1

THE MINIMUM EVOLUTION DISTANCE-BASED APPROACH TO PHYLOGENETIC INFERENCE

Richard Desper and Olivier Gascuel

Distance algorithms remain among the most popular for reconstructing phylogenies, especially for researchers faced with data sets with large numbers of taxa. Distance algorithms are much faster in practice than character or likelihood algorithms, and least-squares algorithms produce trees that have several desirable statistical properties. The fast Neighbor Joining heuristic has proven to be quite popular with researchers, but suffers somewhat from a lack of a statistical foundation. We show here that the balanced minimum evolution approach provides a robust statistical justification and is amenable to fast heuristics that provide topologies superior among the class of distance algorithms. The aim of this chapter is to present a comprehensive survey of the minimum evolution principle, detailing its variants, algorithms, and statistical and combinatorial properties. The focus is on the balanced version of this principle, as it appears quite well suited for phylogenetic inference, from a theoretical perspective as well as through computer simulations.

1.1 Introduction

In this chapter, we present recent developments in distance-based phylogeny reconstruction. Whereas character-based (parsimony or probabilistic) methods become computationally infeasible as data sets grow larger, current distance methods are fast enough to build trees with thousands of taxa in a few minutes on an ordinary computer. Moreover, estimation of evolutionary distances relies on probabilistic models of sequence evolution, and commonly used estimators derive from the maximum likelihood (ML) principle (see Chapter 2, this volume). This holds for nucleotide and protein sequences, but also for gene order data (see Chapter 13, this volume). Distance methods are thus model based, just like full maximum likelihood methods, but computations are simpler as the starting information is the matrix of pairwise evolutionary distances between taxa instead of the complete sequence set.

Although phylogeny estimation has been practiced since the days of Darwin, in the 1960s the accumulation of molecular sequence data gave unbiased

Oliv: "chap01" — 2005/1/20 — 15:28 — page 1 — #1

sequence characters (in contrast with subjective morphological characters) to build phylogenies, and more sophisticated methods were proposed. Cavalli-Sforza and Edwards [9] and Fitch and Margoliash [19] both used standard least-squares projection theory in seeking an optimal topology. While statistically sound, the least-squares methods have typically suffered from great computational complexity, both because finding optimal edge lengths for a given topology was computationally demanding and because a new set of calculations was needed for each topology. This was simplified and accelerated by Felsenstein[18] in the FITCH algorithm [17], and by Makarenkov and Leclerc [35], but heuristic least-squares approaches are still relatively slow, with time complexity in $O(n^4)$ or more, where n is the number of taxa.

In the late 1980s, distance methods became quite popular with the appearance of the Neighbor Joining algorithm (NJ) of Saitou and Nei [40], which followed the same line as ADDTREE [42], but used a faster pair selection criterion. NJ proved to be considerably faster than least-squares approaches, requiring a computing time in $O(n^3)$. Although it was not clear what criterion NJ optimizes, as opposed to the least-squares method, NJ topologies have been considered reasonably accurate by biologists, and NJ is quite popular when used with resampling methods such as bootstrapping. The value of NJ and related algorithms was confirmed by Atteson [2], who demonstrated that this approach is statistically consistent; that is, the NJ tree converges towards the correct tree when the sequence length increases and when estimation of evolutionary distances is itself consistent. Neighbor Joining has spawned similar approaches that improve the average quality of output trees. BIONJ [21] uses a simple biological model to increase the reliability of the new distance estimates at each matrix reduction step, while WEIGHBOR [5] also improves the pair selection step using a similar model and a maximum-likelihood approach.

The 1990s saw the development of minimum evolution (ME) approaches to phylogeny reconstruction. A minimum evolution approach, as first suggested by Kidd and Sgaramella-Zonta [31], uses two steps. First, lengths are assigned to each edge of each topology in a set of possible topologies by some prescribed method. Second, the topology from the set whose sum of lengths is minimal is selected. It is most common to use a least-squares method for assigning edge length, and Rzhetsky and Nei [39] showed that the minimum evolution principle is statistically consistent when using ordinary least-squares (OLS). However, several computer simulations [11, 24, 33] have suggested that this combination is not superior to NJ at approximating the correct topology. Moreover, Gascuel, Bryant and Denis [25] demonstrated that combining ME with *a priori* more reliable weighted least-squares (WLS) tree length estimation can be inconsistent.

In 2000, Pauplin described a simple and elegant scheme for edge and tree length estimation. We have proposed [11] using this scheme in a new "balanced" minimum evolution principle (BME), and have designed fast tree building algorithms under this principle, which only require $O(n^2 \log(n))$ time and have been implemented in the FASTME software. Furthermore, computer

simulations have indicated that the topological accuracy of FASTME is even greater than that of best previously existing distance algorithms. Recently, we explained [12] this surprising fact by showing that BME is statistically consistent and corresponds to a special version of the ME principle where tree length is estimated by WLS with biologically meaningful weights.

The aim of this chapter is to present a comprehensive survey of the minimum evolution principle, detailing its variants, mathematical properties, and algorithms. The focus is on BME because it appears quite well suited for phylogenetic inference, but we shall also describe the OLS version of ME, since it was a starting point from which BME definitions, properties, and algorithms have been developed. We first provide the basis of tree metrics and of the ME framework (Section 1.2). We describe how edge and tree lengths are estimated from distance data (Section 1.3). We survey the agglomerative approach that is used by NJ and related algorithms and show that NJ greedily optimizes the BME criterion (Section 1.4). We detail the insertion and tree swapping algorithms we have designed for both versions of ME (Section 1.5). We present the main consistency results on ME (Section 1.6) and finish by discussing simulation results, open problems and directions for further research (Section 1.7).

1.2 Tree metrics

We first describe the main definitions, concepts, and results in the study of tree metrics (Sections 1.2.1 to 1.2.5); for more, refer to Barthélemy and Guénoche [4] or Semple and Steel [43]. Next, we provide an alternate basis for tree metrics that is closely related to the BME framework (Section 1.2.6). Finally, we present the rationale behind distance-based phylogenetic inference that involves recovering a tree metric from the evolutionary distance estimates between taxa (Section 1.2.7).

1.2.1 *Notation and basics*

A *graph* is a pair $G = (V, E)$, where V is a set of objects called *vertices* or *nodes*, and E is a set of *edges*, that is, pairs of vertices. A *path* is a sequence $(v_0, v_1, \ldots, v_k)$ such that for all i, $(v_i, v_{i+1}) \in E$. A *cycle* is a path as above with $k > 2$, $v_0 = v_k$ and $v_i \neq v_j$ for $0 \leq i < k$. A graph is *connected* if each pair of vertices, $x, y \in V$ is connected by a path, denoted p_{xy}. A connected graph containing no cycles is a *tree*, which shall be denoted by T.

The degree of a vertex v, $\deg(v)$, is defined to be the number of edges containing v. In a tree, any vertex v with $\deg(v) = 1$ is called a *leaf*. We will use the letter L to denote the set of leaves of a tree. Other vertices are called internal. In phylogenetic trees, internal nodes have degree 3 or more. An internal vertex with degree 3 is said to be *resolved*, and when all the internal vertices of a tree are resolved, the tree is said to be fully resolved.

A *metric* is a function with certain properties on unordered pairs from a set. Suppose X is a set. The function d: $X \times X \to \Re$ (the set of real numbers) is

a metric if it satisfies:

1. $d(x, y) \geq 0$ for all x, y, with equality if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all x, y.
3. For all x, y, and z, $d(x, z) \leq d(x, y) + d(y, z)$.

For the remainder of the chapter, we shall use d_{xy} in place of $d(x, y)$. We will assume that $X = L = [n] = \{1, 2, \ldots, n\}$ and use the notation $\mathrm{Met}(n)$ to denote the set of metrics on $[n]$.

Phylogenies usually have lengths assigned to each edge. When the molecular clock holds [49], these lengths represent the time elapsed between the endpoints of the edge. When (as most often) the molecular clock does not hold, the evolutionary distances no longer represent times, but are scaled by substitution rates (or frequencies of mutational events, for example, inversions with gene order data) and the same holds with edge lengths that correspond to the evolutionary distance between the end points of the edges.

Let $T = (V, E)$ be such a tree, with leaf set L, and with l: $E \to \Re^+$ a length function on E. This function induces a *tree metric* on L: for each pair $x, y \in L$, let p_{xy} be the unique path from x to y in T. We define

$$d^T_{xy} = \sum_{e \in p_{xy}} l(e).$$

Where there is no confusion about the identity of T, we shall use d instead of d^T.

In standard graph theory, trees are not required to have associated length functions on their edge sets, and the word *topology* is used to describe the shape of a tree without regard to edge lengths. For our purposes, we shall reserve the word topology to refer to any unweighted tree, and will denote such a tree with calligraphic script $\mathcal{T}$, while the word "tree" and the notation T shall be understood to refer to a tree topology with a length function associated to its edges.

In evolutionary studies, phylogenies are drawn as branching trees deriving from a single ancestral species. This species is known as the *root* of the tree. Mathematically, a rooted phylogeny is a phylogeny to which a special internal node is added with degree 2 or more. This node is the tree root, and is denoted as r; when r has degree 2, it is said to be resolved.

Suppose there is a length function l: $E \to \Re^+$ defining a tree metric d. Suppose further that all leaves of T are equally distant from r, that is, there exists a constant c such that $d_{xr} = c$ for all leaves x. Then d is a special kind of tree metric called spherical or *ultrametric*. When the molecular clock does not hold, this property is lost, and the tree root cannot be defined in this simple way.

