Evolutionary Trees and Ordinal Assertions*

Paul E. Kearney[1], Ryan B. Hayward[2] and Henk Meijer[3]

[1] Department of Computer Science, University of Toronto,
Toronto ON, CANADA M5S 1A4
`kearney@cs.toronto.edu`
[2] Department of Mathematics and Computer Science, University of Lethbridge,
4401 University Drive West, Lethbridge AB, CANADA T1K 3M4
`hayward@cs.uleth.ca`
[3] Department of Computing and Information Science, Queen's University,
Kingston ON, CANADA K7L 3N6
`henk@qucis.queensu.ca`

keywords: distance matrix, evolution, ordinal, phylogeny, species, tree

Abstract. An evolutionary tree, or *phylogeny*, is a tree without degree two vertices whose vertices are labelled by species. An *ordinal assertion*, in an evolutionary context, is a statement of the form "species a and b are as similar as species c and d" and is supported by a distance matrix M of evolutionary data if $M[a,b] \leq M[c,d]$. Many phylogenetic inference methods commit to the assumption that the observed evolutionary dissimilarity $M[s,t]$ approximates the true evolutionary dissimilarity $d_T(s,t)$. However, when M is derived from sequence comparison this assumption can lead to erroneous inferences. A more plausible assumption is that ordinal assumptions supported by the true phylogeny T are also supported by the observed evolutionary dissimilarities M.

A phylogeny T is an *ordinal representation* of the distance matrix M if for all species a, b, c and d labelling T

$$d_T(a,b) \leq d_T(c,d) \text{ if and only if } M[a,b] \leq M[c,d].$$

Ordinal representations are interesting from an evolutionary perspective since they represent ordinal assertions supported by observed data. Ordinal representations are equally interesting from a combinatoric perspective as their analysis is rich in structure. In this paper the applicability of ordinal assertions to phylogenetic inference is discussed and the following results are presented:

- The ordinal representation of a distance matrix is unique. This generalizes the well-known result that no two phylogenies share the same distance matrix [10, 20].
- The ordinal representation of a distance matrix can be found in $O(n^2 log^2(n))$ time. The algorithm presented improves upon an $O(n^3)$ algorithm by Kannan and Warnow [13] that finds binary ordinal representations of distance matrices.
- Under certain conditions weighted ordinal representations can be found in polynomial time.

1 The Assumption of Ordinality

The inference of evolutionary trees, or *phylogenies*, from distance data is a challenging problem both from a computational and biological perspective. Inference methods are often computationally intractable [5, 6] or without clear biological foundation (see [19, 8] for an overview of phylogenetic inference methods). Typically distance data takes the form of a distance matrix M indexed by species

---

* The authors gratefully acknowledge the support of NSERC.

set S such that for each pair of species s and t, $M[s,t]$ is derived by comparing sequence data for s and t. The goal is to infer a phylogeny T labelled by S that is representative of the evolutionary data supported by M.

Definition 1. A *phylogeny T labelled by species set S* is a tree without degree[4] two vertices such that each leaf of T is identified with a unique element of S. T is a *partial* phylogeny if T is permitted vertices of degree two. Since vertices of evolutionary trees represent points of speciation, evolutionary trees do not have degree two vertices.

If the edges of T are assigned positive lengths then T is called *weighted* and $d_T(s,t)$ is the sum of the lengths along the path $P_T(s,t)$ in T from s to t. Otherwise, T is called *unweighted* and $d_T(s,t)$ is the number of edges on the path $P_T(s,t)$. At times d_T is referred to as the *distance matrix of* T. Unless otherwise stated, phylogenies are unweighted and not partial.

If a distance matrix is exactly the distance matrix of some phylogeny then it is called *additive*. Most inference methods such as neighborhood joining [16], Fitch–Margoliash [9, 3], Sattath and Tversky [17], UPGMA [18], and split decomposition [1] are based upon the assumption that the observed evolutionary dissimilarities (M) approximate the true evolutionary dissimilarities (d_T where T is the historically–correct evolutionary tree). We call this the *assumption of additivity*. However, $M[s,t]$ tends to underestimate $d_T(s,t)$ since sequence comparisons give no information about the number of mutations that have occurred at a given site during the evolution of the sequences of s and t. For example, two sequences may differ at a particular site but that difference may be the result of several mutations of the common ancestral site. In an attempt to compensate for this discrepancy *correction transforms* can be applied to M. For example, under the assumption that sequences evolve according to the Jukes–Cantor model [12], the following transformation from the percentage dissimilarity matrix M to a corrected distance matrix M' accounts for anticipated discrepancy:

$$M'[s,t] = -\frac{3}{4} ln(1 - \frac{4}{3} M[s,t]).$$

Nevertheless, there is evidence that inference methods such as neighbor joining and UPGMA perform poorly in simulations even when corrected distances are used [11]. This brings into question the appropriateness of the assumption of additivity.

Recently alternative inference methods using *ordinal assertions* have been examined [15, 14, 13]. An ordinal assertion in an evolutionary context is a statement of the form "species a and b are as similar as species c and d" and is supported by M (supported by T) if $M[a,b] \leq M[c,d]$ (if $d_T(a,b) \leq d_T(c,d)$). Replacing the assumption of additivity is the *assumption of ordinality* which claims that ordinal assertions supported by T where T if the historically–correct phylogeny, are also supported by M. That is, for all species a, b, c and d

$$d_T(a,b) \leq d_T(c,d) \text{ if and only if } M[a,b] \leq M[c,d]. \tag{1}$$

We refer to a (weighted) phylogeny T as a *(weighted) ordinal representation* of M if T and M satisfy (1).

Notice that correction transforms for evolutionary models such as Jukes–Cantor [12] and Felsenstein's [7] implicitly commit to the assumption of ordinality as these correction transforms are monotonic. That is, M and M' are *ordinally equivalent*: $M[a,b] \leq M[c,d]$ if and only if $M'[a,b] \leq M'[c,d]$

---

[4] The degree of a vertex v is the number of vertices adjacent to v.

for all species a, b, c and d. In this sense the assumption of ordinality is prior to the assumption of additivity.

This paper is an investigation into algorithmic and combinatoric properties of ordinal representations. In particular the following results are presented:

- The ordinal representation of a distance matrix is unique. This generalizes the well–known result that no two phylogenies share the same distance matrix [10, 20].
- The ordinal representation of a distance matrix can be found in $O(n^2 log^2(n))$ time. The algorithm presented improves upon an $O(n^3)$ algorithm by Kannan and Warnow [13] that finds binary ordinal representations of distance matrices.
- We discuss how the algorithm for finding ordinal representations can be extended to a polynomial algorithm for finding weighted ordinal representations under certain conditions.

The remainder of the paper is organized as follows. In section 2 *midpath phylogenies* are introduced which are precursors to ordinal representations. In section 3 an algorithm is presented that extends midpath phylogenies to ordinal representations. As a by–product of the algorithm it is shown that ordinal representations are unique (section 3.5). Finally, in section 4 it is shown how weighted ordinal representations can be found under certain conditions.

2 Midpath Phylogenies: A Starting Point

Midpath phylogenies were introduced in [14] as representations of evolutionary data. Here it is shown that an ordinal representation is a midpath phylogeny and, in fact, is an expansion of the *minimal* midpath phylogeny. Hence our understanding of ordinal representations begins with an articulation of these concepts.

Let T_R be an ordinal representation of M, possibly weighted, where M is indexed by S. Let $x, y \in S$ and let w be the edge or vertex of $P_{T_R}(x, y)$ where the midpoint of $P_{T_R}(x, y)$ is located. Observe that $T_R - \{w\}$ satisfies the following *trichotomy*:

- The partial phylogeny of $T_R - \{w\}$ containing x is labelled by $S_{xy} = \{s \in S \mid M[s, x] < M[s, y]\}$.
- The partial phylogeny of $T_R - \{w\}$ containing y is labelled by $S_{yx} = \{s \in S \mid M[s, y] < M[s, x]\}$.
- The partial phylogenies of $T_R - \{w\}$ containing neither x nor y are labelled by $E_{xy} = \{s \in S \mid M[s, x] = M[s, y]\}$.

For any phylogeny T labelled by S and species $x, y \in S$ a vertex or edge w of $P_T(x, y)$ is called the *midpath* for x and y in T, denoted $m(x, y)$, if $T - \{w\}$ satisfies the above trichotomy. If midpaths exist in T for all species pairs then T is called a *midpath phylogeny* of M. An example of a distance matrix M and a midpath phylogeny of M appear in Table 1 and Figure 1.

T is an *edge contraction* of the phylogeny T' at the edge (p, q) if T is obtained by removing p and q from T', creating a new vertex v and letting v's adjacencies in T be the union of p's and q's adjacencies in T'. Conversely, T' is called a *vertex expansion* of T at vertex v.

T' is an *expansion* of T if T' is obtained from T by a (possibly empty) sequence of vertex expansions. T is a *contraction* of T' if T' is an expansion of T.

T is a *minimal* midpath phylogeny of M if no contraction of T is a midpath phylogeny of M.

Theorem 2 [14]. *If it exists, the minimal midpath phylogeny of a distance matrix is unique.*

	Horse	Donkey	Pig	Dog	Rabbit	Monkey	Human	Pigeon
Horse	0	9	43	80	83	136	143	116
Donkey		0	34	71	74	127	134	107
Pig			0	38	53	105	112	78
Dog				0	59	111	118	80
Rabbit					0	84	91	59
Monkey						0	24	118
Human							0	126
Pigeon								0

Table 1. Cytochrome C sequence data appearing in [2].

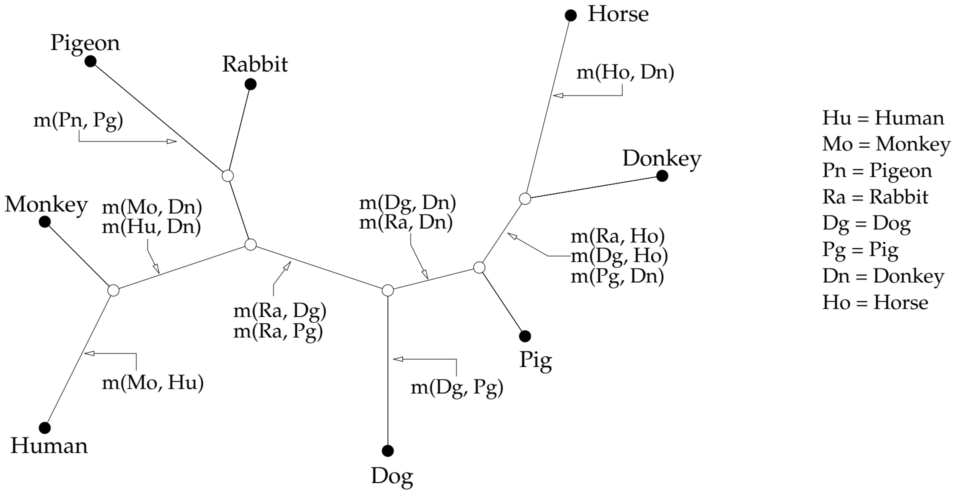

Fig. 1. Midpath phylogeny of the distance matrix appearing in Table 1. Only some of the midpaths are labelled though all exist.

Clearly, an ordinal representation T_R of M is also a midpath phylogeny of M. It follows that T_R is an expansion of the minimal midpath phylogeny T of M by Theorem 2. This is also true when T_R is a weighted ordinal representation of M. An optimal $O(n^2)$ algorithm exists for finding the minimal midpath phylogeny of a distance matrix where n is the number of species [14]. In section 3 an algorithm is presented that expands the minimal midpath phylogeny T of a distance matrix M to an ordinal representation T_R of M. In section 4 a method for expanding the minimal midpath phylogeny to a weighted ordinal representation is discussed.

A final observation is that expansion of the minimal midpath phylogeny is, in general, necessary as is exhibited by the following example.

Let T_1 and T_2 be the phylogenies depicted in Figure 2. If $M = d_{T_1}$ then T_2 is the minimal midpath phylogeny of M but is not an ordinal representation of M since $d_{T_2}(s, y) = d_{T_2}(s, x)$ but $M[s, y] < M[s, x]$.

