for computing the triplet and quartet distance between trees of arbitrary degree. In: *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. p. 1814–1832.
- Brodal G.S., Fagerberg R., Pedersen C.N. 2004. Computing the quartet distance between evolutionary trees in time $O(n \log n)$. *Algorithmica*, 38(2):377–395.
- Buneman P. 1971. The recovery of trees from measure of dissimilarity. In: Hodson F.R., Kendall D.G., Tautau P., editors. *Mathematics and the archeological and historical sciences*. Edinburgh: Edinburgh University Press. p. 387–395.
- Chakerian J., Holmes S. 2012. Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comput. Graph. Stat.* 21(3):581–599.
- Charleston M.A. 1995. Toward a characterization of landscapes of combinatorial optimization problems, with special attention to the phylogeny problem. *J. Comput. Biol.* 2(3):439–450.
- Chor B., Hendy M.D., Holland B.R., Penny D. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Biol. Evol.* 17(10):1529–1541.
- Clay Mathematics Institute 2000. The Millennium Prize Problems: P vs NP problem. <http://www.claymath.org/millennium/>.
- Cleary S., Feragen A., Owen M., Vargas D. 2014a. Multiple principal components analysis in tree space. In: *Asymptotic Statistics on Stratified Spaces*, Volume 44. Oberwolfach Report. p. 33–36.
- Cleary S., Feragen A., Owen M., Vargas D. 2014b. On tree-space principal component analysis. In: *Asymptotic Statistics on Stratified Spaces*, Volume 44. Oberwolfach Report. p. 11–15.
- Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. 2001. *Introduction to Algorithms*. 2nd ed. Cambridge, MA: MIT Press.
- Critchlow D.E., Pearl D.K., Qian C. 1996. The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.* 45(3):323–334.
- Day W. 1985. Optimal algorithms for comparing trees with labeled leaves. *J. Classification* 2:7–28.
- DeGiorgio M., Degnan J.H. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol. Biol. Evol.* 27(3):552–569.
- Diaconis P., Holmes S.P. 2002. Random walks on trees and matchings. *Electron. J. Probab.* 7:2002.
- Ding Y., Grünewald S., Humphries P.J. 2011. On agreement forests. *J. Comb. Theory, Ser. A* 118(7):2059–2065.
- Engström A., Hersh P., Sturmfels B. 2013. Toric cubes. *Rendiconti del Circolo Matematico di Palermo* 62(1):67–78.
- Farris J.S. 1970. Methods for computing Wagner trees. *Syst. Zool.* 19:83–92.
- Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22:240–249.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.* 27(4):401–410.
- Feragen A., Owen M., Petersen J., Wille M.M., Thomsen L.H., Dirksen A., de Bruijne M. 2013. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In: *Information Processing in Medical Imaging*. Springer. p. 74–85.
- Fitch W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* 20:406–416.
- Ford E., St. John K., Wheeler W.C. 2015. Towards improving searches for optimal phylogenies. *Syst. Biol.* 64(1):56–65.
- Foulds L.R., Graham R.L. 1982. The Steiner problem in phylogeny is NP-complete. *Adv. Appl. Math.* 3(1):43–49.
- Ganapathy G., Ramachandran V., Warnow T. 2003. Better hill-climbing searches for parsimony. In: *Proceedings of the Third International Workshop on Algorithms in Bioinformatics (WABI)*, p. 245–258.
- Gill J., Linusson S., Moulton V., Steel M. 2008. A regular decomposition of the edge-product space of phylogenetic trees. *Adv. Appl. Math.* 41(2):158–176.
- Goeffon A., Richer J.-M., Hao J.-K. 2008. Progressive tree neighborhood applied to the maximum parsimony problem. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 5(1):136–145.
- Goloboff P.A., Farris J.S., Nixon K.C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* 24:774–786.
- Gordon K., Ford E., St. John K. 2013. Hamiltonian walks of phylogenetic treespaces. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 10(4):1076–1079.
- Gortana F., Kaim S., von Lupin M. 2014. Isoscope: Exploring mobility. “Urbane Ebenen: Mobilität” class project at the University of Applied Sciences, Potsdam. <http://isoscope.martinvonlupin.de>.
- Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59(3):307–321.
- Gusfield D. 1991. Efficient algorithms for inferring evolutionary trees. *Networks* 21(1):19–28.
- Hendy M., Penny D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* 59(2):277–290.
- Hickey G., Dehne F., Rau-Chaplin A., Blouin C. 2008. SPR distance computation for unrooted trees. *Evol. Bioinformatics* 4:17–27.
- Hillis D.M., Heath T., St. John K. 2005. Analysis and visualization of tree space. *Syst. Biol.* 3:471–482.
- Hillis D.M., Mable B.K., Moritz C. 1996. *Molecular systematics*. Sunderland, Mass: Sinauer Associates.
- Holland B., Huber K., Penny D., Moulton V. 2005. The minmax squeeze: Guaranteeing a minimal tree for population data. *Mol. Biol. Evol.* 22(2):235–242.