1.2.2 *Three-point and four-point conditions*

Consider an ultrametric d derived from a tree T. Let x, y, and z be three leaves of T. Let xy, xz, and yz be defined to be the least common ancestors of x and y, x and z, and y and z, respectively. Note that $d_{xy} = 2d_{x(xy)}$ and analogous equalities hold for d_{xz} and d_{yz}. Without loss of generality, xy is not ancestral

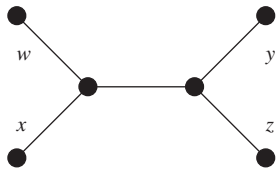

FIG. 1.1. Four-point condition.

to z, and thus $xz = yz$. In this case, $d_{xz} = 2d_{x(xz)} = 2d_{y(yz)} = d_{yz}$. In other words, the two largest of d_{xy}, d_{xz}, and d_{yz} are equal. This can also be written as: for any $x, y, z \in L$,

$$d_{xy} \leq \max\{d_{xz}, d_{yz}\}.$$

This condition is known as the ultrametric inequality or the three-point condition. It turns out [4] that the three-point condition completely characterizes ultrametrics: if d is any metric on any set L satisfying the three-point condition, then there exists a rooted spherical tree T such that $d = d^T$ with L the leaf set of T.

There is a similar characterization of tree metrics in general. Let T be a tree, with tree metric d, and let $w, x, y, z \in L$, the leaf set of T. Without loss of generality, we have the situation in Fig. 1.1, where the path from w to x does not intersect the path from y to z. This configuration implies the (in)equalities:

$$d_{wx} + d_{yz} \leq d_{wy} + d_{xz} = d_{wz} + d_{xy}.$$

In other words, the two largest sums are equal. This can be rewritten as: for all $w, x, y, z \in L$,

$$d_{wx} + d_{yz} \leq \max\{d_{wy} + d_{xz}, d_{wz} + d_{xy}\}.$$

As with the three-point condition, the four-point condition completely characterizes tree metrics [8, 52]. If d is any metric satisfying the four-point condition for all quartets w, x, y, and z, then there is a tree T such that $d = d^T$.

1.2.3 *Linear decomposition into split metrics*

In this section, we consider the algebraic approach to tree metrics. It is common to represent a metric as a symmetric matrix with a null diagonal. Any metric d on the set $[n]$ can be represented as the matrix $\mathbf{D}$ with entries $d_{ij} = d(i, j)$. Let $\mathrm{Sym}(n)$ be the space of symmetric n by n matrices with null diagonals. Note that every metric can be represented by a symmetric matrix, but $\mathrm{Sym}(n)$ also contains matrices with negative entries and matrices that violate the triangle inequality. It is typical to call $\mathrm{Sym}(n)$ the space of *dissimilarity matrices* on $[n]$, and the corresponding functions on $[n]$ are called *dissimilarities*. Let $\mathcal{A}_n$ denote the vector space of dissimilarity functions.

For all $1 \leq i < j \leq n$, let $\mathbf{E}^{(ij)}$ be the matrix with $e^{(ij)}_{ij} = e^{(ij)}_{ji} = 1$, and all other entries equal zero. The set $E = \{\mathbf{E}^{(ij)}\colon 1 \leq i < j \leq n\}$ forms the standard basis for $\mathrm{Sym}(n)$ as a vector space. We shall also express these matrices as vectors

indexed by pairs $1 \leq i < j \leq n$, with $d^{(ij)}$ being a vector with 1 in the (ij) entry, and zero elsewhere. In the following discussion, we will consider other bases for $\mathrm{Sym}(n)$ that have natural relationships to tree metrics.

The consideration of the algebraic structure of tree metrics starts naturally by considering each edge length as an algebraic unit. However, as edges do not have a meaning in the settings of metrics or matrices, our first step is to move from edges to *splits*. A split, roughly speaking, is a bipartition induced by any edge of a tree. Suppose $X \cup Y$ is a non-trivial bipartition of $[n]$; that is, $X \neq \emptyset \neq Y$, and $X \cup Y = [n]$. Such a bipartition is a split, and we will denote it by the notation $X|Y$.

Given the split $X|Y$ of $[n]$, Bandelt and Dress [3] defined the split metric, $\sigma^{X|Y}$ on $[n]$ by

$$\sigma_{ab}^{X|Y} = \begin{cases} 1, & \text{if } |X \cap \{a, b\}| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Any tree topology is completely determined by its splits. Let $e = (x, y)$ be an edge of the topology $\mathcal{T}$. Then define $U_e = \{u \in L : e \in p_{xu}\}$, the set of leaves closer to y than to x, and define $V_e = L \setminus U_e$. We define the set $\mathcal{S}(\mathcal{T})$ to be the set of splits that correspond to edges in $\mathcal{T}$: $\mathcal{S}(\mathcal{T}) = \{U_e \mid V_e : e \in E(\mathcal{T})\}$. For the sake of simplicity, we shall use σ^e to denote $\sigma^{U_e|V_e}$. This set shall prove to be useful as the natural basis for the vector space associated with tree metrics generated by the topology $\mathcal{T}$.

Suppose X is a set of objects contained in a vector space. The vector space *generated by* X, denoted $\langle X \rangle$, is the space of all linear combinations of elements of X. Given a tree topology $\mathcal{T}$, with leaf set $[n]$, let $\mathrm{Met}(\mathcal{T})$ be the set of tree metrics from trees with topology $\mathcal{T}$, and let $\mathcal{A}(\mathcal{T}) = \langle \mathrm{Met}(\mathcal{T}) \rangle$. Any tree metric can be decomposed as a linear sum of split metrics: if d is the metric corresponding to the tree T (of topology $\mathcal{T}$),

$$d = \sum_{e \in E(\mathcal{T})} l^T(e) \sigma^e.$$

Thus $\mathcal{A}(\mathcal{T})$ is a vector space with standard basis $\Sigma(\mathcal{T}) = \{\sigma^e : e \in E(\mathcal{T})\}$.

Note that $\dim \mathcal{A}(\mathcal{T}) = |\Sigma(\mathcal{T})| = |E(\mathcal{T})| \leq 2n - 3$ (with equality when $\mathcal{T}$ is fully resolved), and $\dim \mathcal{A}_n = n(n-1)/2$, and thus for $n > 3$, $\mathcal{A}(\mathcal{T})$ is strictly contained in $\mathcal{A}_n$. Note also that many elements of $\mathcal{A}(\mathcal{T})$ do not define tree metrics, as edge lengths in tree metrics must be non-negative. In fact, the tree metrics with topology $\mathcal{T}$ correspond exactly to the positive cone of $A(\mathcal{T})$, defined by linear combinations of split metrics with positive coefficients.

1.2.4 *Topological matrices*

Let $\mathcal{T}$ be a tree topology with n leaves and m edges, and let $e_1, e_2, \ldots, e_m$ be any enumeration of $E(\mathcal{T})$. Consider the $n(n-1)/2$ by m matrix, $\mathbf{A}^{\mathcal{T}}$, defined by

$$a_{(ij)k}^{\mathcal{T}} = \begin{cases} 1, & \text{if } e_k \in p_{ij}, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose T is a tree of topology $\mathcal{T}$. Let l be the edge length function on E, let $\mathbf{B}$ be the vector with entries $l(e_i)$. Then

$$\mathbf{A}^{\mathcal{T}} \times \mathbf{B} = \mathbf{D}^T,$$

where $\mathbf{D}^T$ is the vector form with entries $d^T_{(ij)}$. This matrix formulation shall prove to be useful as we consider various least-squares approaches to edge length estimation.

1.2.5 *Unweighted and balanced averages*

Given any pair, X, Y, of disjoint subsets of L, and any metric d on L, we use the notation $d_{X|Y}$ to denote the (unweighted) average distance from X to Y under d:

$$d_{X|Y} = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d_{xy}, \tag{1.1}$$

where $|X|$ denotes the number of taxa in the subset X. The average distances shall prove to be useful in the context of solving for optimal edge lengths in a least-squares setting. Given a topology $\mathcal{T}$ with leaf set L, and a metric d on L, it is possible to recursively calculate all the average distances for all pairs A, B of disjoint subtrees of $\mathcal{T}$. If $A = \{a\}$, and $B = \{b\}$, we observe that $d_{A|B} = d_{ab}$. Suppose one of A, B has more than one element. Without loss of generality, B separates into two subtrees B_1 and B_2, as shown in Fig. 1.2, and we calculate

$$d_{A|B} = \frac{|B_1|}{|B|} d_{A|B_1} + \frac{|B_2|}{|B|} d_{A|B_2}. \tag{1.2}$$

It is easy to see that equations (1.1) and (1.2) are equivalent. Moreover, the same equations and notation apply to define $\delta_{A|B}$, that is, the (unweighted) average of distance estimates between A and B.

Pauplin [38] replaced equation (1.2) by a "balanced" average, using 1/2 in place of $|B_1|/|B|$ and $|B_2|/|B|$ for each calculation. Given a topology $\mathcal{T}$, we recursively define $d^{\mathcal{T}}_{A|B}$: if $A = a$, and $B = b$, we similarly define $d^{\mathcal{T}}_{A|B} = d_{ab}$, but

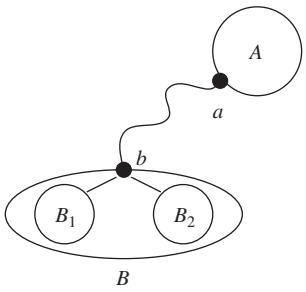

FIG. 1.2. Calculating average distances between subtrees.

if $B = B_1 \cup B_2$ as in Fig. 1.2,

$$d^{\mathcal{T}}_{A|B} = \frac{1}{2} d^{\mathcal{T}}_{A|B_1} + \frac{1}{2} d^{\mathcal{T}}_{A|B_2}. \tag{1.3}$$

For any fully resolved topology $\mathcal{T}$, consideration of these average distances leads us to a second basis for $\mathcal{A}(\mathcal{T})$, which we consider in the next section.

The balanced average uses weights related to the topology $\mathcal{T}$. Let τ_{ab} denote the topological distance (i.e. the number of edges) between taxa a and b, and τ_{AB} the topological distance between the roots of A and B. For any topology $\mathcal{T}$, equation (1.3) leads directly to the identity:

$$d^{\mathcal{T}}_{A|B} = \sum_{a \in A, b \in B} 2^{\tau_{AB} - \tau_{ab}} d_{ab}, \tag{1.4}$$

where

$$\sum_{a \in A, b \in B} 2^{\tau_{AB} - \tau_{ab}} = 1.$$

We thus see that the balanced average distance between a pair of subtrees places less weight on pairs of taxa that are separated by numerous edges; this observation is consistent with the fact that long evolutionary distances are poorly estimated (Section 1.2.7).