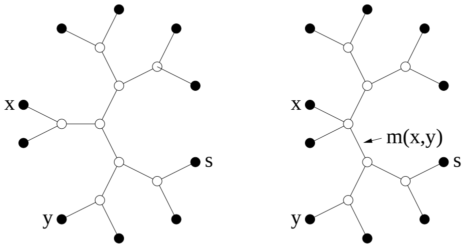

Fig. 2. Minimal midpath phylogenies are not always ordinal representations.

3 Expanding the Minimal Midpath Phylogeny

An ordinal representation T_R of M is an expansion of the minimal midpath phylogeny T of M. In this section we detail an algorithm that expands T to T_R. Before presenting the details of the expansion algorithm some notation is introduced. The following definitions apply to both partial and non–partial phylogenies.

If a phylogeny T is rooted at vertex r then this is denoted (T, r). Rooting a phylogeny provides an orientation of the phylogeny. If v is a vertex of (T, r) then $S(v)$ denotes the set of all species s such that $v \in P_T(s, r)$, that is, v is the least common ancestor of species in $S(v)$. If u is adjacent to v and $u \in P_T(v, r)$ then u is the *parent* of v in (T, r) and v is a *child* of u. A *shallowest species* of (T, r) is any species s closest to r and the *shallowest depth* of (T, r) is $d_T(r, s)$. If T is rooted at r and all species are equidistant from r then T is an *ultrametric* of *depth* $d_T(r, s)$, for any species s in T.

If T is a phylogeny and W is a subset of T's vertices then $T[W] = \cup_{w_i, w_j \in W} P_T(w_i, w_j)$ is the *subtree of* T *induced by* W. Given two subtrees T_1 and T_2 of T, $d_T(T_1, T_2)$ is the minimum distance in T between a vertex of T_1 and a vertex of T_2. In particular, $d_T(v, (p, q)) = min\{d_T(v, p), d_T(v, q)\}$ where v is a vertex of T and (p, q) is an edge of T. Define $P_T(T_1, T_2)$ to be the shortest path in T connecting T_1 and T_2 assuming T_1 and T_2 are disjoint. An *internal vertex* of T is a vertex that does not have degree one.

A phylogeny T labelled by S is *ordinally consistent* with the distance matrix M if $d_T(a, b) \leq d_T(c, d)$ if and only if $M[a, b] \leq M[c, d]$ for all species a, b, c and $d \in S$ where S is a subset of the species indexing M. Two phylogenies T_1 and T_2 are *ordinally equivalent* if $d_{T_1}(a, b) \leq d_{T_1}(c, d)$ if and only if $d_{T_2}(a, b) \leq d_{T_2}(c, d)$ for all species a, b, c and d labelling T_1 and T_2.

A concept that will be used repeatedly in the following sections is the correspondence between the location of a midpath $m(x, y)$ in T and the location of the midpoint of $P_{T_R}(x, y)$ in T_R. We refer to this as the *midpath–midpoint correspondence.* Before detailing the correspondence, it is necessary to relate a vertex or edge of T to the expansion of that vertex or edge within an expansion T' of T.

If T' is an *expansion* of T then the *expansion mapping*, denoted $exp_{T'}$, that maps vertices and edges of T to subtrees of T' is defined such that

- for each $v \in T$, $exp_{T'}(v)$ is the subtree of T' that resulted from expansion of v and
- for each edge $(p, q) \in T$, $exp_{T'}((p, q))$ is the edge of T' that connects $exp_{T'}(p)$ and $exp_{T'}(p)$ in T'.

An expansion mapping is illustrated in Figure 3.

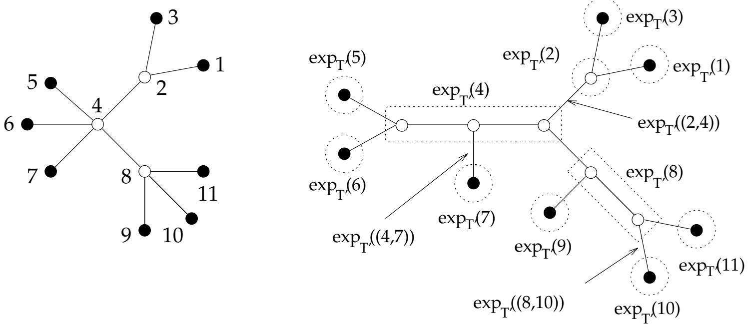

Fig. 3. A phylogeny T, an expansion T' of T and the expansion mapping $exp_{T'}$.

The midpath–midpoint correspondence is as follows:

- If the midpath for x and y in T is the edge (p,q) then $exp_{T_R}((p,q))$ is the midpoint of $P_{T_R}(x,y)$.
- If the midpath for x and y in T is the vertex v then there is a vertex in $exp_{T_R}(v)$ that is the midpoint of $P_{T_R}(x,y)$.

Hence, the midpath $m(x,y)$ in T specifies where the midpoint of $P_{T_R}(x,y)$ is located. This correspondence follows immediately from the trichotomy property of section 2.

The remainder of this section is organized as follows. An overview of the expansion algorithm is given in section 3.1. Section 3.2 introduces the concept of a *switch*. Switches play a central role in the combinatorics of ordinal representations. Section 3.3 presents the expansion algorithm in detail including the One–Two Theorem. Proofs of lemmas supporting the One–Two Theorem are presented in section 3.4. Implications of the One–Two Theorem are discussed in section 3.5. In particular, the ordinal representation of a distance matrix is unique. Finally, an analysis of the expansion algorithm's running time is given in section 3.6.

3.1 Expansion Algorithm Overview

Root the minimal midpath phylogeny T of M at an arbitrary internal vertex r. An ordinal representation T_R of M is obtained by expanding internal vertices of (T,r) such that a vertex is expanded only when all of its children have been expanded. Hence, to explain how the expansion algorithm proceeds it suffices to detail how an arbitrary internal vertex v of T is expanded.

To provide a context for the ensuing discussion, consider the locality of a vertex v in (T,r) and the locality of $exp_{T_R}(v)$ in T_R (see Figure 4). Let $v_1, \ldots, v_m$ denote the children of v in (T,r) and let u be the parent of v. In accordance with the notation introduced above, $T[S(v)]$ denotes the subtree of (T,r) rooted at v whereas $T_R[S(v)]$ is the expansion of $T[S(v)]$ in T_R.

To obtain $T_R[S(v)]$ from $T[S(v)]$ it is helpful to understand the structure of $T_R[S(v)]$. $T_R[S(v)]$ satisfies the following four conditions:

Condition 1: $T_R[S(v)]$ has at most one degree two vertex and it is a vertex of $exp_{T_R}(v)$.

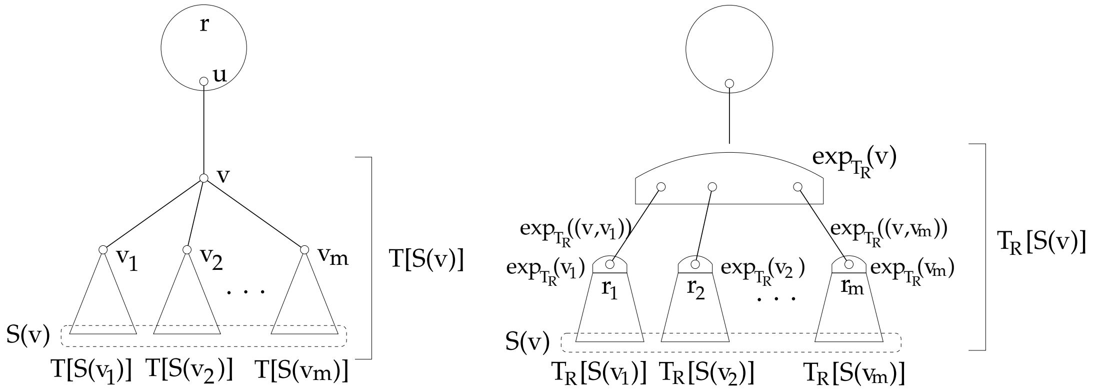

Fig. 4. The locality of v in the rooted minimal midpath phylogeny (T, r) and the locality of $exp_{T_R}(v)$ in the ordinal representation T_R.

Condition 1 is necessary since there is an edge (p, q) connecting $T_R[S(v)]$ to $T_R - T_R[S(v)]$ where $p \in exp_{T_R}(v)$ and $q \in exp_{T_R}(u)$, and so, if p has degree 3 in T_R then p has degree two in $T_R[S(v)]$. Conversely, $T_R[S(v)]$ cannot have more than one degree two vertex since T_R does not have degree two vertices.

Condition 2: $T_R[S(v)]$ is ordinally consistent with M.

Condition 2 follows from the fact that T_R is ordinally consistent with M.

Condition 3: If $m(x, y) = (v, v_i)$ then $exp_{T_R}((v, v_i))$ is the midpoint of $P_{T_R}(x, y)$.

Condition 3 follows from the fact that, assuming $x \in S(v_i)$, for all species $s \in S(v_i)$, $d_{T_R}(s, x) < d_{T_R}(s, y)$ and for all species $s \notin S(v_i)$, $d_{T_R}(s, y) < d_{T_R}(s, x)$.

Condition 4: If $m(x, y) = v$ then there is a vertex $w \in exp_{T_R}(v)$ such that w is the midpoint of $P_{T_R}(x, y)$. Furthermore, if $x \in S(v_i)$ and $y \in S(v_j)$ then $d_{T_R}(exp_{T_R}(v_i), exp_{T_R}(v_j)) = 2$.

Condition 4 follows from the fact that if $m(x, y) = v$ where $x \in S(v_i)$ and $y \in S(v_j)$ then for all $s \in S(v_i)$, $d_{T_R}(s, x) < d_{T_R}(s, y)$, for all $s \in S(v_j)$, $d_{T_R}(s, y) < d_{T_R}(s, x)$ and for all $s \notin S(v_i) \cup S(v_j)$, $d_{T_R}(s, x) = d_{T_R}(s, y)$. Hence, the midpoint of $P_{T_R}(x, y)$ must be an internal vertex of $P_{T_R}(exp_{T_R}(v_i), exp_{T_R}(v_j)) \subseteq exp_{T_R}(v)$.

If $d_{T_R}(exp_{T_R}(v_i), exp_{T_R}(v_j)) > 2$ then there are internal vertices p_1 and p_2 of $P_{T_R}(exp_{T_R}(v_i), exp_{T_R}(v_j))$ such that there are species $s_1, s_2 \notin S(v_i) \cup S(v_j)$ and $P_{T_R}(s_1, s_2) \cap P_{T_R}(x, y) = P_{T_R}(p_1, p_2)$. However, not both $d_{T_R}(s1, x) = d_{T_R}(s1, y)$ and $d_{T_R}(s2, x) = d_{T_R}(s2, y)$ as required. It follows by contradiction that $d_{T_R}(exp_{T_R}(v_i), exp_{T_R}(v_j)) = 2$.

Definition 3. *Solution*(v) is the set of expansions of $T[S(v)]$ that satisfy Conditions 1 through 4.

For each vertex v of T the algorithm determines *Solution*(v) dynamically such that *Solution*(v) is determined after *Solution*(v_1), *Solution*(v_2), … *Solution*(v_m) have been determined. Clearly, any phylogeny of *Solution*(r) is an ordinal representation of M as it satisfies Condition 2 and $S(r) = S$.

3.2 Switches

An inherent difficulty the expansion algorithm must address is that several expansions of $T[S(v)]$ may satisfy Conditions 1 through 4. For example, consider the phylogenies T_1, T and T_2 depicted in Figure 5. Let M be the distance matrix of T_1. Observe that T is the minimal midpath phylogeny of M and that both T_1 and T_2 are expansions of T at vertex r that satisfy Conditions 1 through 4, in particular, T_1 and T_2 are ordinally equivalent.

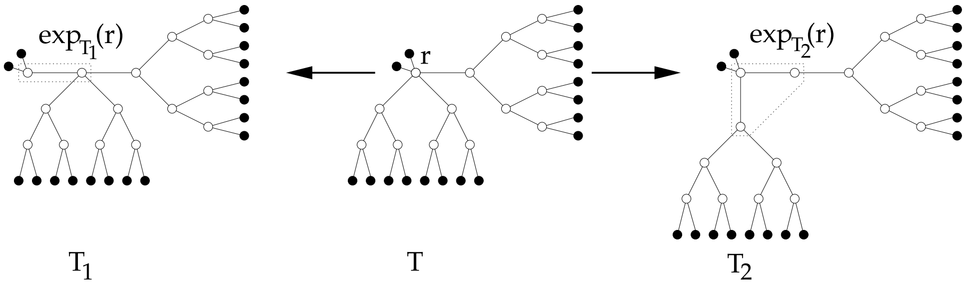

Fig. 5. T_1 and T_2 are expansions of T that are ordinally equivalent.