- Humphries P.J., Wu T. 2013. On the neighborhoods of trees. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 10(3):721–728.
- Kim J. 2000. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol. Phylogenet. Evol.* 17(1):58–75.
- Kirkup B., Kim J. 2000. From rolling hills to jagged mountains: scaling of heuristic searches for phylogenetic estimation. *Mol. Biol. Evol.* (In revision).
- Kuhner M.K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11(3):459–468.
- Li M., Tromp J., Zhang L. 1996. Some notes on the nearest neighbour interchange distance. In: *COCOON '96: Proceedings of the Second Annual International Conference on Computing and Combinatorics*. London, UK: Springer. p. 343–351.
- Maddison D.R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool.* 40(3):315–328.
- Margush T., McMorris F. 1981. Consensus n-trees. *Bull. Math. Biol.* 43:239–244.
- McMorris F., Meronk D., Neumann D. 1983. A view of some consensus methods for trees. In: *Numerical Taxonomy: Proceedings of the NATO Advanced Study Institute on Numerical Taxonomy*. Berlin: Springer-Verlag.
- Miller E., Owen M., Provan J.S. 2015. Polyhedral computational geometry for averaging metric phylogenetic trees. *Adv. Appl. Math.* 68:51–91.
- Money D., Whelan S. 2012. Characterizing the phylogenetic tree-search problem. *Syst. Biol.* 61(2):228–239.

- Moulton V., Steel M. 1999. Retractions of finite distance functions onto tree metrics. *Disc. Appl. Math.* 91(1):215–233.
- Moulton V., Steel M. 2004. Peeling phylogenetic ‘oranges’. *Adv. Appl. Math.* 33(4):710–727.
- Nye T.M. 2014. An algorithm for constructing principal geodesics in phylogenetic treespace. *IEEE Trans. Comput. Biol. Bioinformatics* 11(2):304–315.
- Nye T.M. 2015. Convergence of random walks to brownian motion in phylogenetic tree-space. arXiv preprint arXiv:1508.02906.
- Nye T.M.W. 2011. Principal components analysis in the space of phylogenetic trees. *Ann. Stat.* 39(5):2716–2739.
- Owen M., Provan J.S. 2011. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8:2–13.
- Park H.J., Sul S.-J., Williams T.L. 2010. Large-scale analysis of phylogenetic search behavior. In *Advances in Experimental Medicine and Biology*. Springer. p. 35–42.
- Robinson D. 1971. Comparison of labeled trees with valency three. *J. Combinatorial Theory B* 11:105–119.
- Roch S. 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 3(1):92–94.
- Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
- Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61(3):539–542.
- Rudin W. 1987. *Real and complex analysis*. New York, NY: McGraw Hill.
- Sand A., Holt M.K., Johansen J., Fagerberg R., Brodal G.S., Pedersen, C.N., Mailund T. 2013. Algorithms for computing the triplet and quartet distances for binary and general trees. *Biology* 2(4): 1189–1209.
- Sanderson M.J., McMahon M.M., Stamatakis A., Zwickl D.J., Steel M. 2015. Impacts of terraces on phylogenetic inference. *Syst. Biol.* syv024.
- Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. *Science* 333:448–450.
- Sankoff D., Abel Y., Hein J. 1994. A tree: a window: a hill; generalization of nearest-neighbor interchange in phylogenetic optimization. *J. Class.* 11(2):209–232.
- Sautou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
- Schröder E. 1870. Vier combinatorische probleme. *Z. Math. Phys.* 15:361–376.
- Schuh R.T., Polhemus J.T. 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the leptoDOMORPHA (hemiptera). *Syst. Zool.* 29:1–26.
- Semple C. and M. Steel. 2003. *Phylogenetics*, Volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford: Oxford University Press.
- Song Y.S. 2003. On the combinatorics of rooted binary phylogenetic trees. *Ann. Combinatorics* 7(3):365–379.
- Stamatakis A., Blagojevic F., Antonopoulos C.D., Nikolopoulos D.S. 2007. Exploring new search algorithms and hardware for phylogenetics: RAXML meets the IBM Cell. *J. VLSI Signal Process. Syst.* 48(3):271–286.
- Steel M.A. 1994. The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biol.* 43(4):560–564.
- Sturm K.-T. 2003. Probability measures on metric spaces of nonpositive curvature. In: *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, Volume 338 of *Contemporary Mathematics*. American Mathematical Society, Providence, RI. p. 357–390.
- Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30(12):2725–2729.
- Urheim E., Ford E., St. John K. 2016. Characterizing local optima for maximum parsimony. *Bull. Math. Biol.* 78(5):1058–1075. (On-line publication: 30 May 2016).
- Wheeler W.C. 2012. *Systematics: a course of lectures*. Oxford, UK: Wiley-Blackwell.
- Whidden C., Beiko R.G., Zeh N. 2013. Fixed-parameter algorithms for maximum agreement forests. *SIAM J. Comput.* 42(4):1431–1466.
- Whidden C., Matsen F.A. 2015. Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.* syv006.