1.2.6 *Alternate balanced basis for tree metrics*

The split metrics are not the only useful basis for studying tree metrics. Desper and Vingron [13] have proposed a basis related to unweighted averages, which is well adapted to OLS tree fitting. In this section, we describe a basis related to balanced averages, well suited for balanced length estimation.

Let e be an arbitrary internal edge of any given topology $\mathcal{T}$, and let w, x, y, and z be the four edges leading to subtrees W, X, Y, and Z, as in Fig. 1.3(a). Let B^e be the tree with a length of 2 on e and length $-1/2$ on the four edges w, x, y, and z. Let β^e be the dissimilarity associated to B^e, which is equal to

$$\beta^e = 2\sigma^e - \frac{1}{2}\sigma^w - \frac{1}{2}\sigma^x - \frac{1}{2}\sigma^y - \frac{1}{2}\sigma^z. \tag{1.5}$$

Now consider e as in Fig. 1.3(b), and let B^e be defined to have a length of $\frac{3}{2}$ on e, and a length of $-\frac{1}{2}$ on y and z. Let β^e be the dissimilarity associated with B^e,

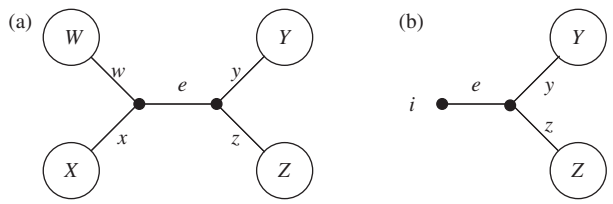

FIG. 1.3. Internal and external edge configurations.

that is,

$$\beta^e = \frac{3}{2}\sigma^e - \frac{1}{2}\sigma^y - \frac{1}{2}\sigma^z. \tag{1.6}$$

Let $\beta^{e'}_{U_e|V_e}$ be the balanced average distance between the sets of the bipartition $U_e \mid V_e$ when the dissimilarity is $\beta^{e'}$, where e' is any edge from $\mathcal{T}$. It is easily seen that

$$\beta^{e'}_{U_e|V_e} = 1 \quad \text{when } e = e', \quad \text{else } \beta^{e'}_{U_e|V_e} = 0. \tag{1.7}$$

Let $\mathcal{B}(\mathcal{T}) = \{\beta^e\colon e \in E(\mathcal{T})\}$. Then $\beta(\mathcal{T})$ is a set of vectors that are mutually independent, as implied by equation (1.7). To prove independence, we must prove that $v = \sum_e c_e\beta^e = 0$ implies $c_e = 0$ for all e. Let e' be any edge of $\mathcal{T}$ and consider the balanced average distance in the e' direction: $v_{U_{e'}|V_{e'}} = \sum_e c_e\beta^e_{U_{e'}|V_{e'}} = c_{e'} = 0$. Thus, $c_{e'} = 0$ for all e', and independence is proven. Since $\mathcal{B}(\mathcal{T})$ is a linearly independent set of the correct cardinality, it forms a basis for $\mathcal{A}(\mathcal{T})$. In other words, any tree metric can be expressed uniquely in the form

$$d = \sum_e d^{\mathcal{T}}_{U_e|V_e}\beta^e, \tag{1.8}$$

which is another useful decomposition of tree metrics. From this decomposition, we see that the length of T is the weighted sum of lengths of the B^es, that is,

$$l(T) = \sum_e d^{\mathcal{T}}_{U_e|V_e} l(B^e).$$

Note that $l(B^e) = 0$ for any internal edge e, while $l(B^e) = 1/2$ for any external edge e. Thus

$$l(T) = \frac{1}{2}\sum_{i\in L} d^{\mathcal{T}}_{\{i\}|L\setminus\{i\}}. \tag{1.9}$$

Returning to the expressions of equation (1.5) and equation (1.6), we can decompose d as

$$\begin{aligned} d = &\sum_{e \text{ external}} d^{\mathcal{T}}_{U_e|V_e}\left(\frac{3}{2}\sigma^e - \frac{1}{2}\sigma^y - \frac{1}{2}\sigma^z\right) \\ &+ \sum_{e \text{ internal}} d^{\mathcal{T}}_{U_e|V_e}\left(2\sigma^e - \frac{1}{2}\sigma^w - \frac{1}{2}\sigma^x - \frac{1}{2}\sigma^y - \frac{1}{2}\sigma^z\right), \end{aligned}$$

that is,

$$\begin{aligned} d = &\sum_{e \text{ external}} \left(\frac{3}{2}d^{\mathcal{T}}_{U_e|V_e} - \frac{1}{2}d^{\mathcal{T}}_{U_y|V_y} - \frac{1}{2}d^{\mathcal{T}}_{U_z|V_z}\right)\sigma^e \\ &+ \sum_{e \text{ internal}} \left(2d^{\mathcal{T}}_{U_e|V_e} - \frac{1}{2}d^{\mathcal{T}}_{U_w|V_w} - \frac{1}{2}d^{\mathcal{T}}_{U_x|V_x} - \frac{1}{2}d^{\mathcal{T}}_{U_y|V_y} - \frac{1}{2}d^{\mathcal{T}}_{U_z|V_z}\right)\sigma^e. \end{aligned} \tag{1.10}$$

Because the representation given by equation (1.8) is unique, equation (1.10) gives us formulae for edge lengths: for internal edges,

$$l(e) = 2d^{\mathcal{T}}_{U_e|V_e} - \frac{1}{2}d^{\mathcal{T}}_{U_w|V_w} - \frac{1}{2}d^{\mathcal{T}}_{U_x|V_x} - \frac{1}{2}d^{\mathcal{T}}_{U_y|V_y} - \frac{1}{2}d^{\mathcal{T}}_{U_z|V_z}, \tag{1.11}$$

and for external edges,

$$l(e) = \frac{3}{2}d^{\mathcal{T}}_{U_e|V_e} - \frac{1}{2}d^{\mathcal{T}}_{U_y|V_y} - \frac{1}{2}d^{\mathcal{T}}_{U_z|V_z}. \tag{1.12}$$

We shall see that these formulae (1.9, 1.11, 1.12) correspond to the estimates found by Pauplin via a different route. We shall also provide another combinatorial interpretation of formula (1.9) due to Semple and Steel [44].

1.2.7 *Tree metric inference in phylogenetics*

Previous sections (1.2.1 to 1.2.6) describe the mathematical properties of tree metrics. Inferring the tree corresponding to a given tree metric is simple. For example, we can use the four-point condition and closely related ADDTREE algorithm [42] to reconstruct the tree topology, and then formulae (1.11) and (1.12) to obtain the edge lengths. However, in phylogenetics we only have evolutionary distance estimates between taxa, which do not necessarily define a tree metric. The rationale of the distance-based approach can thus be summarized as follows [16].

The true Darwinian tree T is unknown but well defined, and the same holds for the evolutionary distance that corresponds to the number of evolutionary events (e.g. substitutions) separating the taxa. This distance defines a tree metric d corresponding to T with positive weights (numbers of events) on edges. Due to hidden (parallel or convergent) events, the true number of events is unknown and greater than or equal to the observed number of events. Thus, the distance-based approach involves estimating the evolutionary distance from the differences we observe today between taxa, assuming a stochastic model of evolution. Such models are described in this volume, in Chapter 2 concerning sequences and substitution events, and in Chapter 13 concerning various genome rearrangement events.

Even when the biological objects and the models vary, the basic principle remains identical: we first compute an estimate Δ of $\mathbf{D}$, the metric associated with T, and then reconstruct an estimate $\hat{T}$ of T using Δ. The estimated distance matrix Δ no longer exactly fits a tree, but is usually very close to a tree. For example, we extracted from TreeBASE (www.treebase.org) [41] 67 Fungi sequences (accession number M520), used DNADIST with default options to calculate a distance matrix, and used NJ to infer a phylogeny. The tree $\hat{T}$ obtained in this (simple) way explains more than 98% of the variance in the distance matrix (i.e. $\sum_{i,j}(\delta_{ij} - d^{\hat{T}}_{ij})^2 / \sum_{i,j}(\delta_{i,j} - \overline{\delta})^2$ is about 2%, where $\overline{\delta}$ is the average value of δ_{ij}). In other words, this tree and the distance matrix are extremely close, and the mere principle of the distance approach appears fully justified in

this case. Numerous similar observations have been made with aligned sequences and substitution models.

In the following, we shall not discuss evolutionary distance estimation, which is dealt with in other chapters and elsewhere (e.g. [49]), but this is clearly a crucial step. An important property that holds in all cases is that estimation of short distances is much more reliable than estimation of long distances. This is simply due to the fact that with long distances the number of hidden events is high and is thus very hard to estimate. As we shall see (Section 1.3.7 and Chapter 13, this volume), this feature has to be taken into account to design accurate inference algorithms. Even if the estimated distance matrix Δ is usually close to a tree, tree reconstruction from such an approximate matrix is much less obvious than in the ideal case where the matrix perfectly fits a tree. The next sections are devoted to this problem, using the minimum evolution principle.

1.3 Edge and tree length estimation

In this section, we consider edge and tree length estimation, given an input topology and a matrix of estimated evolutionary distances. We first consider the least-squares framework (Sections 1.3.1 to 1.3.3), then the balanced approach (Sections 1.3.5 and 1.3.6), and finally show that the latter is a special case of weighted least-squares that is well suited for phylogenetic inference.

For the rest of this section, Δ will be the input matrix, $\mathcal{T}$ the input topology, and $\mathbf{A}$ will refer to the topological matrix $\mathbf{A}^{\mathcal{T}}$. We shall also denote as $\hat{l}$ the length estimator obtained from Δ, $\hat{T}$ the tree with topology $\mathcal{T}$ and edge lengths $\hat{l}(e)$, $\hat{\mathbf{B}}$ the vector of edge length estimates, and $\hat{\mathbf{D}} = (\hat{d}_{ij})$ the distance matrix corresponding to the tree metric $d^{\hat{T}}$. Depending on the context, Δ and $\hat{\mathbf{D}}$ will sometimes be in vector form, that is, $\Delta = (\delta_{(ij)})$ and $\hat{\mathbf{D}} = (\hat{d}_{(ij)})$.