To achieve an efficient algorithm it is necessary that the size of $Solution(v)$ remains polynomial. In fact, it is shown in section 3.3 that if the size of $Solution(v)$ is greater than one then $Solution(v)$ is a *switch*:

Definition 4. An (n,k)-*switch* W is a pair of ordinally equivalent partial phylogenies $W[ON]$ and $W[OFF]$ with the following structural properties:

Let (U_1, u_1), (U_2, u_2), …, (U_n, u_n) and (U', u') be utrametrics each of depth k and let (U, u) be a rooted phylogeny. A *spine* is formed by joining (U, u) and (U', u') by a path $u = w_0$, w_1, …, $w_{n+1} = u'$ of length $n+1$. $W[ON]$ and $W[OFF]$ are defined as follows:

- u_1, …, u_n are identified with w_1, …, w_n, respectively, in $W[ON]$.
- u_1, …, u_n are adjacent to w_0, …, w_{n-1}, respectively, in $W[OFF]$.

So that we may refer to certain components of a switch W we write $W = (U, U_1, U_2, \ldots, U_n, U', w_{n+1}, w_n)$. The vertices w_n and w_{n+1} are distinguished as they later play an important role.

Switches are best described by pictures. In Figure 6 all species are located at the leaves of U_1, U_2, U' and U. U_1, U_2, and U' are ultrametrics of depth 4 whereas U is an ultrametric of depth 1. It can be verified by comparing the distance matrices of the two partial phylogenies that they are ordinally equivalent. Figure 7 depicts a $(1,3)$–switch and a $(3,5)$–switch. Notice that the partial phylogenies T_1 and T_2 of Figure 5 form the $(1,3)$–switch of Figure 7.

3.3 Details of the Expansion Algorithm: Determining *Solution*(v)

In this section we detail how $Solution(v)$ is determined given that $Solution(v_1)$, …, $Solution(v_m)$ have been determined. An important invariant that the algorithm maintains is the One–Two Theorem:

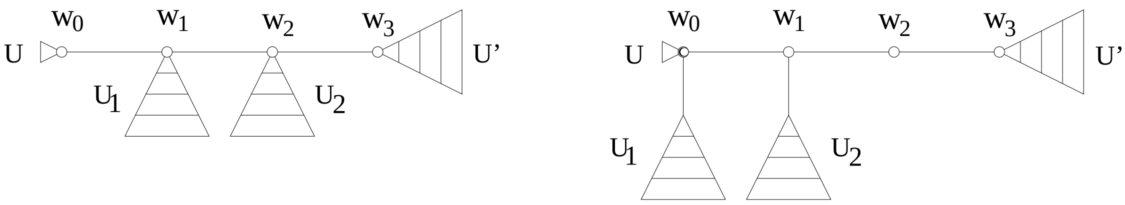

Fig. 6. A $(2,4)$-switch $W = (U, U_1, U_2, U', w_3, w_2)$ where $W[ON]$ is on the left and $W[OFF]$ is on the right.

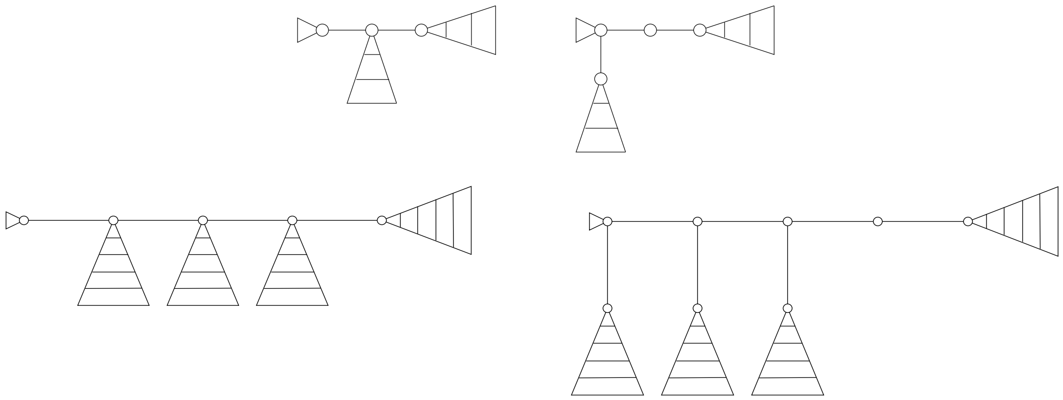

Fig. 7. A $(1,3)$–switch above and a $(3,5)$–switch below.

One–Two Theorem. *For each vertex v of T, if $|Solution(v)| > 1$ then $Solution(v)$ is a switch.*

A proof of the One–Two Theorem appears in section 3.5. It is assumed, by induction, that the One–Two Theorem holds for $Solution(v_1), \ldots, Solution(v_m)$. That is, for each i either $Solution(v_i) = \{T_i\}$ or $Solution(v_i)$ is a switch $\{T_i[ON], T_i[OFF]\}$. We proceed by simultaneously describing the algorithm and arguing for its correctness. The exposition concludes with a summary of the algorithm.

Define (T_i, r_i) to be a *selection* of $Solution(v_i)$ if $T_i \in Solution(v_i)$ and $r_i \in exp_{T_i}(v_i)$. A selection (T_i, r_i) of $Solution(v_i)$ is a *subsolution* if there is some $T_v \in Solution(v)$ such that $T_v[S(v_i)] = T_i$ and there exists an edge $(r_i, w_i) \in T_v$ for some vertex $w_i \in exp_{T_v}(v)$. In particular, each $T_v \in Solution(v)$ contains subsolutions $(T_1, r_1), \ldots, (T_m, r_m)$ selected from $Solution(v_1), \ldots, Solution(v_m)$ respectively. Our intent is to eliminate selections from each $Solution(v_i)$ that are not subsolutions of some $T_v \in Solution(v)$.

Let (T_i, r_i) and (T_j, r_j) be selections of $Solution(v_i)$ and $Solution(v_j)$, respectively. If (T_i, r_i) and (T_j, r_j) are subsolutions of $T_v \in Solution(v)$ then the distance $d_{T_v}(r_i, r_j)$ can be determined via a *midpath test*:

- Pick $s \in S(v_i)$ and $t \in S(v_j)$ arbitrarily and locate $m(s,t)$ in T. Observe that either $m(s,t) \in T[S(v_i)]$, $m(s,t) = (v, v_i)$, $m(s,t) = v$, $m(s,t) = (v, v_j)$ or $m(s,t) \in T[S(v_j)]$.
- If $m(s,t) \in T[S(v_i)]$ then the midpoint w of $P_{T_v}(s,t)$ is located in T_i by the midpath–midpoint

correspondence. Since T_v is ordinally consistent with M it must be that $d_{T_v}(s, w) = d_{T_v}(w, t) = d_{T_i}(w, r_i) + d_{T_v}(r_i, r_j) + d_{T_j}(r_j, t)$. All of these distances are known except $d_{T_v}(r_i, r_j)$ which can then be derived.
- If $m(s, t) \in T[S(v_j)]$ then $d_{T_v}(r_i, r_j)$ can be derived as above.
- If $m(s, t) = (v, v_i)$ then $\mathit{exp}_{T_v}((v, v_i))$ is the midpoint of $P_{T_v}(s, t)$ since T_v satisfies Condition 3, and so, $d_{T_v}(s, \mathit{exp}_{T_v}((v, v_i))) = d_{T_v}(\mathit{exp}_{T_v}((v, v_i)), t) = d_{T_v}(r_i, r_j) - 1 + d_{T_j}(r_j, t)$. All of these distances are known except $d_{T_v}(r_i, r_j)$ which can then be derived.
- If $m(s, t) = (v, v_j)$ then $d_{T_v}(r_i, r_j)$ can be derived as above.
- If $m(s, t) = v$ then $d_{T_v}(r_i, r_j) = 2$ since T_v satisfies Condition 4.

Given that $d_{T_v}(r_i, r_j)$ has been determined, define $\mathit{join}((T_i, r_i), (T_j, r_j))$ to be the partial phylogeny obtained by joining T_i and T_j by a path from r_i to r_j of length $d_{T_v}(r_i, r_j)$. In particular, $\mathit{join}((T_i, r_i), (T_j, r_j)) = T_v[S(v_i) \cup S(v_j)]$. We refer to $\mathit{join}((T_i, r_i), (T_j, r_j))$ as a $\mathit{Solution}(v_i) \times \mathit{Solution}(v_j)$ *join* and say that (T_i, r_i) and (T_j, r_j) *participate* in $\mathit{join}((T_i, r_i), (T_j, r_j))$. Observe that $\mathit{join}((T_i, r_i), (T_j, r_j))$ can be defined regardless of whether or not (T_i, r_i) and (T_j, r_j) are subsolutions of some $T_v \in \mathit{Solution}(v)$.

If (T_i, r_i) and (T_j, r_j) are subsolutions of $T_v \in \mathit{Solution}(v)$ then $J = \mathit{join}((T_i, r_i), (T_j, r_j))$ satisfies Conditions 2 through 4 since T_v satisfies Conditions 2 through 4. Furthermore, only internal vertices of $P_J(r_i, r_j)$ have degree two in J since T_v satisfies Condition 1.

Condition 1': Only internal vertices of $P_J(r_i, r_j)$ have degree two in $J = \mathit{join}((T_i, r_i), (T_j, r_j))$.

Define $\mathit{Candidates}(i, j)$ to be the set of all $\mathit{Solution}(v_i) \times \mathit{Solution}(v_j)$ joins that satisfy Conditions 1' and 2 through 4. If a selection (T_i, r_i) of $\mathit{Solution}(v_i)$ does not participate in a join in $\mathit{Candidates}(i, j)$ then (T_i, r_i) is not a subsolution. This technique is used to identify selections that are not subsolutions.

By induction both $\mathit{Solution}(v_i)$ and $\mathit{Solution}(v_j)$ satisfy the One–Two Theorem and so are each either size 1 or a switch. Under this assumption the following analysis of $\mathit{Candidates}(i, j)$ shows that there is one selection of $\mathit{Solution}(v_j)$ that participates in all joins in $\mathit{Candidates}(i, j)$ and at most two selections of $\mathit{Solution}(v_i)$ that participate in joins in $\mathit{Candidates}(i, j)$, or vice–versa. Hence, all but one selection of $\mathit{Solution}(v_j)$ are eliminated as subsolutions and all but two selections of $\mathit{Solution}(v_i)$ are eliminated as subsolutions, or vice–versa. The fundamental structural result utilized to reach this conclusion is the Joining Lemma:

Joining Lemma. *Let* $J = \mathit{join}((T_i, r_i), (T_j, r_j))$ *and* $J' = \mathit{join}((T'_i, r'_i), (T'_j, r'_j))$ *be two* $\mathit{Solution}(v_i)$ $\times \mathit{Solution}(v_j)$ *joins that satisfy Condition 1' and Conditions 2 through 4 such that* $\mathit{Solution}(v_i)$ *and* $\mathit{Solution}(v_j)$ *are each either size 1 or a switch and* $(T_i, r_i) \neq (T'_i, r'_i)$*. Then* $(T_j, r_j) = (T'_j, r'_j)$ *and if* $\mathit{Solution}(v_i) = \{T_i\}$ *then* $d_{T_i}(r_i, r'_i) = 1$*.*

Proof. See section 3.4. □

If $J = \mathit{join}((T_i, r_i), (T_j, r_j))$ and $J' = \mathit{join}((T'_i, r'_i), (T'_j, r'_j))$ such that $J, J' \in \mathit{Candidates}(i, j)$ then by the Joining Lemma either $(T_i, r_i) = (T'_i, r'_i)$ or $(T_j, r_j) = (T'_j, r'_j)$. It follows that there is either a selection of $\mathit{Solution}(v_i)$ or a selection of $\mathit{Solution}(v_j)$ that participates in all joins in $\mathit{Candidates}(i, j)$. Assume $(T_j, r_j) \in \mathit{Solution}(v_j)$ is this selection.