1.3.1 *The least-squares (LS) approach*

Using this notation, we observe that $\hat{\mathbf{D}} = \mathbf{A}\hat{\mathbf{B}}$, and the edge lengths are estimated by minimizing the difference between the observation Δ and $\hat{\mathbf{D}}$. The OLS approach involves selecting edge lengths $\hat{\mathbf{B}}$ minimizing the squared Euclidean fit between Δ and $\hat{\mathbf{D}}$:

$$\mathrm{OLS}(\hat{T}) = \sum_{i,j} (\hat{d}_{ij} - \delta_{ij})^2 = (\hat{\mathbf{D}} - \Delta)^t (\hat{\mathbf{D}} - \Delta).$$

This yields:

$$\hat{\mathbf{B}} = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \Delta. \tag{1.13}$$

However, this approach implicitly assumes that each estimate δ_{ij} has the same variance, a false supposition since large distances are much more variable than short distances (Section 1.2.7). To address this problem, Fitch and Margoliash [19], Felsenstein [18], and others have proposed using a WLS

approach, that is, minimizing

$$\mathrm{WLS}(\hat{T}) = \sum_{i,j} \frac{(\hat{d}_{ij} - \delta_{ij})^2}{v_{ij}} = (\hat{\mathbf{D}} - \Delta)^t \mathbf{V}^{-1} (\hat{\mathbf{D}} - \Delta),$$

where $\mathbf{V}$ is the diagonal $n(n-1)/2 \times n(n-1)/2$ matrix containing the variances v_{ij} of the δ_{ij} estimates. This approach yields

$$\hat{\mathbf{B}} = (\mathbf{A}^t \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^t \mathbf{V}^{-1} \Delta. \tag{1.14}$$

OLS is a special case of WLS, which in turn is a special case of generalized least-squares (GLS) that incorporates the covariances of the δ_{ij} estimates [7, 47]. When the full variance–covariance matrix is available, GLS estimation is the most reliable and WLS is better than OLS. However, GLS is rarely used in phylogenetics, due to its computational cost and to the difficulty of estimating the covariance terms. WLS is thus a good compromise. Assuming that the variances are known and the covariances are zero, equation (1.14) defines the minimum-variance estimator of edge lengths.

Direct solutions of equations (1.13) and (1.14) using matrix calculations requires $O(n^4)$ time. A method requiring only $O(n^3)$ time to solve the OLS version was described by Vach [50]. Gascuel [22] and Bryant and Waddell [6] provided algorithms to solve OLS in $O(n^2)$ time. Fast algorithms for OLS are based on the observation of Vach [50]: If $\hat{T}$ is the tree with edge lengths estimated using OLS equation (1.13), then for every edge e in $E(\hat{T})$ we have:

$$\hat{d}_{U_e|V_e} = \delta_{U_e|V_e}. \tag{1.15}$$

In other words, the average distance between the components of every split is identical in the observation Δ and the inferred tree metric.

1.3.2 *Edge length formulae*

Equation (1.15) provides a system of linear equations that completely determines edge length estimates in the ordinary least squares framework. Suppose we seek to assign a length to the internal edge e shown in Fig. 1.3(a), which separates subtrees W and X from subtrees Y and Z. The OLS length estimate is then [39]:

$$\hat{l}(e) = \frac{1}{2}[\lambda(\delta_{W|Y} + \delta_{X|Z}) + (1-\lambda)(\delta_{W|Z} + \delta_{X|Y}) - (\delta_{W|X} + \delta_{Y|Z})], \tag{1.16}$$

where

$$\lambda = \frac{|W||Z| + |X||Y|}{|W \cup X||Y \cup Z|}. \tag{1.17}$$

If the same way, for external edges (Fig. 1.3(b)) the OLS length estimate is given by

$$\hat{l}(e) = \frac{1}{2}(\delta_{\{i\}|Y} + \delta_{\{i\}|Z} - \delta_{Y|Z}). \tag{1.18}$$

These edge length formulae allow one to express the total length of all edges, that is, the tree length estimate, as a linear sum of average distances between pairs of subtrees.

1.3.3 *Tree length formulae*

A general matrix expression for tree length estimation is obtained from the equations in Section 1.3.1. Letting $\mathbf{1}$ be a vector of 1s, we then have

$$\hat{l}(T) = \mathbf{1}^t(\mathbf{A}^t\mathbf{V}^{-1}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{V}^{-1}\Delta. \tag{1.19}$$

However, using this formula would require heavy computations. Since the length of each edge in a tree can be expressed as a linear sum of averages between the four subtrees incident to the edge (presuming a fully resolved tree), a minor topological change will leave most edge lengths fixed, and will allow for an easy recalculation of the length of the tree. Suppose T is the tree in Fig. 1.3(a) and T' is obtained from T by swapping subtrees X and Y across the edge e, which corresponds to a nearest neighbour interchange (NNI, see Section 1.5 for more details). Desper and Gascuel [11] showed that the difference in total tree lengths (using OLS estimations) can be expressed as

$$\begin{aligned}\hat{l}(T) - \hat{l}(T') =& \frac{1}{2}[(\lambda - 1)(\delta_{W|Y} + \delta_{X|Z}) - (\lambda' - 1)(\delta_{W|X} + \delta_{Y|Z}) \\ & - (\lambda - \lambda')(\delta_{W|Z} + \delta_{X|Y})],\end{aligned} \tag{1.20}$$

where λ is as in equation (1.17), and

$$\lambda' = \frac{|W||Z| + |X||Y|}{|W \cup Y||X \cup Z|}.$$

We shall see in Section 1.5 that equation (1.20) allows for very fast algorithms, both to build an initial tree and to improve this tree by topological rearrangements.

1.3.4 *The positivity constraint*

The algebraic edge length assignments given in Sections 1.3.1 and 1.3.2 have the undesirable property that they may assign negative "lengths" to several of the edges in a tree. Negative edge lengths are frowned upon by evolutionary biologists, since evolution cannot proceed backwards [49]. Moreover, when using a pure least-squares approach, that is, when not only the edge lengths are selected using a least-squares criterion but also the tree topology, allowing for negative edge lengths gives too many degrees of freedom and might result in suboptimal trees using negative edge lengths to produce a low apparent error. Imposing positivity is thus desirable when reconstructing phylogenies, and Kuhner and Felsenstein [32] and others showed that FITCH (a pure LS method) has better topological accuracy when edge lengths are constrained to be non-negative.

Adding the positivity constraint, however, removes the possibility of using matrix algebra (equations 1.13 and 1.14) to find a solution. One might be tempted to simply use matrix algebra to find the optimal solution, and then set negative lengths to zero, but this jury-rigged approach does not provide an

optimal solution to the constrained problem. In fact, the problem at hand is non-negative linear regression (or non-negative least-squares, that is, NNLS), which involves projecting the observation Δ on the positive cone defined by $\mathcal{A}(\mathcal{T})$, instead of on the vector space $\mathcal{A}(\mathcal{T})$ itself as in equations (1.13) and (1.14). In general, such a task is computationally difficult, even when relatively efficient algorithms exist [34]. Several dedicated algorithms have been designed for tree inference, both to estimate the edge lengths for a given tree topology [4] and to incorporate the positivity constraint all along tree construction [18, 26, 29, 35]. But all of these procedures are computationally expensive, with time complexity in $O(n^4)$ or more, mostly due to the supplementary cost imposed by the positivity constraint.

In contrast, minimum evolution approaches do not require the positivity constraint. Some authors have suggested that having negative edges might result in trees with underestimated length, as tree length is obtained by summing edge lengths. In fact, having internal edges with negative lengths tends to give longer trees, as a least-squares fit forces these negative lengths to be compensated for by larger positive lengths on other edges. Trees with negative edges thus tend to be discarded when using the minimum evolution principle. Simulations [22] confirm this and, moreover, we shall see in Section 1.3.5 that the balanced framework naturally produces trees with positive edge lengths without any additional computational cost.

1.3.5 *The balanced scheme of Pauplin*

While studying a quick method for estimating the total tree length, Pauplin [38] proposed to simplify equations (1.16) and (1.18) by using weights $\frac{1}{2}$ and the balanced average we defined in Section 1.2.6. He obtained the estimates for internal edges:

$$\hat{l}(e) = \frac{1}{4}(\delta^{\mathcal{T}}_{W|Y} + \delta^{\mathcal{T}}_{X|Z} + \delta^{\mathcal{T}}_{W|Z} + \delta^{\mathcal{T}}_{W|Y}) - \frac{1}{2}(\delta^{\mathcal{T}}_{W|X} + \delta^{\mathcal{T}}_{Y|Z}), \tag{1.21}$$

and for external edges:

$$\hat{l}(e) = \frac{1}{2}(\delta^{\mathcal{T}}_{\{i\}|Y} + \delta^{\mathcal{T}}_{\{i\}|Z} - \delta^{\mathcal{T}}_{X|Y}). \tag{1.22}$$

Using these formulae, Pauplin showed that the tree length is estimated using the simple formula

$$\hat{l}(T) = \sum_{\{i,j\}\subset L} 2^{1-\tau_{ij}}\delta_{ij}. \tag{1.23}$$

In fact, equations (1.21), (1.22), and (1.23) are closely related to the algebraic framework introduced in Section 1.2.6. Assume that a property dual of Vach's [50] theorem (15) for OLS is satisfied in the balanced settings, that is, for every edge $e \in E(\mathcal{T})$:

$$\hat{d}^{\mathcal{T}}_{U_e|V_e} = \delta^{\mathcal{T}}_{U_e|V_e}.$$

We then obtain from equation (1.8) the following simple expression:

$$\hat{\mathbf{D}} = \sum_e \hat{d}^{\mathcal{T}}_{U_e|V_e}\beta^e = \sum_e \delta^{\mathcal{T}}_{U_e|V_e}\beta^e.$$

As a consequence, equations (1.9), (1.11), and (1.12) can be used as estimators of tree length, internal edge length, and external edge length, respectively, simply by turning the balanced averages of $\mathbf{D}$ into those of Δ, that is, $d^{\mathcal{T}}_{X|Y}$ becomes $\delta^{\mathcal{T}}_{X|Y}$. These estimators are consistent by construction (if $\Delta = \mathbf{D}$ then $\hat{\mathbf{D}} = \mathbf{D}$) and it is easily checked (using equations (1.3) and (1.4)) that these estimators are the same as Pauplin's defined by equations (1.21), (1.22), and (1.23). The statistical properties (in particular the variance) of these estimators are given in Section 1.3.7.

Moreover, we have shown [11] that the balanced equivalent of equation (1.20) is

$$\hat{l}(T) - \hat{l}(T') = \frac{1}{4}(\delta^{\mathcal{T}}_{W|X} + \delta^{\mathcal{T}}_{Y|Z} - \delta^{\mathcal{T}}_{W|Y} - \delta^{\mathcal{T}}_{X|Z}). \tag{1.24}$$

Equation (1.24) implies a nice property about balanced edge lengths.

Suppose we use balanced length estimation to assign edge lengths corresponding to the distance matrix Δ to a number of tree topologies, and consider a tree T such that $\hat{l}(T') > \hat{l}(T)$ for any tree T' that can be reached from T by one nearest neighbour interchange (NNI). Then $\hat{l}(e) > 0$ for every internal edge $e \in T$, and $\hat{l}(e) \geq 0$ for every external edge of T.

The proof of this theorem is obtained using equations (1.24) and (1.21). First, consider an internal edge $e \in T$. Suppose e separates subtrees W and X from Y and Z as in Fig. 1.3(a). Since T is a local minimum under NNI treeswapping, the value of equation (1.24) must be negative, that is, $\delta^{\mathcal{T}}_{W|X} + \delta^{\mathcal{T}}_{Y|Z} < \delta^{\mathcal{T}}_{W|Y} + \delta^{\mathcal{T}}_{X|Z}$. A similar argument applied to the other possible NNI across e leads to the analogous inequality $\delta^{\mathcal{T}}_{W|X} + \delta^{\mathcal{T}}_{Y|Z} < \delta^{\mathcal{T}}_{W|Z} + \delta^{\mathcal{T}}_{W|Y}$. These two inequalities force the value of $\hat{l}(e)$ to be positive according to equation (1.21). Now, suppose there were an external edge e with $\hat{l}(e) < 0$. Referring to equation (1.22), it is easy to see that a violation of the triangle inequality would result, contradicting the metric nature of Δ implied by the commonly used methods of evolutionary distance estimation.

1.3.6 *Semple and Steel combinatorial interpretation*

Any tree topology defines circular orderings of the taxa. A circular ordering can be thought of as a (circular) list of the taxa encountered in order by an observer looking at a planar embedding of the tree. For example (Fig. 1.4), the tree $((1,2),3,(4,5))$ induces the four orderings $(1,2,3,4,5)$, $(1,2,3,5,4)$, $(2,1,3,4,5)$, and $(2,1,3,5,4)$.

As one traverses the tree according to the circular order, one passes along each edge exactly twice—once in each direction. Thus, adding up the leaf-to-leaf distances resulting from all pairs of leaves adjacent in the circular order will yield a sum equal to twice the total length of the tree. For example, using

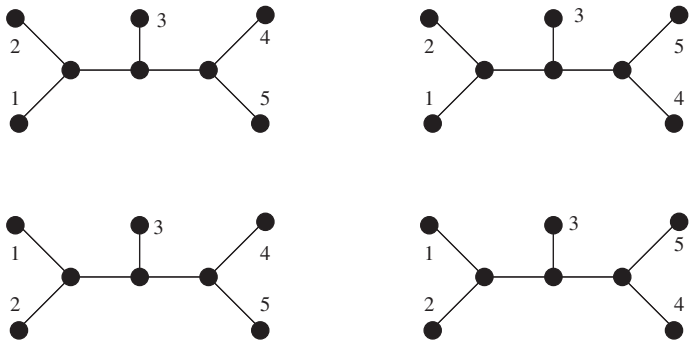

FIG. 1.4. Circular orders of a five-leaf tree.

$(1, 2, 3, 4, 5)$ (which results from the tree in the upper left of Fig. 1.4), we get $l(T) = (d_{12} + d_{23} + d_{34} + d_{45} + d_{51})/2$.

In general, this equality holds for each circular order: given an order $o = (o(1), o(2), \ldots, o(n))$,

$$l(T) = l(d, o) = \frac{1}{2}\left(d_{o(1)o(n)} + \sum_{i=1}^{n-1} d_{o(i)o(i+1)}\right).$$

As we average over $o \in C(T)$, the set of circular orders associated with the tree T, we observe

$$l(T) = \frac{1}{|C(T)|} \sum_{o \in C(T)} l(d, o). \tag{1.25}$$

Semple and Steel [44] have shown that this average is exactly equation (1.9), which becomes Pauplin's formula (1.23) when substituting the d_{ij}s with the δ_{ij} estimates. Moreover, they showed that this result can be generalized to unresolved trees. Let u be any internal node of $\mathcal{T}$, and $\deg(u)$ be the degree of u, that is, 3 or more. Then the following equality holds:

$$l(T) = \sum_{\{i,j\} \subset L} \lambda_{ij} d_{ij}, \tag{1.26}$$

where

$$\begin{aligned} \lambda_{ij} &= \prod_{u \in p_{ij}} (\deg(u) - 1)^{-1}, \quad \text{when } i \neq j, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

1.3.7 *BME: a WLS interpretation*

The WLS approach (equation (1.14)) takes advantage of the variances of the estimates. It is usually hard (or impossible) to have the exact value of these variances, but it is well known in statistics that approximate values are sufficient to obtain reliable estimators. The initial suggestion of Fitch and Margoliash [19], and the default setting in the programs FITCH [18] and PAUP* [48], is to

assume variances are proportional to the squares of the distances, that is, to use $v_{ij} \propto \delta_{ij}^2$. Another common approximation (e.g. [21]) is $v_{ij} \propto \delta_{ij}$. However, numerous studies [7, 36, 37, 47] suggest that variance grows exponentially as a function of evolutionary distance and, for example, Weighbor [5] uses this more suitable approximation.

Desper and Gascuel [12] recently demonstrated that the balanced scheme corresponds to $v_{ij} \propto 2^{\tau_{ij}}$, that is, variance grows exponentially as a function of the topological distance between taxa i and j. Even when topological and evolutionary distances differ, they are strongly correlated, especially when the taxa are homogeneously sampled, and our topology-based approximation is likely capturing most of above-mentioned exponential approximations. Moreover, assuming that the matrix $\mathbf{V}$ is diagonal with $v_{ij} \propto 2^{\tau_{ij}}$, Pauplin's formula (1.23) becomes identical to matrix equation (1.19) and defines the minimum variance tree length estimator. Under this assumption, the edge and tree lengths given by BME are thus as reliable as possible. Since we select the shortest tree, reliability in tree length estimation is of great importance and tends to minimize the probability of selecting a wrong tree. This WLS interpretation then might explain the strong performance of the balanced minimum evolution method.

1.4 The agglomerative approach

In this section, we consider the agglomerative approach to tree building. Agglomerative algorithms (Fig. 1.5) work by iteratively finding pairs of neighbours in the tree, separating them from the rest of the tree, and reducing the size of the problem by treating the new pair as one unit, then recalculating a distance matrix with fewer entries, and continuing with the same approach on the smaller data set.

The basic algorithms in this field are UPGMA (unweighted pair group method using arithmetic averages) [45] and NJ (Neighbor Joining) [40]. The UPGMA algorithm assumes that the distance matrix is approximately ultrametric, while the NJ algorithm does not. The ultrametric assumption allows UPGMA to be quite simple.

1.4.1 *UPGMA and WPGMA*

Given an input distance matrix Δ with entries δ_{ij},

1. Find i, j such that $i \neq j$, δ_{ij} is minimal.
2. Create new node u, connect i and j to u with edges whose lengths are $\delta_{ij}/2$.

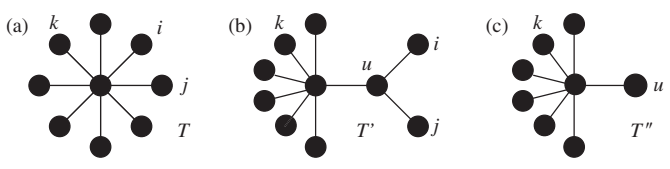

FIG. 1.5. Agglomerative algorithms: (a) find neighbours in star tree; (b) insert new node to join neighbours; (c) continue with smaller star tree.

3. If i and j are the only two entries of Δ, stop and return tree.
4. Else, build a new distance matrix by removing i and j, and adding u, with δ_{uk} defined as the average of δ_{ik} and δ_{jk}, for $k \neq i, j$.
5. Return to Step 1 with smaller distance matrix.

Step 4 calculates the new distances as the average of two distances that have been previously calculated or are original evolutionary distance estimates. In UPGMA, this average is unweighted and gives equal weight to each of the original estimates covered by the i and j clusters, that is, $\delta_{uk} = (|i|\delta_{ik} + |j|\delta_{jk})/(|i| + |j|)$, where $|x|$ is the size of cluster x. In WPGMA the average is weighted (or balanced) regarding original estimates and gives the same weight to each cluster, that is, $\delta_{uk} = (\delta_{ik} + \delta_{jk})/2$. Due to ambiguity (weight of the clusters/weight of the original distance estimates), these two algorithms are often confused for one another and some commonly used implementations of "UPGMA" in fact correspond to WPGMA. In biological studies it makes sense to use a balanced approach such as WPGMA, since a single isolated taxon often gives as much information as a cluster containing several remote taxa [45].

However, the ultrametric (molecular clock) assumption is crucial to Step 1. If Δ is a tree metric but not an ultrametric, the minimal entry might not represent a pair of leaves that can be separated from the rest of the tree as a subtree. To find a pair of neighbours, given only a matrix of pairwise distances, the Neighbor Joining algorithm of Saitou and Nei [40] uses a minimum evolution approach, as we shall now explain.