Suppose $\mathit{Solution}(v_i)$ is a switch $T_i = (U, U_1, \ldots, U_p, U', r_{ON}, r_{OFF})$. By the Switch Root Lemma there are at most two selections of $\mathit{Solution}(v_i)$ that can participate in $\mathit{Solution}(v_i) \times \mathit{Solution}(v_j)$ joins satisfying Conditions 1' and 2 through 4, namely, $(T_i[ON], r_{ON})$ and $(T_i[OFF], r_{OFF})$.

Switch Root Lemma. *Let* $Solution(v_i)$ *be a* (p,h)*-switch* $T_i = (U,\, U_1,\, \ldots,\, U_p,\, U',\, r_{ON},\, r_{OFF})$ *such that* $J = join((T_i[x], w), (T_j, r_j))$ *is a* $Solution(v_i) \times Solution(v_j)$ *join satisfying Conditions 1' and 2 through 4 where* $x \in \{ON, OFF\}$*. Then* $w \equiv r_x$*.*

Proof. See section 3.4. □

Suppose $Solution(v_i) = \{T_i\}$ and more than two $Solution(v_i)$ selections participate in joins in $Candidates(i,j)$. In particular, let $J, J', J'' \in Candidates(i,j)$ such that $J = join((T_i, r_i), (T_j, r_j))$, $J' = join((T_i, r_i'), (T_j, r_j))$ and $J'' = join((T_i, r_i''), (T_j, r_j))$. By the Joining Lemma, $d_{T_i}(r_i, r_i') = d_{T_i}(r_i, r_i'') = d_{T_i}(r_i', r_i'') = 1$. However, this implies that $\{r_i, r_i', r_i''\}$ induces a cycle in T_i contradicting T_i being a tree. It follows that no more than two $Solution(v_i)$ selections participate in joins in $Candidates(i,j)$.

Hence, there is one selection of $Solution(v_j)$ that participates in all joins in $Candidates(i,j)$ and at most two selections of $Solution(v_i)$ that participate in joins in $Candidates(i,j)$, or vice–versa. Let $W = \{1, \ldots, m\}$ and let $Selections$ denote a set of selections that is initially empty. Repeat the following procedure until W is reduced to a set of size 1.

- Select i and j from W.
- Compute $Candidates(i,j)$.
- By the above there is some selection (T_i, r_i) of $Solution(v_i)$ or some selection (T_j, r_j) of $Solution(v_j)$ that participates in all joins in $Candidates(i,j)$. Assume the former.
- Remove i from W and add (T_i, r_i) to $Selections$.

After $m-1$ iterations $Selections$ contains $m-1$ selections such that if $(T_i, r_i) \in Selections$ then (T_i, r_i) is the only selection of $Solution(v_i)$ that can be a subsolution. Without loss of generality, assume W is reduced to $\{m\}$. Then for some i, $Candidates(i,m)$ was computed above, furthermore, there are no more than two $Solution(v_m)$ selections (T_m, r_m) and (T_m', r_m') that participate in joins in $Candidates(i,m)$, and so, (T_m, r_m) and (T_m', r_m') are the only selections of $Solution(v_m)$ that can be subsolutions.

Let T_v denote the element of $Solution(v)$, if it exists, for which the elements of $Selections \cup \{(T_m, r_m)\}$ are subsolutions and let T_v' denote the element of $Solution(v)$, if it exists, for which the elements of $Selections \cup \{(T_m', r_m')\}$ are subsolutions. Using midpath tests, $d_{T_v}(a,b)$ can be determined for every $a, b \in \{r_1, \ldots, r_m\}$. Similarly, $d_{T_v'}(a,b)$ can be determined for every $a, b \in \{r_1, \ldots, r_{m-1}, r_m'\}$. T_v (and T_v') can be constructed efficiently and uniquely from this distance information [4, 20]. The process by which T_v is recovered from its set of subsolutions $Selections \cup \{(T_m, r_m)\}$ is denoted $T_v = recon(Selections \cup \{(T_m, r_m)\})$. Similarly for T_v'. If T_v and T_v' satisfy Conditions 1 through 4 then they are admitted to $Solution(v)$.

Algorithm Summary

1. Find the minimal midpath phylogeny T of M. Root T at some arbitrary internal vertex r.
2. For each vertex v of T, if v is a leaf then $Solution(v) = \{v\}$, otherwise, suppose v has children v_1, …, v_m such that $Solution(v_1)$, …, $Solution(v_m)$ have been determined. $Solution(v)$ is computed as follows:
   (a) $W = \{1, \ldots, m\}$
   (b) $Selections = \emptyset$
   (c) While $|W| > 1$ do

```
      i. Select $i, j$ from $W$
     ii. Compute $Candidates(i, j)$
    iii. If $(T_i, r_i)$ participates in all joins of $Candidates(i, j)$ then
           A. $W = W - \{i\}$
           B. $Selections = Selections \cup \{(T_i, r_i)\}$
           C. $partner(j) = i$
     iv. else $(T_j, r_j)$ participates in all joins of $Candidates(i, j)$ then
           A. $W = W - \{j\}$
           B. $Selections = Selections \cup \{(T_j, r_j)\}$
           C. $partner(i) = j$
  (d) end while
  (e) If $W = \{j\}$ and $(T_j, r_j)$, $(T'_j, r'_j)$ are the only selections of $Solution(v_j)$ participating in joins
      of $Candidates(partner(j), j)$ then
      i. Compute $T_v = recon(Selections \cup \{(T_j, r_j)\})$
     ii. Compute $T'_v = recon(Selections \cup \{(T'_j, r'_j)\})$
  (f) If $T_v$ $(T'_v)$ satisfy Conditions 1 through 4 then admit $T_v$ $(T'_v)$ to $Solution(v)$.
3. Return any phylogeny from $Solution(r)$.
```

3.4 Technical Developments

In this section we present the proofs of the structural lemmas supporting the expansion algorithm appearing in section 3.3.

Shallowest Depth Lemma. *A rooted partial phylogeny where only the root is permitted to have degree two and the shallowest depth is k has species pairs that are distances $0, 2, \ldots, 2k$ apart.*

Proof. The Lemma is proven by induction on the depth of the rooted phylogeny. The depth of a rooted phylogeny is the largest distance from a species to the root. Clearly the Lemma holds for phylogenies of depth 0.

To prove the general case, let T be a phylogeny rooted at r with shallowest depth k and depth $d > 0$. Let $v_1, \ldots, v_m$ be the children of r in (T, r) and define $T_i = T[S(v_i)]$ for $1 \leq i \leq m$.

Since T has a species at shallowest depth k, one of $(T_1, v_1), \ldots, (T_m, v_m)$, say (T_1, v_1), has a species s at shallowest depth $k - 1$. Since T_1 has depth less than d and the root v_1 of T_1 does not have degree two in T_1, there are species pairs in T_1 realizing distances $0, 2, \ldots, 2k - 2$ by induction. It remains to exhibit a pair of species in T distance $2k$ apart.

If any (T_i, v_i) where $i \geq 2$ has a species t at shallowest depth $k - 1$ then $d_T(s, t) = 2k$. Otherwise, (T_2, v_2) has shallowest depth $l \geq k$ and depth less than d. By induction there are species pairs in T_2 realizing distances $0, 2, \ldots, 2k, \ldots, 2l$. $\square$

Joining Lemma. *Let $J = join((T_i, r_i), (T_j, r_j))$ and $J' = join((T'_i, r'_i), (T'_j, r'_j))$ be two $Solution(v_i)$ $\times Solution(v_j)$ joins that satisfy Conditions 1' and 2 through 4 such that $Solution(v_i)$ and $Solution(v_j)$ are each either size 1 or a switch and $(T_i, r_i) \neq (T'_i, r'_i)$. Then $(T_j, r_j) = (T'_j, r'_j)$ and if $Solution(v_i) = \{T_i\}$ then $d_{T_i}(r_i, r'_i) = 1$.*

Proof. Define

$$\hat{T}_i = \begin{cases} \quad T_i & \text{, if } \mathit{Solution}(v_i) = \{T_i\} \\ \text{spine of } T_i & \text{, if } \mathit{Solution}(v_i) \text{ is a } (p,h)\text{–switch } T_i = (U,\, U_1,\, \ldots,\, U_p,\, U',\, r_{ON},\, r_{OFF}) \end{cases}$$

and similarly define

$$\hat{T}_j = \begin{cases} \quad T_j & \text{, if } \mathit{Solution}(v_j) = \{T_j\} \\ \text{spine of } T_j & \text{, if } \mathit{Solution}(v_j) \text{ is a } (q,k)\text{–switch } T_j = (W,\, W_1,\, \ldots,\, W_q,\, W',\, s_{ON},\, s_{OFF}). \end{cases}$$

Let $\hat{S}_i$ denote the species of $\hat{T}_i$ and let $\hat{S}_j$ denote the species of $\hat{T}_j$. Note that by definition of switch, $J[\hat{S}_i] = J'[\hat{S}_i]$ and $J[\hat{S}_j] = J'[\hat{S}_j]$.

If $T_i \neq T_i'$ then $\mathit{Solution}(v_i)$ is a (p,h)–switch $T_i = (U,\, U_1,\, \ldots,\, U_p,\, U',\, r_{ON},\, r_{OFF})$, and by the Switch Root Lemma, we may assume $J = ((T_i[ON], r_{ON}), (T_j, r_j))$ and $J' = ((T_i[OFF], r_{OFF}), (T_j', r_j'))$. Similarly, if $T_j \neq T_j'$ then $\mathit{Solution}(v_j)$ is a (q,k)–switch $T_j = (W,\, W_1,\, \ldots,\, W_q,\, W',\, s_{ON},\, s_{OFF})$, and by the Switch Root Lemma, $J = ((T_i, r_i), (T_j[x], s_x))$ and $J' = ((T_i', r_i'), (T_j[y], s_y))$ where $\{x, y\} = \{ON, OFF\}$.

Define $A = \{s \in \hat{S}_i \mid r_i \in P_{\hat{T}_i}(s, r_i')\}$, $B = \{s \in \hat{S}_i \mid r_i' \in P_{\hat{T}_i}(s, r_i)\}$, $C = \{s \in \hat{S}_j \mid r_j \in P_{\hat{T}_j}(s, r_j')\}$ and $D = \{s \in \hat{S}_j \mid r_j' \in P_{\hat{T}_j}(s, r_j)\}$. Let a, b, c and d be shallowest depth species of $(\hat{T}_i[A], r_i)$, $(\hat{T}_i[B], r_i')$, $(\hat{T}_j[C], r_j)$ and $(\hat{T}_j[D], r_j')$ respectively. For brevity, let $l = d_J(r_i, r_j)$, $l' = d_{J'}(r_i', r_j')$, $w_i = d_{\hat{T}_i}(r_i, r_i')$ and $w_j = d_{\hat{T}_j}(r_j, r_j')$. The joins J and J' are depicted in Figure 8.

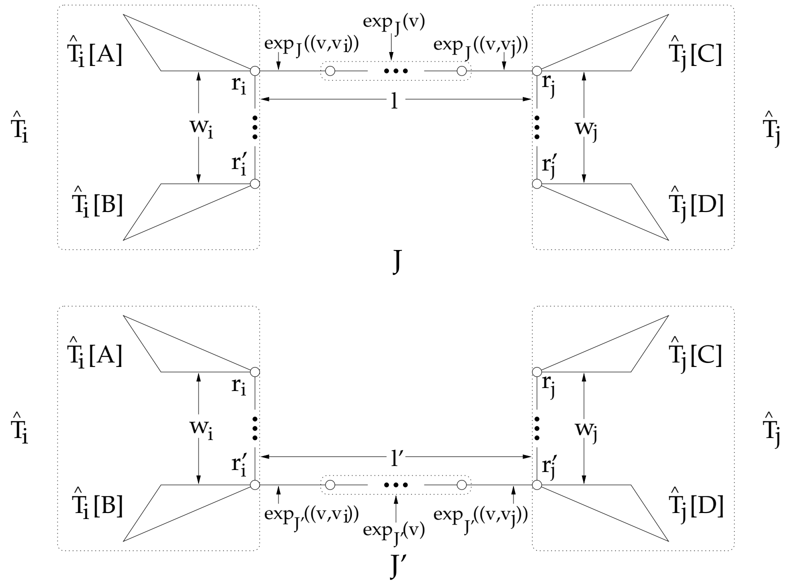

Fig. 8. The $\mathit{Solution}(v_i) \times \mathit{Solution}(v_j)$ joins J and J'.