1.4.2 *NJ as a balanced minimum evolution algorithm*

To select the pair of taxa to be agglomerated, NJ tests each topology created by connecting a taxon pair to form a subtree (Fig. 1.5(b)) and selects the topology with minimal length. As this process is repeated at each step, NJ can be seen as a greedy algorithm minimizing the total tree length, and thus complying with the minimum evolution principle. However, the way the tree length is estimated by NJ at each step is not well understood. Saitou and Nei [40] showed that NJ's criterion corresponds to the OLS length estimation of the topology shown in Fig. 1.5(b), assuming that every leaf (cluster) contains a unique taxon. Since clusters may contain more than one taxon after the first step, this interpretation is not entirely satisfactory. But we shall see that throughout the process, NJ's criterion in fact corresponds to the balanced length of topology as shown in Fig. 1.5(b), which thus implies that NJ is better seen as the natural greedy agglomerative approach to minimize the balanced minimum evolution criterion.

We use for this purpose the general formula (1.26) of Semple and Steel to estimate the difference in length between trees T and T' in Fig. 1.5. Each of the leaves in T and T' is associated to a subtree either resulting from a previous agglomeration, or containing a single, original, taxon that has yet to be agglomerated. In the following, every leaf is associated to a "subtree." Each of these leaf-associated subtrees is binary and identical in T and T', and we can thus define the balanced average distance between any subtree pair, which

has the same value in T and T'. Furthermore, the balanced average distances thus defined correspond to the entries in the current distance matrix, as NJ uses the balanced approach for matrix reduction, just as in WPGMA Step 4. In the following, A and B denote the two subtrees to be agglomerated, while X and Y are two subtrees different from A and B and connected to the central node (Fig. 1.5). Also, let r be the degree of the central node in T, and a, b, x, and y be any original taxa in A, B, X, and Y, respectively.

Using equation (1.26), we obtain:

$$\hat{l}(T) - \hat{l}(T') = \sum_{\{i,j\} \subset L} (\lambda_{ij} - \lambda'_{ij}) \delta_{ij},$$

where the coefficients λ and λ' are computed in T and T', respectively. The respective coefficients differ only when the corresponding taxon pair is not within a single subtree A, B, X, or Y; using this, the above equation becomes:

$$\begin{aligned}\hat{l}(T) - \hat{l}(T') = &\sum_{\{a,b\}} (\lambda_{ab} - \lambda'_{ab}) \delta_{ab} + \sum_{\{a,x\}} (\lambda_{ax} - \lambda'_{ax}) \delta_{ax} \\ &+ \sum_{\{b,x\}} (\lambda_{bx} - \lambda'_{bx}) \delta_{bx} + \sum_{\{x,y\}} (\lambda_{xy} - \lambda'_{xy}) \delta_{xy}.\end{aligned}$$

Using now the definition of the λ's and previous remarks, we have:

$$\begin{aligned}\hat{l}(T) - \hat{l}(T') =&((r-1)^{-1} - 2^{-1}) \delta^T_{AB} + ((r-1)^{-1} - (2(r-2))^{-1}) \\ &\times \sum_X (\delta^T_{AX} + \delta^T_{BX}) + ((r-1)^{-1} - (r-2)^{-1}) \sum_{\{X,Y\}} \delta^T_{XY}.\end{aligned}$$

Letting I and J be any of the leaf-associated subtrees, we finally obtain:

$$\begin{aligned}\hat{l}(T) - \hat{l}(T') = &- 2^{-1} \delta^T_{AB} + 2^{-1} (r-2)^{-1} \left(\sum_{I \neq A} \delta^T_{AI} + \sum_{I \neq B} \delta^T_{BI} \right) \\ &+ ((r-1)^{-1} - (r-2)^{-1}) \sum_{\{I,J\}} \delta^T_{IJ}.\end{aligned}$$

The last term in this expression is independent of A and B, while the first two terms correspond to Studier and Keppler's [46] way of writing NJ's criterion [20]. We thus see that, all through the process, minimizing at each step the balanced length of T' is the same as selecting the pair A, B using NJ's criterion. This proves that NJ greedily optimizes a global (balanced minimum evolution) criterion, contrary to what has been written by several authors.

1.4.3 *Other agglomerative algorithms*

The agglomerative approach to tree metrics was first proposed by Sattath and Tversky [42] in ADDTREE. This algorithm uses the four-point condition (Section 1.2.2) to select at each step the pair of taxa to be agglomerated, and is therefore relatively slow, with time complexity in $O(n^4)$. NJ's $O(n^3)$ was thus

important progress and the speed of NJ, combined with its good topological accuracy, explains its popularity. To improve NJ, two lines of approach were pursued.

The first approach was to explicitly incorporate the variances and covariances of δ_{ij} estimates in the agglomeration scheme. This was first proposed in BIONJ [21], which is based on the approximation $v_{ij} \propto \delta_{ij}$ (Section 1.3.7) and on an analogous model for the covariances; BIONJ uses these (co)variances when computing new distances (Step 4 in algorithm of Section 1.4.1) to have more reliable estimates all along the reconstruction process. The same scheme was used to build the proper OLS version of NJ, which we called UNJ (Unweighted Neighbor Joining) [22], and was later generalized to any variance–covariance matrix of the δ_{ij}s [23]. Weighbor [5] followed the same line but using a better exponential model of the variances [36] and, most importantly, a new maximum-likelihood based pair selection criterion. BIONJ as well as Weighbor then improved NJ thanks to better statistical models of the data, but kept the same agglomerative algorithmic scheme.

The second approach that we describe in the next section involves using the same minimum evolution approach as NJ, but performing a more intensive search of the tree space via topological rearrangement.

1.5 Iterative topology searching and tree building

In this section, we consider rules for moving from one tree topology to another, either by adding a taxon to an existing tree, or by swapping subtrees. We shall consider topological transformations before considering taxon insertion, as selecting the best insertion point is achieved by iterative topological rearrangements. Moreover, we first describe the OLS versions of the algorithms, before their BME counterparts, as the OLS versions are simpler.

1.5.1 *Topology transformations*

The number of unrooted binary tree topologies with n labelled leaves is $(2n-5)!!$, where $k!! = k*(k-2)*\cdots*1$ for k odd. This number grows large far too quickly (close to n^n) to allow for exhaustive topology search except for small values of n. Thus, heuristics are typically relied upon to search the space of topologies when seeking a topology optimal according to any numerical criterion. The following three heuristics are available to users of PAUP* [48]. Tree bisection reconnection (TBR) splits a tree by removing an edge, and then seeks to reconnect the resulting subtrees by adding a new edge to connect some edge in the first tree with some edge in the second tree. Given a tree T, there are $O(n^3)$ possible new topologies that can be reached with one TBR. Subtree pruning regrafting (SPR) removes a subtree and seeks to attach it (by its root) to any other edge in the other subtree. (Note that an SPR is a TBR where one of the new insertion points is identical to the original insertion point.) There are $O(n^2)$ SPR transformations from a given topology. We can further shrink the search space by requiring the new insertion point to be along an edge adjacent to the original insertion

point. Such a transformation is known as an NNI, and there are $O(n)$ NNI transformations from a given topology. Although there are comparatively few NNIs, this type of transformation is sufficient to allow one to move from any binary topology to any other binary topology on the same leaf set simply by a sequence of NNIs.

1.5.2 *A fast algorithm for NNIs with OLS*

Since there are only $O(n)$ NNI transformations from a given topology, NNIs are a popular topology search method. Consider the problem of seeking the minimum evolution tree among trees within one NNI of a given tree. The naive approach would be to generate a set of topologies, and separately solve OLS for each topology. This approach would require $O(n^3)$ computations, because we would run the $O(n^2)$ OLS edge length algorithm $O(n)$ times.

Desper and Gascuel [11] have presented a faster algorithm for simulaneously testing, in $O(n^2)$ time, all of the topologies within one NNI of an initial topology. This algorithm, FASTNNI, is implemented in the program FASTME. Given a distance matrix Δ and a tree topology $\mathcal{T}$:

1. Pre-compute average distances Δ_{avg} between non-intersecting subtrees of $\mathcal{T}$. Initialize $h_{\min} = 0$. Initialize $e_{\min} \in E(T)$.
2. Starting with $e_{\min}$, loop over edges $e \in E(T)$. For each edge e, use equation (1.20) and the matrix Δ_{avg} to calculate $h_1(e)$ and $h_2(e)$, the relative differences in total tree length resulting from each of the two possible NNIs. Let $h(e)$ be the greater of the two. If $h_i(e) = h(e) > h_{\min}$, set $e_{\min} = e$, $h_{\min} = h(e)$, and the indicator variable $s = i$.
3. If $h_{\min} = 0$, stop and exit. Otherwise, perform NNI at $e_{\min}$ in direction pointed to by the variable s.
4. Recalculate entries of Δ_{avg}. Return to Step 2.

Step 1 of FASTNNI can be achieved in $O(n^2)$ time using equation (1.2). Each calculation of equation (1.20) in Step 2 can be done in constant time, and, because there is only one new split in the tree after each NNI, each recalculation of Δ_{avg} in Step 4 can be done in $O(n)$ time. Thus, algorithm requires $O(n^2)$ time to reach Step 2, and an additional $O(n)$ time for each NNI. If s swaps are performed, the total time required is $O(n^2 + sn)$.

1.5.3 *A fast algorithm for NNIs with BME*

The algorithm presented in Section 1.5.2 can be modified to also be used to search for a minimum evolution tree when edges have balanced lengths. The modified algorithm, FASTBNNI, is the same as FASTNNI, with the following exceptions:

1. Instead of calculating the vector of unweighted averages, we calculate the vector $\Delta^{\mathcal{T}}_{\text{avg}}$ of balanced averages.