Dependencies among the values l, l', w_i and w_j can be derived by considering the locations of midpaths $m(a,c)$, $m(a,d)$, $m(b,c)$ and $m(b,d)$ in the minimal midpath phylogeny T. In particular, for any $x \in S(v_i)$ and $y \in S(v_j)$, $m(x,y) \in T[S(v_i)] \cup \{(v,v_i), v, (v,v_j)\} \cup T[S(v_j)]$.

For example, suppose $m(a,c) \in T[S(v_i)]$. It follows by midpath–midpoint correspondence that the midpoint w of $P_J(a,c)$ occurs in $J[S(v_i)]$ and the midpoint w' of $P_{J'}(a,c)$ occurs in $J'[S(v_i)]$. This yields

$$d_J(a,w) = d_J(w,r_i) + d_J(r_i,r_j) + d_J(r_j,c) \tag{2}$$

and

$$d_{J'}(a,w') = d_{J'}(w',r_i) + d_{J'}(r_i,r_i') + d_{J'}(r_i',r_j') + d_{J'}(r_j,r_j') + d_{J'}(r_j,c). \tag{3}$$

But $d_J(a,w) = d_{J'}(a,w')$, $d_J(w,r_i) = d_{J'}(w',r_i)$ and $d_J(r_j,c) = d_{J'}(r_j,c)$ since $J[\hat{S}_i] = J'[\hat{S}_i] = \hat{T}_i$ and $J[\hat{S}_j] = J'[\hat{S}_j] = \hat{T}_j$. Furthermore, $d_{J'}(r_i,r_i') = d_{\hat{T}_i}(r_i,r_i')$ and $d_{J'}(r_j,r_j') = d_{\hat{T}_j}(r_j,r_j')$. Combining equalities (2) and (3) yields $d_J(r_i,r_j) = d_{J'}(r_i',r_j') + d_{\hat{T}_i}(r_i,r_i') + d_{\hat{T}_j}(r_j,r_j')$, or in terms of the shortened notation, $l = l' + w_i + w_j$.

For another example, suppose $m(b,d) = v$. Since J and J' satisfy Condition 4, $l = 2$, $l' = 2$, the midpoint of $P_J(b,d)$ is the vertex w adjacent to r_i and r_j, and the midpoint of $P_{J'}(b,d)$ is the vertex w' adjacent to r_i' and r_j'. This yields

$$d_J(b,r_i') + w_i + 1 = w_j + d_J(r_j',d) + 1$$

and

$$d_{J'}(b,r_i') + 1 = d_{J'}(r_j',d) + 1.$$

It follows from $d_J(b,r_i') = d_{J'}(b,r_i')$ and $d_J(r_j',d) = d_{J'}(r_j',d)$ that $w_i = w_j$ (in addition to $l = l' = 2$ from above).

Using this technique dependencies among l, l', w_i and w_j can be derived for each possible location of midpaths $m(a,c)$, $m(a,d)$, $m(b,c)$ and $m(b,d)$ in T. This *midpath survey* appears in Table 2. The midpath survey can be used to derive restrictions on midpath locations. For example, if $m(b,c) \in T[S(v_i)]$ then $l + w_i = l' + w_j$ and if $m(a,d) = (v,v_j)$ then $l = l' + w_i + w_j$. Observe that combined these two dependencies imply $w_i = 0$. However, by assumption $(T_i, r_i) \neq (T_i', r_i')$. Hence, if $Solution(v_i) = \{T_i\}$ then $T_i = T_i'$ implying that $r_i \neq r_i'$. Otherwise $Solution(v_i)$ is a switch, and so by the argument above, $r_i \equiv r_{ON}$ and $r_i' \equiv r_{OFF}$ and so $w_i = 1$. It follows that $w_i = d_{\hat{T}_i}(r_i,r_i') > 0$. We conclude that not both $m(b,c) \in T[S(v_i)]$ and $m(a,d) = (v,v_j)$. In this way the midpath survey is utilized in the ensuing proof to analyze the structure of J and J'.

We proceed with a case-by-case analysis based upon the location of $m(a,d)$ in T:

Case $m(a,d) \in T[S(v_i)]$:

By the midpath survey $l + w_j = l' + w_i$. If $m(b,d) \in T[S(v_i)]$ then $l' = l + w_i + w_j$, and so, $w_i = 0$ — a contradiction. It follows that $m(b,d) \notin T[S(v_i)]$, hence, the midpoint of $P_J(a,d)$ is in $\hat{T}_i$ but the midpoint of $P_J(b,d)$ is not in $\hat{T}_i$. This implies that

$$d_J(a,r_i) \geq d_J(b,r_i) = d_J(b,r_i') + w_i \tag{4}$$

and since the midpoint of $P_J(a,d)$ is in $\hat{T}_i$

$$d_J(a,r_i) \geq d_J(d,r_i) = d_J(d,r_j') + w_j + l. \tag{5}$$

	$m(a,c)$	$m(a,d)$	$m(b,c)$	$m(b,d)$
$T[S(v_i)]$	$l = l' + w_i + w_j$	$l + w_j = l' + w_i$	$l + w_i = l' + w_j$	$l' = l + w_i + w_j$
(v, v_i)	$l + w_i = l' + w_j$	$l' = l + w_i + w_j$	$l = l' + w_i + w_j$	$l + w_j = l' + w_i$
v	$l = l' = 2$ $w_i = w_j$	$w_i = 0$ $w_j = 0$	$w_i = 0$ $w_j = 0$	$l = l' = 2$ $w_i = w_j$
(v, v_j)	$l + w_j = l' + w_i$	$l = l' + w_i + w_j$	$l' = l + w_i + w_j$	$l + w_i = l' + w_j$
$T[S(v_j)]$	$l = l' + w_i + w_j$	$l + w_j = l' + w_i$	$l + w_i = l' + w_j$	$l' = l + w_i + w_j$

Table 2. Midpath location and the resulting dependency among l, l', w_i and w_j.

Combining inequalities (4) and (5) yields

$$2d_J(a, r_i) \geq d_J(b, d) > d_{J'}(b, d). \tag{6}$$

The last inequality follows from the fact that $d_J(b, d) - d_{J'}(b, d) = l + w_i + w_j - l'$, however, $l + w_j = l' + w_i$ from above implying $d_J(b, d) - d_{J'}(b, d) > 0$ since $w_i > 0$.

It is shown that $(\hat{T}_i[A], r_i)$ has no degree two vertices except, possibly, r_i. This allows the Shallowest Depth Lemma to be applied to $(\hat{T}_i[A], r_i)$. Suppose $Solution(v_i) = \{T_i\}$, and so, $(\hat{T}_i[A], r_i) = (T_i[A], r_i)$. By assumption only internal vertices of $P_{J'}(r_i', r_j')$ have degree two in J', and so, $(T_i[A], r_i)$ has no degree two vertices except, possibly, r_i.

Suppose $Solution(v_i)$ is a switch $T_i = (U, U_1, \ldots, U_p, U', r_{ON}, r_{OFF})$, and so, $(\hat{T}_i[A], r_i) = (U', r_{ON})$. Since $T_i[OFF] \in Solution(v_i)$, $T_i[OFF]$ satisfies Condition 1. The degree of r_{OFF} is 2 in $T_i[OFF]$ implying that no vertex in (U', r_{ON}) has degree two except, possibly, r_{ON}.

The Shallowest Depth Lemma can be applied to $(\hat{T}_i[A], r_i)$, and so, there are species pairs in A distances 2, 4, ..., $2d_{\hat{T}_i}(a, r_i)$ apart. Observe that $d_{\hat{T}_i}(a, r_i) = d_J(a, r_i)$. It follows from inequality 6 that there are species $a_1, a_2 \in A$ such that either $d_{\hat{T}_i}(a_1, a_2) = d_J(b, d)$, if $d_J(b, d)$ is even, or, $d_{\hat{T}_i}(a_1, a_2) = d_J(b, d) - 1$, if $d_J(b, d)$ is odd.

However, $d_J(b, d) > d_{J'}(b, d)$ but $d_J(a_1, a_2) = d_{J'}(a_1, a_2) = d_{\hat{T}_i}(a_1, a_2)$. So if $d_J(b, d)$ is even then $d_J(a_1, a_2) = d_J(b, d)$ whereas $d_{J'}(a_1, a_2) > d_{J'}(b, d)$. This contradicts both J and J' being ordinally consistent with M as required by Condition 2. If $d_J(b, d)$ is odd then $d_J(b, d) > d_J(a_1, a_2)$ whereas $d_{J'}(a_1, a_2) \geq d_{J'}(b, d)$. This also contradicts both J and J' being ordinally consistent with M as required by Condition 2.

Case $m(a, d) = (v, v_i)$**:**

By the midpath survey $l' = l + w_i + w_j$. The only location for $m(a, c)$ that does not result in $w_i = 0$ (a contradiction) is $m(a, c) = (v, v_i)$. By the midpath survey $l + w_i = l' + w_j$, and so, $w_j = 0$. This implies $r_j = r_j'$, and so, $T_j = T_j'$. Finally, $w_j = 0$ also implies $l' = l + w_i$.

Case $m(a, d) = v$**:**

By the midpath survey $w_i = 0$ which, as argued above, is a contradiction.

Case $m(a, d) = (v, v_j)$**:**

By the midpath survey $l = l' + w_i + w_j$. The only location for $m(b, d)$ that does not result in $w_i = 0$ (a contradiction) is $m(b, d) = (v, v_i)$. By the midpath survey $l + w_j = l' + w_i$, and so, $w_j = 0$. This implies $r_j = r_j'$, and so, $T_j = T_j'$. Finally, $w_j = 0$ also implies $l = l' + w_i$.

Case $m(a, d) \in T[S(v_j)]$**:**

Suppose $Solution(v_j) = \{T_j\}$. Then $(\hat{T}_j[D], r'_j) = (T_j[D], r'_j)$ and $(\hat{T}_j[D], r'_j)$ has no degree two vertices except, possibly, r'_j. This follows since, by assumption, only internal vertices of $P_J(r_i, r_j)$ have degree two in J, and so, $(T_j[D], r'_j)$ has no degree two vertices except, possibly, r'_j. An argument analogous to the argument presented in the case $m(a,d) \in T[S(v_i)]$ results in a contradiction where a, d and $(\hat{T}_i[A], r_i)$ are replaced by d, a and $(\hat{T}_j[D], r'_j)$.

Suppose $Solution(v_j)$ is a (q,k)–switch $T_j = (W, W_1, \ldots, W_q, W', s_{ON}, s_{OFF})$. As noted at the beginning of the proof, $J = ((T_i, r_i), (T_j[x], s_x))$ and $J' = ((T'_i, r'_i), (T_j[y], s_y))$ where $\{x, y\} = \{ON, OFF\}$.

If $x \equiv OFF$ and $y \equiv ON$ then $(\hat{T}_j[D], r'_j) = (W', s_{ON})$ has no degree two vertices except, possibly, s_{ON}. Again, an argument analogous to the argument presented in the case $m(a,d) \in T[S(v_i)]$ results in a contradiction where a, d and $(\hat{T}_i[A], r_i)$ are replaced by d, a and $(\hat{T}_j[D], r'_j)$.

Otherwise, suppose $x \equiv ON$ and $y \equiv OFF$. Since $m(a,d) \in T[S(v_j)]$, by the midpath survey $l + w_j = l' + w_i$. If $m(b,c) \in \{(v, v_i), v, (v, v_j)\}$ then either $w_i = 0$ or $w_j = 0$ which are contradictions. Otherwise, $m(b,c) \in T[S(v_i)] \cup T[S(v_j)]$ and so $l + w_i = l' + w_j$ implying $w_i = w_j$ and $l = l'$. In fact, since $Solution(v_j)$ is a switch, $d_{\hat{T}_j}(s_{ON}, s_{OFF}) = w_j = 1$.

Let $p_0, p_1, \ldots, p_q, p_{q+1}$ denote the path in $\hat{T}_j$ connecting the root (p_0) of W to the root (p_{q+1}) of W' (see Definition 4).