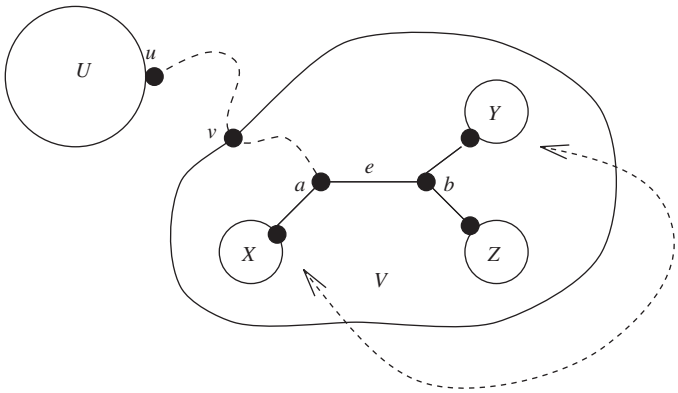

FIG. 1.6. Average calculation after NNI.

2. While comparing the current topology with possible new tree topologies, we use equation (1.24) instead of equation (1.20) to calculate the possible improvement in tree length.
3. Step 3 remains unchanged.
4. Instead of recalculating only averages relating to the new split $WY \mid XZ$, (e.g. $\Delta^{\mathcal{T}}_{WY|U}$ for some set $U \subset X \cup Z$), we also need to recalculate the averages relating to $\Delta^{\mathcal{T}}_{U|V}$ for all splits where U or V is contained in one of the four subtrees W, X, Y, or Z.

As with FASTNNI, Step 1 only requires $O(n^2)$ computations, and Step 2 requires $O(n)$ computations for each pass through the loop. To understand the need for a modification to Step 4, consider Fig. 1.6.

Let us suppose U is a subtree contained in W, and V is a subtree containing X, Y, and Z. Let a, b, u, and v be as in the figure. When making the transition from T to T', by swapping subtrees X and Y, the relative contribution of $\Delta^{\mathcal{T}}_{U|X}$ to $\Delta^{\mathcal{T}'}_{U|V}$ is halved, and the contribution of $\Delta^{\mathcal{T}}_{U|Y}$ is doubled, because Y is one edge closer to U, while X is one edge further away. To maintain an accurate matrix of averages, we must calculate

$$\Delta^{\mathcal{T}'}_{U|V} = \Delta^{\mathcal{T}}_{U|V} + 2^{-2-\tau_{av}}(\Delta^{\mathcal{T}}_{U|Y} - \Delta^{\mathcal{T}}_{U|X}). \tag{1.27}$$

Such a recalculation must be done for each pair U, V, where U is contained in one of the four subtrees and V contains the other three subtrees. To count the number of such pairs, consider tree roots u, v: if we allow u to be any node, then v must be a node along the path from u to e, that is, there are at most $\operatorname{diam}(\mathcal{T})$ choices for v and $n \operatorname{diam}(\mathcal{T})$ choices for the pair (u, v). Thus, each pass through Step 4 will require $O(n \operatorname{diam}(\mathcal{T}))$ computations.

The value of $\operatorname{diam}(\mathcal{T})$ can range from $\log n$ when $\mathcal{T}$ is a balanced binary tree to n when $\mathcal{T}$ is a "caterpillar" tree dominated by one central path. If we

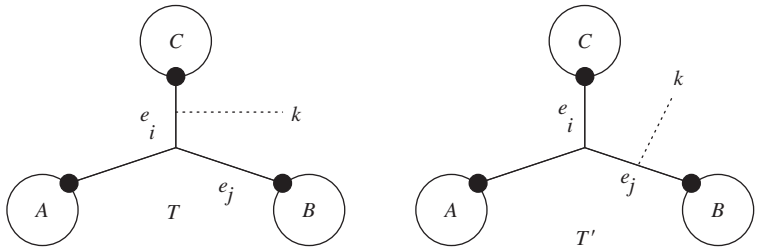

FIG. 1.7. Inserting a leaf into a tree; T' is obtained from T by NNI of k and A.

select a topology from the uniform distribution on the space of binary topologies, we would expect $\text{diam}(\mathcal{T}) = O(\sqrt{n})$, while the more biologically motivated Yule-Harding distribution [28, 51] on the space of topologies would lead to an expected diameter in $O(\log n)$. Thus, s iterations of FASTBNNI would require $O(n^2 + sn \log n)$ computations, presuming a tree with a biologically realistic diameter.

1.5.4 *Iterative tree building with OLS*

In contrast to the agglomerative scheme, many programs (e.g. FITCH, PAUP*, FASTME) use an approach iteratively adding leaves to a partial tree. Consider Fig. 1.7. The general approach is:

1. Start by constructing T_3, the (unique) tree with three leaves.
2. For $k = 4$ to n,
   (a) Test each edge of T_{k-1} as a possible insertion point for the taxon k.
   (b) Based on optimization criterion (e.g. sum of squares, minimum evolution), select the optimal edge $e = (u, v)$.
   (c) Form tree T_k by removing e, adding a new node w, and edges (u, w), (v, w), and (w, k).
3. (Optional) Search space of topologies closely related to T_n using operations such NNIs or global tree swapping.

Insertion approaches can vary in speed from very fast to very slow, depending on the amount of computational time required to test each possible insertion point, and on how much post-processing topology searching is done. The naive approach would use any $O(n^2)$ algorithm to recalculate the OLS edge lengths for each edge in each test topology. This approach would take $O(k^2)$ computations for each edge, and thus $O(k^3)$ computations for each pass through Step 2(a). Summing over k, we see that the naive approach would result in a slow $O(n^4)$ algorithm.

The FASTME program of Desper and Gascuel [11] requires only $O(k)$ computations on Step 2(a) to build a greedy minimum evolution tree using OLS edge lengths. Let Δ be the input matrix, and Δ^k_{avg} be the matrix of average distances

between subtrees in T_k.

1. Start by constructing T_3, the (unique) tree with three leaves; initialize Δ^3_{avg}, the matrix of average distances between all pairs of subtrees in T_3.
2. For $k = 4$ to n,
   (a) We first calculate $\delta_{\{k\}|A}$, for each subtree A of T_{k-1}.
   (b) Test each edge $e \in T_{k-1}$ as a possible insertion point for k.
      i. For all $e \in E$, we will let $f(e)$ to be the cost of inserting k along the edge e.
      ii. Root T_{k-1} at r, an arbitrary leaf, let e_r be the edge incident to r.
      iii. Let $c_r = f(e_r)$, a constant we will leave uncalculated.
      iv. We calculate $g(e) = f(e) - c_r$ for each edge e. Observe $g(e_r) = 0$. Use a top–down search procedure to loop over the edges of T_{k-1}. Consider $e = e_j$, whose parent edge is e_i (see Fig. 1.7). Use equation (1.20) to calculate $g(e_j) - g(e_i)$. (This is accomplished by substituting A, B, C, and $\{k\}$ for W, X, Y, and Z, respectively.) Since $g(e_i)$ has been recorded, this calculation gives us $g(e_j)$.
      v. Select $e_{\min}$ such that $g(e_{\min})$ is minimal.
   (c) Form T_k by breaking $e_{\min}$, adding a new node w_k and edges connecting w_k to the vertices of $e_{\min}$ and to k. Update the matrix Δ^k_{avg} to include average distances in T_k between all pairs of subtrees separated by at most three edges.
3. FASTNNI post-processing (Section 1.5.2).

Let us consider the time complexity of this algorithm. Step 1 requires constant time. Step 2 requires $O(k)$ time in 2(a), thanks to equation (1.2), constant time for each edge considered in 2(b)iv for a total of $O(k)$ time, and $O(k)$ time for 2(c). Indeed, updating Δ^k_{avg} from $\Delta^{k-1}_{\text{avg}}$ only requires $O(k)$ time because we do not update the entire matrix. Thus T_k can be created from T_{k-1} in $O(k)$ time, which leads to $O(n^2)$ computations for the entire construction process. Adding Step 3 leads to a total cost of $O(n^2 + sn)$, where s is the number of swaps performed by FASTNNI from the starting point T_n.

1.5.5 *From OLS to BME*

Just as FASTBNNI is a slight variant of the FASTNNI algorithm for testing NNIs, we can easily adapt the greedy OLS taxon-insertion algorithm of Section 1.5.4 to greedily build a tree, using balanced edge lengths instead of OLS edge lengths. The only differences involve calculating balanced averages instead of unweighted averages.

1. In Step 2(a), we calculate $\delta^{\mathcal{T}_{k-1}}_{\{k\}|A}$ instead of $\delta_{\{k\}|A}$, using equation (1.3).
2. In Step 2(b)iv, we use equation (1.24) instead of equation (1.20) to calculate $g(e_j)$.
3. In Step 2(c), we need to calculate $\delta^{\mathcal{T}_k}_{X|Y}$ for each subtree X containing k, and each subtree Y disjoint from X.

4. Instead of FASTNNI post-processing, we use FASTBNNI post-processing.

The greedy balanced insertion algorithm is a touch slower than its OLS counterpart. The changes to Step 2(a) and 2(b) do not increase the running time, but the change to Step 2(c) forces the calculation of $O(k \operatorname{diam}(T_k))$ new average distances. With the change to FASTBNNI, the total cost of this approach is $O(n^2 \operatorname{diam}(T) + sn \operatorname{diam}(T))$ computations, given s iterations of FASTBNNI. Simulations [11] suggest that $s \ll n$ for a typical data set; thus, one could expect a total of $O(n^2 \log n)$ computations on average.

1.6 Statistical consistency

Statistical consistency is an important and desired property for any method of phylogeny reconstruction. Statistical consistency in this context means that the phylogenetic tree output by the algorithm in question converges to the true tree with correct edge lengths, when the number of sites increases and when the model used to estimate the evolutionary distances is the correct one. Whereas the popular character-based parsimony method has been shown to be statistically inconsistent in some cases [15], many popular distance methods have been shown to be statistically consistent. We first discuss positive results with the OLS and balanced versions of the minimum evolution principle, then provide negative results, and finally present the results of Atteson [2] that provide a measure of the convergence rate of NJ and related algorithms.