Since $m(a,d) \in T[S(v_j)]$, the midpoint of $P_{J'}(a,d)$ is in $\hat{T}_j$. It follows that

$$d_{J'}(d, p_0) \geq d_{J'}(a, r_i) + w_i + l' - q. \tag{7}$$

It is now shown that the midpoint of $P_{J'}(c,d) \in \hat{T}_j[D]$. Suppose the midpoint of $P_{J'}(c,d) = p_g$, $0 < g \leq q$. Since $d_{J'}(c, p_q) = d_J(c, p_q)$ and $d_{J'}(p_q, d) = d_J(p_q, d)$, p_g is also the midpoint of $P_J(c,d)$. But then $d_J(s,d) = d_J(s,c)$ for all $s \in W_g$ yet $d_{J'}(s,d) < d_{J'}(s,c)$ for all $s \in W_g$. It follows that not both J and J' are ordinally consistent with M violating Condition 2. Similarly, a contradiction occurs if the midpoint of $P_{J'}(c,d)$ is the edge (p_{g-1}, p_g) for some g such that $0 < g \leq q$. It must be that the midpoint of $P_{J'}(c,d)$ is in $\hat{T}_j[D] = W$ or in $\hat{T}_j[C] = W'$.

If $m(a,c) \in T[S(v_j)]$ then by the midpath survey $l = l' + w_i + w_j$ implying $w_j = 0$ — a contradiction. It follows that the midpoint of $P_{J'}(a,c)$ is not in $\hat{T}_j$. Since the midpoint of $P_{J'}(a,d)$ is in $\hat{T}_j$ it must be that $d_{\hat{T}_j}(d, s_{OFF}) > d_{\hat{T}_j}(c, s_{OFF})$ and so the midpoint of $P_{J'}(c,d)$ is in $\hat{T}_j[D]$. It follows that

$$d_{J'}(d, p_0) \geq d_{J'}(c, s_{ON}) + q + w_j. \tag{8}$$

Combining inequalities (7) and (8) yields

$$2d_{J'}(d, p_0) \geq d_{J'}(a,c) > d_J(a,c). \tag{9}$$

The last inequality holds since $l = l'$ and $w_i = w_j = 1$.

The Shallowest Depth Lemma can be applied to $(\hat{T}_j[D], p_0) = (W, p_0)$ since W has no degree two vertices except, possibly, p_0. This is true since $T_j[OFF] \in Solution(v_j)$, and so, $T_i[OFF]$ satisfies Condition 1. However, s_{OFF} is degree two in $T_j[OFF]$ implying that no vertices of W have degree two in $T_i[OFF]$.

It follows that there are species pairs in D distances $2, 4, \ldots, 2d_{J'}(d, p_0)$ apart. In particular, there are species $w_1, w_2 \in D$ such that if $d_{J'}(a,c)$ is even then $d_{\hat{T}_j}(w_1, w_2) = d_{J'}(a,c)$ and if $d_{J'}(a,c)$ is odd then $d_{\hat{T}_j}(w_1, w_2) = d_{J'}(a,c) - 1$.

However, $d_{J'}(a,c) > d_J(a,c)$ by inequality (9 yet $d_J(w_1,w_2) = d_{J'}(w_1,w_2) = d_{\hat{T}_j}(w_1,w_2)$. So if $d_{J'}(a,c)$ is even then $d_{J'}(w_1,w_2) = d_{J'}(a,c)$ whereas $d_J(w_1,w_2) > d_J(a,c)$. This contradicts both J and J' being ordinally consistent with M violating Condition 2. If $d_{J'}(a,c)$ is odd then $d_{J'}(a,c) > d_{J'}(w_1,w_2)$ whereas $d_J(w_1,w_2) \geq d_J(a,c)$. This also contradicts both J and J' being ordinally consistent with M.

This concludes the case–by–case argument. Thus far it has been shown that $(T_j, r_j) = (T'_j, r'_j)$ and either $l = l' + w_i$ or $l' = l + w_i$. It remains to prove $w_i = 1$ if $Solution(v_i) = \{T_i\}$.

Let $p \in P_{T_i}(r_i, r'_i)$ such that p is adjacent to r_i. Let s be any species such that $\{p\} = P_{T_i}(s, r_i) \cap P_{T_i}(s, r'_i) \cap P_{T_i}(r_i, r'_i)$. Dependencies among l, l' and w_i can be derived by considering all possible locations for $m(c,s)$ along $P_T(c,s)$:

If $m(c,s) \in T[S(v_i)] \cup \{(v,v_j)\} \cup T[S(v_j)]$ then $l+2 = l'+w_i$. The case–by–case argument proved either $l = l' + w_i$ or $l' = l + w_i$. The former yields a contradiction whereas the latter implies $w_i = 1$.

If $m(c,s) = (v, v_i)$ then $l + w_i = l' + 2$. The case–by–case argument proved either $l = l' + w_i$ or $l' = l + w_i$. The latter yields a contradiction whereas the former implies $w_i = 1$.

Finally, if $m(c,s) = v$ then $l' = l$ implying $w_i = 0$ contradicting the assumption $w_i > 0$. We conclude that $w_i = 1$. $\square$

Switch Root Lemma. *Let $Solution(v_i)$ be a (p,h)-switch $T_i = (U, U_1, \ldots, U_p, U', r_{ON}, r_{OFF})$ such that $J = join((T_i[x], w), (T_j, r_j))$ is a $Solution(v_i) \times Solution(v_j)$ join satisfying Conditions 1' and 2 through 4 where $x \in \{ON, OFF\}$. Then $w \equiv r_x$.*

Proof. If $x \equiv OFF$ then $w \equiv r_{OFF}$ since r_{OFF} is degree two in $T_i[OFF]$ and J satisfies Condition 1'.

Suppose $x \equiv ON$ and observe that since $T_i[OFF] \in Solution(v_i)$ and the degree of r_{OFF} is two in $T_i[OFF]$, it must be that $r_{OFF} \in exp_{T_i[OFF]}(v_i)$. Hence, if $d_{T_i[OFF]}(r_{OFF}, u_1) \leq d_{T_i[OFF]}(r_{OFF}, u_2)$ where $u_1, u_2 \in S(v_i)$ then $d_J(s, u_1) \leq d_J(s, u_2)$ for any $s \in S(v_j)$. Consequently, $d_J(w, u_1) \leq d_J(w, u_2)$.

It follows that $d_J(w, u') = d_J(w, u'')$ for all $u', u'' \in U'$ since $d_{T_i[OFF]}(r_{OFF}, u') = d_{T_i[OFF]}(r_{OFF}, u'')$. Also, $d_J(w, u') < d_J(w, u_p)$, for all $u' \in U'$, $u_p \in U_p$ since $d_{T_i[OFF]}(r_{OFF}, u') < d_{T_i[OFF]}(r_{OFF}, u_p)$. The only vertex w in $T_i[ON]$ satisfying these constraints is r_{ON}. $\square$

3.5 The One–Two Theorem and its Implications

Let M be a distance matrix and T the minimal midpath phylogeny of M rooted at r. Applying the One–Two Theorem to $Solution(r)$ yields either $Solution(r)$ has size 1, and so M has a unique ordinal representation, or $Solution(r)$ is a switch T_R. However, $T_R[OFF]$ has a degree two vertex by Definition 4, and so, is not an ordinal representation of M. It follows that $T_R[ON]$ is the unique ordinal representation of M. As a corollary to the One–Two Theorem we have:

Theorem 5. *If M has an ordinal representation then it is unique.*

This generalizes the well–known result that the distance matrix d_T of a phylogeny T is unique to T [10, 20]. In fact, the ordinal assertions supported by d_T are unique to T.

One–Two Theorem. *For each vertex v of T, if $|Solution(v)| > 1$ then $Solution(v)$ is a switch.*

Proof. Recall from section 3.3 that if $|Solution(v)| > 1$ then $Solution(v) = \{T_v, T'_v\}$ where $T_v = recon(Selections \cup \{(T_m, r_m)\})$ and $T'_v = recon(Selections \cup \{(T'_m, r'_m)\})$. Here it is shown that $\{T_v, T'_v\}$ is a switch.

Define

$$\hat{T_m} = \begin{cases} T_m & \text{, if } Solution(v_m) = \{T_m\} \\ \text{spine of } T_m & \text{, if } Solution(v_m) \text{ is a } (q,h)\text{–switch } T_m = (U,\, U_1,\, \ldots,\, U_q,\, U',\, r_{ON},\, r_{OFF}) \end{cases}$$

and let $\hat{S_m}$ be the species of $\hat{T_i}$. Let $T' = T_v[S(v) - S(v_m)]$. Since (T_1, r_1), …, (T_{m-1}, r_{m-1}) are subsolutions of both T_v and T'_v, $T' = T'_v[S(v) - S(v_m)]$ as well. Let p be the vertex of T' adjacent to some vertex in $T_v - T'$ and let p' be the vertex of T' adjacent to some vertex in $T'_v - T'$. In particular, $T_v = (T_m, r_m) \cup P_{T_v}(r_m, p) \cup T'$ and $T'_v = (T'_m, r'_m) \cup P_{T'_v}(r'_m, p') \cup T'$.

Define $A = \{s \in \hat{S_m} | r_m \in P_{\hat{T_m}}(s, r'_m)\}$, $B = \{s \in \hat{S_m} | r'_m \in P_{\hat{T_m}}(s, r_m)\}$, $C = \{s \in S(v) - S(v_m) | p \in P_{T'}(s, p')\}$ and $D = \{s \in S(v) - S(v_m) | p' \in P_{T'}(s, p)\}$. For brevity let $l = d_{T_v}(r_m, p)$, $l' = d_{T'_v}(r'_m, p')$ and $w = d_{T'}(p, p')$. Finally, let a, b, c and d be species in A, B, C and D respectively. Observe that $1 \leq l, l' \leq 2$ since T_v and T'_v satisfy Condition 1.

Note that if $Solution(v_m)$ is a switch $(U, U_1, \ldots, U_q, U', r_{ON}, r_{OFF})$ then by the Switch Joining Lemma $\{r_m, r'_m\} = \{ON, OFF\}$. Otherwise $Solution(v_m) = \{T_m\}$, and by the Joining Lemma, $d_{T_m}(r_m, r'_m) = 1$. Hence, in either case, $d_{\hat{T_m}}(r_m, r'_m) = 1$.

As in the proof of the Joining Lemma, dependencies among l, l' and w are derived by examining the locations of midpaths $m(a,c)$, $m(a,d)$, $m(b,c)$ and $m(b,d)$. The reader should consult the proof of the Joining Lemma for details on how dependencies among l, l' and w are actually derived. Observe that the midpaths $m(a,c)$, $m(a,d)$, $m(b,c)$ and $m(b,d)$ are located in $T[S(v_m)] \cup \{(v, v_m)\} \cup T[S(v) - S(v_m)]$. A midpath survey of dependencies among l, l' and w resulting from midpath locations appears in Table 3.

	$m(a,c)$	$m(a,d)$	$m(b,c)$	$m(b,d)$
$T[S(v_m)]$	$l = l' + w + 1$	$l + w = l' + 1$	$l + 1 = l' + w$	$l' = l + w + 1$
(v, v_m)	$l + 1 = l' + w$	$l' = l + w + 1$	$l = l' + w + 1$	$l + w = l' + 1$
$T[S(v) - S(v_m)]$	$l = l' + w + 1$	$l + w = l' + 1$	$l + 1 = l' + w$	$l' = l + w + 1$

Table 3. Midpath location and the resulting dependency among l, l', and w.

Case $m(a,d) = (v, v_m)$:
By the midpath survey $l' = l + w + 1$. Since $1 \leq l, l' \leq 2$, it follows that $l' = 2$, $l = 1$ and $w = 0$. Consequently, the only midpath location for $m(a,c)$ that does not imply $w < 0$ (a contradiction) is (v, v_m). In fact, this must be true for all $a \in A$ and $c \in C$. Since $w = 0$, it follows that $p = p'$ and $C = D$, and so, $\hat{T_m}[A]$ and T' are ultrametrics. Furthermore, since $m(a,c) = (v, v_m)$ and $l = 1$, the midpoint of $P_{T_v}(a,c)$ is (r_m, p) by Condition 3 implying that $\hat{T_m}[A]$ and T' are ultrametrics of the same depth.

If $Solution(v_m) = \{T_m\}$ then since $\hat{T_m}[A]$ and T' are ultrametrics of some depth h and T_v and T'_v are ordinally equivalent, $Solution(v)$ is a $(1,h)$–switch $(T_m[B], T_m[A], T', p, z)$ where z is the vertex of T'_v adjacent to p and r'_m. See Figure 9.

Otherwise $Solution(v_m)$ is a (q,h)–switch $T_m = (U, U_1, \ldots, U_q, U', r_{ON}, r_{OFF})$. Above it was argued that $\{r_m, r'_m\} = \{r_{ON}, r_{OFF}\}$. If $r_m \equiv r_{ON}$ and $r'_m \equiv r_{OFF}$ then $\hat{T_m}[A] \equiv U'$ is an ultrametric of depth h implying that T' is an ultrametric of depth h as well. It follows that $Solution(v)$ is a $(q+1,h)$–switch $(U, U_1, \ldots, U_q, U', T', p, z)$ where z is the vertex of T'_v adjacent to p and r'_m. See Figure 9.

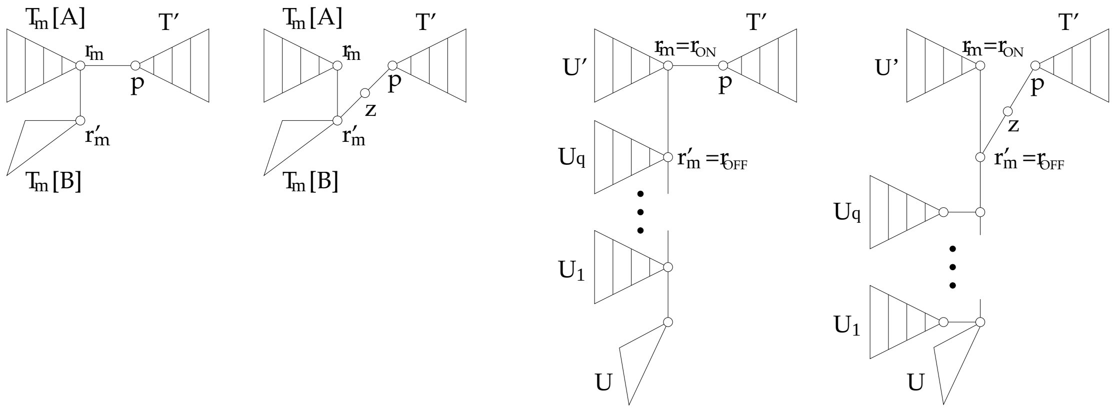

Fig. 9. T_v and T'_v on the left when $Solution(v_m) = \{T_m\}$ and on the right when $Solution(v_m)$ is a (q,h)–switch $T_m = (U, U_1, \ldots, U_q, U', r_{ON}, r_{OFF})$.

If $r_m \equiv r_{OFF}$ and $r'_m \equiv r_{ON}$ then let p_0 denote the root of the ultrametric U. Since $U \subseteq \hat{T_m}[A]$, and $\hat{T_m}[A]$ is an ultrametric, U is an ultrametric of some depth k. Since $d_{\hat{T_m}}(p_0, r_{OFF}) = q$, the depth of $(\hat{T_m}[A], r_{OFF})$ and the depth of (T', p) is $k+q$. We compare the depth of (T', p) to the depth h of the ultrametric U'.

If $k + q > h$ then by the Shallowest Depth Lemma there are species $w_1, w_2 \in T'$ such that $d_{T_v}(w_1, w_2) = d_{T'_v}(w_1, w_2) = 2h + 2$. However, for any $u' \in U'$ and $u_q \in U_q$,

$$d_{T'_v}(u', u_q) = 2h + 1 < d_{T'_v}(w_1, w_2) = 2h + 2 = d_{T_v}(w_1, w_2) < 2h + 3 = d_{T_v}(u', u_q).$$

This contradicts both T_v and T'_v satisfying Condition 2. Note that for the Shallowest Depth Lemma to be applied to T', T' cannot have degree two vertices except, possibly, at its root p. This is indeed the case since $l' = 2$, and so, $P_{T'_v}(r'_m, p)$ has a degree two vertex. Since T'_v satisfies Condition 1, T' has no degree two vertices except, possibly, for p.

If $k + q < h$ then by the Shallowest Depth Lemma there are species $w_1, w_2 \in U'$ such that $d_{T_v}(w_1, w_2) = d_{T'_v}(w_1, w_2) = 2k + 2q + 2$. However, for any $u \in U$ and $s \in T'$,

$$d_{T_v}(u, s) = 2k + 2q + 1 < d_{T_v}(w_1, w_2) = 2h + 2q + 2 = d_{T'_v}(w_1, w_2) < 2k + 2q + 3 = d_{T'_v}(u, s).$$

This contradicts both T_v and T'_v satisfying Condition 2. Note that for the Shallowest Depth Lemma to be applied to U', U' cannot have degree two vertices except, possibly, at its root r_{ON}. This is indeed the case since $l' = 2$, and so, $P_{T'_v}(r'_m, p)$ has a degree two vertex. Since T'_v satisfies Condition 1, U' has no degree two vertices except, possibly, for r_{ON}.

If $k + q = h$ then observe that $d_{T_v}(a, b) = d_{T_v}(a, c)$ but $d_{T'_v}(a, b) < d_{T'_v}(a, c)$. This contradicts both T_v and T'_v satisfying Condition 2.

Case $m(b, c) = (v, v_m)$:
This case is symmetric with the above case where a and d are replaced by b and c.

Case $m(a, d), m(b, c) \in T[S(v_m)] \cup T[S(v) - S(v_m)]$:
By the midpath survey $l + w = l' + 1$ and $l + 1 = l' + w$ implying that $w = 1$ and $l = l'$. Consequently, the only non-contradictory location for $m(a, c)$ and $m(b, d)$ is (v, v_m), for all $a \in A$, $b \in B$, $c \in C$ and $d \in D$. It follows that $T'[C]$ and $T'[D]$ are ultrametrics of some depths h and k, respectively. Observe that, for all $c_1, c_2 \in C$, the midpoint of $P_{T'}(c_1, c_2) = p$. By the midpath–midpoint correspondence (section 3.1), $m(c_1, c_2) = v$ and by the same argument, $m(d_1, d_2) = v$. Suppose $z \neq p \in (T'[C], p)$ has degree two. Let $c_1, c_2 \in C$ such that $p, z \in P_{T'[C]}(p, z)$. Since $z \in exp_{T'[C]}(v)$, $c_1 \in S(v_i)$, $c_2 \in S(v_j)$ such that $i \neq j$. Because $m(c_1, c_2) = v$ and $J = join((T_i, r_i), (T_j, r_j))$ satisfies Condition 4, $d_J(c_1, z) = d_J(c_2, z)$. This contradicts p and z being distinct. It follows that only p can have degree two in $(T'[C], p)$, and by a similar argument, only p' can have degree two in $(T'[D], p')$.

Observe that either $h > k + 1$ or $k > h + 1$ since the midpoint of $P_{T_v}(c, d)$ cannot be p, (p, p') or p' since $d_{T_v}(a, p) < d_{T_v}(a, p')$ yet $d_{T'_v}(a, p') < d_{T'_v}(a, p)$. Assume that $h > k + 1$ without loss of generality.

Suppose p has degree 2 in T'. Since T'_v satisfies Condition 1 it must be that $l = l' = 1$. Since $l' = 1$ and $m(b, d) = (v, v_m)$, by midpath–midpoint correspondence the midpoint of $P_{T'_v}(b, d)$ is $(r'_m, p') = exp_{T'_v}((v, v_m))$. It follows that $d_{\hat{T}_m}(b, r'_m) = k$. Observe that the shallowest depth of (T', p) is $k + 1$, and so, by the Shallowest Depth Lemma there are species $w_1, w_2 \in S(v) - S(v_m)$ such that $d_{T_v}(w_1, w_2) = d_{T'_v}(w_1, w_2) = 2k + 2$. However,

$$d_{T'_v}(b, d) = 2k + 1 < d_{T'_v}(w_1, w_2) = 2k + 2 = d_{T_v}(w_1, w_2) < 2k + 3 = d_{T_v}(b, d)$$

which contradicts both T_v and T'_v satisfying Condition 2.

Suppose p' has degree 2 in T'. Since T_v satisfies Condition 1 it must be that $l = l' = 1$. Observe that the shallowest depth of $(T'[C], p)$ is $h > k + 1$, and so, by the Shallowest Depth Lemma there are species $w_1, w_2 \in C$ such that $d_{T_v}(w_1, w_2) = d_{T'_v}(w_1, w_2) = 2k + 2$. However,

$$d_{T'_v}(b, d) = 2k + 1 < d_{T'_v}(w_1, w_2) = 2k + 2 = d_{T_v}(w_1, w_2) < 2k + 3 = d_{T_v}(b, d)$$

which contradicts both T_v and T'_v satisfying Condition 2.

Finally, suppose T' has no degree two vertices. If $l = l' = 1$ then either of the above arguments can be applied, otherwise, $l = l' = 2$. Observe that the shallowest depth of $(T'[C], p)$ is $h > k + 1$, and so, by the Shallowest Depth Lemma there are species $w_1, w_2 \in C$ such that $d_{T_v}(w_1, w_2) = d_{T'_v}(w_1, w_2) = 2k + 4$. However,

$$d_{T'_v}(b, d) = 2k + 3 < d_{T'_v}(w_1, w_2) = 2k + 4 = d_{T_v}(w_1, w_2) < 2k + 5 = d_{T_v}(b, d)$$

which contradicts both T_v and T'_v satisfying Condition 2. □

The proof of the One–Two Theorem details how switches “grow”. If $Solution(v_i)$ is not a switch for $1 \leq i \leq m$ but $Solution(v)$ is a switch then $Solution(v)$ is a $(1, h)$–switch. That is, a switch has been created. Otherwise, if $Solution(v_m)$ is a (q, h)–switch and $Solution(v)$ is a switch then $Solution(v)$ is a $(q + 1, h)$–switch. Hence, a switch has been extended by appending an ultrametric of depth h.

3.6 Complexity Analysis

Elements of the algorithm are indexed in accordance with the algorithm summary appearing at the end of section 3.3. Let n be the number of species in S.

Let D_{ON} and D_{OFF} be two dimensional arrays indexed by S that are initially empty. As the algorithm progresses intervertex distance values become known and are inserted into these arrays. Since $Solution(v)$ may be a switch $\{T_v[ON], T'_v[OFF]\}$, $D_{ON}[s,t]$ is assigned $d_{T_v[ON]}(s,t)$ and $D_{OFF}[s,t]$ is assigned $d_{T_v[OFF]}(s,t)$, for every $s,t \in S(v)$. Two operations are performed on D_{ON} and D_{OFF}. The first is when $D_{ON}[s,t]$ and $D_{OFF}[s,t]$ are initialized. The second occurs when $Solution(v_i)$ is a switch $\{T_i[ON], T_i[OFF]\}$ and it has been determined that either $T_i[ON]$ or $T_i[OFF]$ is not a subsolution of some $T_v \in Solution(v)$. Assuming the latter, $D_{OFF}[s,t]$ is reset to $D_{ON}[s,t]$ for all $s,t \in S(v_i)$ such that $D_{OFF}[s,t] \neq D_{ON}[s,t]$. Each of these operations can occur just once, hence, the complexity of maintaining D_{ON} and D_{OFF} is $O(n^2)$.

Step 1 of the algorithm requires $O(n^2)$ as a minimal midpath phylogeny can be found in $O(n^2)$ time [14]. The complexity of step 2 is a function of the complexity of determining $Solution(v)$ for each vertex v of T. The while loop of step c iterates $m_v - 1$ times where m_v denotes the number of v's children. The complexity of each iteration is a function of the complexity of computing $Candidates(i,j)$.

Computing $Candidates(i,j)$ requires testing that each $Solution(v_i) \times Solution(v_j)$ join satisfies Conditions 1' and 2 through 4. First we bound the number of $Solution(v_i) \times Solution(v_j)$ joins:

Lemma 6. *For each vertex v in T and $T_v \in Solution(v)$, $|exp_{T_v}(v)| = O(log(n))$.*

Proof. Let $T_v \in Solution(v)$ and assume $|exp_{T_v}(v)| \geq 3$. Let w_1, …, w_m be the vertices of $exp_{T_v}(v)$ that are adjacent to vertices in $T_v - exp_{T_v}(v)$. For each w_i define $S(w_i) = \{s \in S(v) |\ d_{T_v}(s,w_i) < d_{T_v}(s,w_j)$ for all $w_j \neq w_i\}$.

Define x_i to be the shallowest species in $(T_v[S(w_i)], w_i)$ for each w_i. If the midpoint of $P_{T_v}(x_i,x_j)$ is on $P_{T_v}(w_i,w_j)$ then $m(x_i,x_j) = v$ since $P_{T_v}(x_i,x_j) \subseteq exp_{T_v}(v)$. However, by Condition 4, $w_i = w_j$ — a contradiction. It follows that the midpoint of $P_{T_v}(x_i,x_j)$ is in $T_v[S(w_i)] - \{w_i\}$ or is in $T_v[S(w_j)] - \{w_j\}$. Assuming the latter,

$$d_{T_v}(x_i,w_i) + d_{T_v}(w_i,w_j) < d_{T_v}(w_j,x_j). \tag{10}$$

Denote this by $x_i < x_j$ and observe that this results in a total order of x_1, …, x_m. Without loss of generality assume $x_1 < x_2 < \ldots < x_m$.

Since x_m is a shallowest depth species of $T_v[S(w_m)]$, $|S(w_m)| \geq 2^{d_{T_v}(w_m,x_m)-1}$ (note that w_m might have degree two). Employing the above recurrence for $d_{T_v}(w_m,x_m)$ we have by (10)

$$|S(w_m)| \geq 2^{d_{T_v}(w_m,x_m)-1} > 2^{d_{T_v}(w_{m-1},w_m)+d_{T_v}(w_{m-1},x_{m-1})-1}$$

and by repeated application of (10)

$$|S(w_m)| > 2^{d_{T_v}(w_{m-1},w_m)+\ldots+d_{T_v}(w_1,w_2)+d_{T_v}(w_1,x_1)-1} \geq 2^{m-1}. \tag{11}$$

The last inequality follows since $d_{T_v}(w_i,w_{i+1}) \geq 1$ for $1 \leq i \leq m-1$.

Since $n \geq |S(w_m)|$ and $|S(w_m)| > 2^{m-1}$ it follows that $log(n) > m-1$. Furthermore, since $exp_{T_v}(v)$ has m leaves and at most one degree two vertex by Condition 1, $exp_{T_v}(v)$ has no more than $2m-1$ vertices. It follows that $|exp_{T_v}(v)| = O(log(n))$. □

Lemma 6 implies that the number of $Solution(v_i)$ selections and the number of $Solution(v_j)$ selections is $O(log(n))$, and so, the number of $Solution(v_i) \times Solution(v_j)$ joins is $O(log^2(n))$.

Let $J = join((T_i, r_i), (T_j, r_j))$ be a $Solution(v_i) \times Solution(v_j)$ join. The complexity of forming J is the complexity of determining $d_J(r_i, r_j)$ using a midpath test. If the midpath test uses midpath $m(x, y)$ to determine $d_J(r_i, r_j)$ then it takes $O(n)$ time to locate the midpoint of $P_J(x, y)$ and then using D_{ON} and/or D_{OFF} to determine $d_J(r_i, r_j)$ in constant time.

To determine if J satisfies Condition 1' a traversal of J is performed that searches for degree two vertices not internal to $P_J(r_i, r_j)$. This requires $O(n)$ time.

To determine if J satisfies Condition 2 a *consistency check* is performed:

> Given a phylogeny T with species set S and $W \subseteq S \times S$, the *rank table R for T restricted to W* is defined such that $R(d) = \{(s, t) \in W | d_T(s, t) = d\}$ where $0 \leq d \leq n$. An *unrestricted* rank table for T is one where $W = S \times S$. R ranks pairs in W with respect to their distance in T. A rank table R is *ordinally consistent* with M if $M[a, b] \leq M[c, d]$ if and only if $(a, b) \in R(d)$, $(c, d) \in R(d')$ and $d \leq d'$, for all pairs $(a, b), (c, d) \in W$. Clearly T is ordinally consistent with M if and only if the unrestricted rank table for T is ordinally consistent with M.
>
> It is assumed that unrestricted rank tables are maintained for each $T_v \in Solution(v)$, for every $v \in T$. In particular there are rank tables R_{T_i} and R_{T_j} for T_i and T_j, respectively. Since T_i and T_j satisfy Condition 2, their rank tables are ordinally consistent with M.
>
> A rank table R' for J restricted to $S(v_i) \times S(v_j)$ is formed by computing $d_J(s, t)$ for every $s \in S(v_i)$ and $t \in S(v_j)$ and then placing (s, t) into $R'(d_J(s, t))$. Forming R' and verifying that R' is ordinally consistent with M requires $O(|S(v_i)|\ |S(v_j)|)$ time since there are $O(|S(v_i)|\ |S(v_j)|)$ elements in R'. Amortized throughout the algorithm's execution this contributes $O(n^2 log^2(n))$ time as every pair of species is placed in a rank table R' at most $O(log^2(n))$ times.
>
> The unrestricted rank table R_J for J is then the union of R_{T_1}, R_{T_j} and R', that is, $R_J(d) = R_{T_i}(d) \cup R_{T_j}(d) \cup R'(d)$, for $0 \leq d \leq n$. However, to verify that R_J is ordinally consistent with M it suffices to create a rank table R_{ij} for J restricted to W such that for each $0 \leq d \leq n$, W contains a pair from each of $R_{T_i}(d)$, $R_{T_j}(d)$ and $R'(d)$, if such pairs exist. Consequently, $|R_{ij}(d)| \leq 3$ for $0 \leq d \leq n$. It follows that it can be confirmed in $O(n)$ time that R_{ij} is ordinally consistent with M. Hence it can be verified in $O(n)$ time that J satisfies Condition 2.

To determine if J satisfies Conditions 3 and 4 it suffices to verify the location of the midpoint of $P_J(x, y)$ for each midpath $m(x, y)$ such that $m(x, y) \in \{(v, v_i), v, (v, v_j)\}$. The verification for each midpath takes constant time given D_{ON} and D_{OFF}. Furthermore, the verification for each midpath is repeated at most $O(log^2(n))$ times since there are at most $O(log^2(n))$ $Solution(v_i) \times Solution(v_j)$ joins. This complexity amortized throughout the execution of the algorithm contributes $O(n^2 log^2(n))$ to the overall time complexity.

Step f requires that T_v and T'_v be constructed. This requires $O(m_v^2)$ time [4, 20]. It must also be verified that T_v and T'_v satisfy Conditions 1 through 4. However, this adds only a constant factor to the overall complexity analyzed thus far.

Finally, step 3 of the algorithm requires the verification that an element of $Solution(r)$ does not have degree two vertices. This has complexity $O(n)$.

The overall complexity of the algorithm is $O(n^2 + \Sigma_{v \in T} C_v + n^2 log^2(n) + n)$ where the first term is the cost of step 1 as well as the cost of maintaining D_{ON} and D_{OFF}, C_v denotes the cost of computing $Solution(v)$, the third term is the amortized cost of verifying Conditions 3 and 4 for all midpaths as well as the amortized cost of building all R' tables and the fourth term is the cost of step 3.

It remains to determine C_v. By the above analysis $C_v = O(m_v n log^2(n) + m_v^2)$ and so $\Sigma_{v \in T} C_v = O(n^2 log^2(n) + n^2)$ since $\Sigma_{v \in T} m_v = n$. It follows that the overall complexity of the algorithm is $O(n^2 log^2 n)$.

4 Discussion

A weighted ordinal representation T_R of M is a midpath phylogeny of M, and so, T_R is an expansion of the minimal midpath phylogeny T of M. By defining an appropriate linear inequality program where the variables are edge weights, it can be determined in polynomial time if an expansion T' of T can be edge–weighted so that it is a weighted ordinal representation of M. It follows that if the minimal midpath phylogeny T of M has a polynomial number of binary expansions then a weighted ordinal representation of M can be found in polynomial time. Hence, if the minimal midpath phylogeny is nearly binary then weighted ordinal representations can be found efficiently. It remains an open problem as to whether weighted ordinal representations can be found in polynomial time in the unrestricted case.

Here midpath phylogenies have been used only as a means to an end. However, midpath phylogenies are themselves a tree representation of evolutionary data, and so, are of independent interest. Loosely speaking, an ordinal representation is a metrical tree representation of ordinal assertions supported by M whereas a midpath phylogeny is a topological tree representation of ordinal assertions supported by M. Often it is the topology of an evolutionary tree that is of primary interest and not the amount of evolution. The reader is referred to [14] for more results on midpath phylogenies.

This paper is an effort to better understand the relationship between a phylogeny T and the ordinal assertions supported by T. The understanding of this relationship is vital to the development of ordinal methods for phylogenetic inference. To illustrate, consider the relationship between a phylogeny T and its distance matrix d_T: d_T is unique to T, and furthermore, T can be efficiently recovered from d_T. These results, along with the assumption of additivity, motivate the use of distance data for the inference of phylogeny and have led to the development of many inference methods.

In this paper it has been shown that the ordinal assertions supported by d_T are unique to T, furthermore, T can be efficiently recovered from the ordinal assertions supported by d_T. These results indicate that, under the assumption of ordinality, ordinal data can play an important role in the inference of phylogeny. Future work includes the development of ordinal methods for phylogenetic inference.

Acknowledgments

The authors would like to thank Derek Corneil for many helpful suggestions.

References

1. H.-J. Bandelt and A. W. M. Dress, Split decomposition: A new and useful approach to phylogenetic analysis of distance data, *Mol. Phyl. Evol.*, **1** (1992), 242–252.
2. W. A. Beyer, T. F. Smith, M. L. Stein, and S. M. Ulam, A molecular sequence metric and evolutionary trees, *Math. Bios.*, **19** (1974), 9–25.
3. L. L. Cavalli–Sforza and A. W. F. Edwards, Phylogenetic analysis: Models and estimation procedures, *Evolution*, **32** (1967), 550–570.
4. J. C. Culberson and P. Rudnicki, A fast algorithm for constructing trees from distance matrices, *Information Processing Letters*, **30** (1989), 215–220.
5. William H. E. Day, Computational complexity of inferring phylogenies from dissimilarity matrices, *Bull. of Math. Biol.*, **49**:4 (1987), 461–467.
6. M. Farach, S. Kannan and T. Warnow, A robust model for finding evolutionary trees, *Algorithmica*, **13**:1 (1995), 157–179.
7. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, **17** (1981), 368–376.
8. J. Felsenstein, Numerical methods for inferring evolutionary trees, *Quarterly Review of Biology*, **57**:4 (1982), 379–404.
9. W. M. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science*, **155** (1967), 279–284.
10. S. L. Hakimi and S. S. Yau, Distance matrix of a graph and its realizability, *Quarterly of Applied Mathematics*, **22** (1964), 305–317.
11. D. Hillis, J. Huelsenbeck and D. Swofford, Hobgoblin of phylogenetics? *Nature*, **369** (1994), 363–364.
12. T. H. Jukes and C. R. Cantor, Evolution of protein molecules, in *Mammalian Protein Metabolism* (H. N. Munro ed.), Academic Press, New York, 1969.
13. S. Kannan and T. Warnow, Tree reconstruction from partial orders, *SIAM J. Comp.*, **24**:3 (1995), 511–519.
14. P. Kearney, A six–point condition for ordinal matrices, to appear in *Journal of Computational Biology*, (1997).
15. P. Kearney, R. B. Hayward and H. Meijer, Inferring evolutionary trees from ordinal data, *Proceedings of the 8th Annual ACM–SIAM Symposium on Discrete Algorithms*, (1997), 418–426.
16. N. Saitou and M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4** (1987), 406–425.
17. S. Sattath and A. Tversky, Additive similarity trees, *Psychometrika*, **42** (1977), 319–345.
18. R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman, San Francisco, 1963.
19. David L. Swofford, Gary J. Olsen, Peter J. Waddell and David M. Hillis, Phylogenetic Inference, in *Molecular Systematics*, 2nd Edition (David M. Hillis, Craig Moritz and Barbara K. Mable eds.), Sinauer Associates, Sunderland, Massachusetts, 1996.
20. M. S. Waterman, T. F. Smith, M. Singh and W. A. Beyer, Additive evolutionary trees. *J. Theoret. Biol.*, **64** (1977), 199–213.

This article was processed using the LaTeX macro package with LLNCS style