1.6.1 *Positive results*

A seminal paper in the field of minimum evolution is the work of Rzhetsky and Nei [39], demonstrating the consistency of the minimum evolution approach to phylogeny estimation, when using OLS edge lengths. Their proof was based on this idea: if T is a weighted tree of topology $\mathcal{T}$, and if the observation Δ is equal to d^T (i.e. the tree metric induced by T), then for any wrong topology $\mathcal{W}$, $\hat{l}(\mathcal{W}) > \hat{l}(\mathcal{T}) = l(T)$. In other words, T is the shortest tree and is thus the tree inferred using the ME principle. Desper and Gascuel [12] have used the same approach to show that the balanced minimum evolution method is consistent.

The circular orders of Section 1.3.6 lead to an easy proof of the consistency of BME (first discussed with David Bryant and Mike Steel). Assume $\Delta = d^T$ and consider any wrong topology $\mathcal{W}$. Per Section 1.3.6, we use $C(\mathcal{W})$ to denote the set of circular orderings of $\mathcal{W}$, and let $\hat{l}(\Delta, o, \mathcal{W})$ be the length estimate of $\mathcal{W}$ from Δ under the ordering o for $o \in C(\mathcal{W})$. The modified version of equation (1.25) yields the balanced length estimate of $\mathcal{W}$:

$$\hat{l}(\mathcal{W}) = \frac{1}{|C(\mathcal{W})|} \sum_{o \in C(\mathcal{W})} \hat{l}(\Delta, o, \mathcal{W}).$$

If $o \in C(\mathcal{W}) \cap C(\mathcal{T})$, then $\hat{l}(\Delta, o, \mathcal{W}) = \hat{l}(\Delta, o, \mathcal{T}) = l(T)$. If $o \in C(\mathcal{W}) \setminus C(\mathcal{T})$, then some edges of T will be double counted in the sum producing $\hat{l}(\Delta, o, \mathcal{W})$.

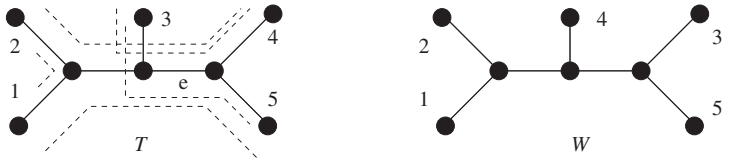

FIG. 1.8. Wrong topology choice leads to double counting edge lengths.

For example, if $\mathcal{T}$ and $\mathcal{W}$ are as shown in Fig. 1.8, and $o = (1, 2, 4, 3, 5) \in C(\mathcal{W}) \setminus C(\mathcal{T})$, then $\hat{l}(\Delta, o, \mathcal{W})$ (represented in Fig. 1.8 by dashed lines) counts the edge e twice. It follows that $\hat{l}(\mathcal{W}) > \hat{l}(\mathcal{T})$.

1.6.2 *Negative results*

Given the aforementioned proofs demonstrating the statistical consistency of the minimum evolution approach in selected settings, it is tempting to hope that minimum evolution would be a consistent approach for any least-squares estimation of tree length. Having more reliable tree length estimators, for example, incorporating the covariances of the evolutionary distance estimates, would then yield better tree inference methods based on the ME principle. Sadly, we have shown [25] that this is not always the case. Using a counter-example we showed that the ME principle can be inconsistent even when using WLS length estimation, and this result extends to various definitions of tree length, for example, only summing the positive edge length estimates while discarding the negative ones. However, our counter-example for WLS length estimation was artificial in an evolutionary biology context, and we concluded, "It is still conceivable that minimum evolution combined with WLS good practical results for realistic variance matrices." Our more recent results with BME confirm this, as BME uses a special form of WLS estimation (Section 1.3.7) and performs remarkably well in simulations [12].

On the other hand, in reference [25] we also provided a very simple 4-taxon counter-example for GLS length estimation, incorporating the covariances of distance estimates (in contrast to WLS). Variances and covariances in this counter-example were obtained using a biological model [36], and were thus fully representative of real data. Using GLS length estimation, all variants of the ME principle were shown to be inconsistent with this counter-example, thus indicating that any combination of GLS and ME is likely a dead end.

1.6.3 *Atteson's safety radius analysis*

In this section, we consider the question of algorithm consistency, and the circumstances under which we can guarantee that a given algorithm will return the correct topology $\mathcal{T}$, given noisy sampling of the metric d^T generated by some tree T with topology $\mathcal{T}$. As we shall see, NJ, a simple agglomerative heuristic approach based on the BME, is optimal in a certain sense, while more sophisticated algorithms do not possess this particular property.

Given two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ of identical dimensions, some standard measures of the distance between them include the L_p norms. For any real value of $p \geq 1$, the L_p distance between $\mathbf{A}$ and $\mathbf{B}$ is defined to be

$$\|\mathbf{A} - \mathbf{B}\|_p = \left(\sum_{i,j} (a_{ij} - b_{ij})^p \right)^{1/p}.$$

For $p = 2$, this is the standard Euclidean distance, and for $p = 1$, this is also known as the "taxi-cab" metric. Another related metric is the L_∞ norm, defined as

$$\|\mathbf{A} - \mathbf{B}\|_\infty = \max_{i,j} |a_{ij} - b_{ij}|.$$

A natural question to consider when approaching the phylogeny reconstruction problem is: given a distance matrix Δ, is it possible to find the tree T such that $\|d^T - \Delta\|_p$ is minimized? Day [10] showed that this problem is NP-hard for the L_1 and L_2 norms. Interestingly, Farach *et al.* [14] provided an algorithm for solving this problem in polynomial time for the L_∞ norm, but for the restricted problem of ultrametric approximation (i.e. $\|d^T - \Delta\|_\infty$ is minimized over the space of ultrametrics). Agarwala *et al.* [1] used the ultrametric approximation algorithm to achieve an approximation algorithm for the L_∞ norm: if $\epsilon = \min_T \|d^T - \Delta\|_\infty$, where d^T ranges over all tree metrics, then the single pivot algorithm of Agarwala *et al.* produces a tree T' whose metric $d^{T'}$ satisfies $\|d^{T'} - \Delta\|_\infty \leq 3\epsilon$.

The simplicity of the L_∞ norm also allows for relatively simple analysis of how much noise can be in a matrix Δ that is a sample of the metric d^T while still allowing accurate reconstruction of the tree T. We define the *safety radius* of an algorithm to be the maximum value ρ such that, if e is the shortest edge in a tree T, and $\|\Delta - d^T\|_\infty < \rho\, l(e)$, then the algorithm in question will return a tree with the same topology as T.

It is immediately clear that no algorithm can have a safety radius greater than $\frac{1}{2}$: consider the following example from [2]. Suppose $e \in T$ is an internal edge with minimum length $l(e)$. Let W, X, Y, and Z be four subtrees incident to e, such that W and X are separated from Y and Z, as in Fig. 1.9. Let d be a metric:

$$\begin{aligned}
d_{ij} &= d^T_{ij} - \frac{l(e)}{2}, && \text{if } i \in W, j \in Y \quad \text{or} \quad i \in X, j \in Z, \\
d_{ij} &= d^T_{ij} + \frac{l(e)}{2}, && \text{if } i \in W, j \in X \quad \text{or} \quad i \in Y, j \in Z, \\
d_{ij} &= d^T_{ij}, && \text{otherwise.}
\end{aligned}$$

d is graphically realized by the network N in Fig. 1.9, where the edge e has been replaced by two pairs of parallel edges, each with a length of $l(e)/2$.

Moreover, consider the tree T' which we reach from T by a NNI swapping X and Y, and keeping the edge e with length $l(e)$. Then it is easily seen that

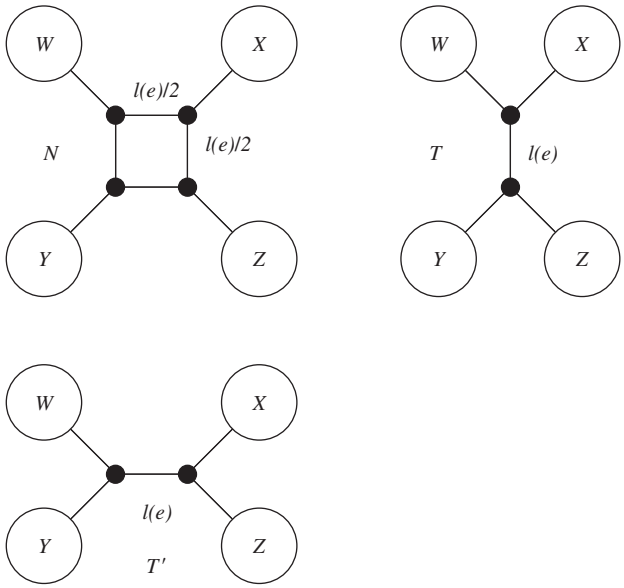

FIG. 1.9. Network metric equidistant from two tree metrics.

$\|d^T - d\|_\infty = l(e)/2 = \|d^{T'} - d\|_\infty$. Since d is equidistant to d^T and $d^{T'}$, no algorithm could guarantee finding the correct topology, if d is the input metric.

Atteson [2] proved that NJ achieves the best possible safety radius, $\rho = \frac{1}{2}$. If d^T is a tree metric induced by T, Δ is a noisy sampling of d^T, and $\epsilon = \max_{i,j} |d^T_{ij} - \delta_{ij}|$, then NJ will return a tree with the same topology as T, providing all edges of T are longer than 2ϵ. In fact, this result was proven for a variety of NJ related algorithms, including UNJ, BIONJ, and ADDTREE, and is a property of the agglomerative approach, when this approach is combined with NJ's (or ADDTREE's) pair selection criterion. An analogous optimality property was recently shown concerning UPGMA and related agglomerative algorithms for ultrametric tree fitting [27]. In contrast, the 3-approximation algorithm only has been proven to have a safety radius of $\frac{1}{8}$.

1.7 Discussion

We have provided an overview of the field of distance algorithms for phylogeny reconstruction, with an eye towards the balanced minimum evolution approach. The BME algorithms are very fast—faster than Neighbor Joining and sufficiently fast to quickly build trees on data sets with thousands of taxa. Simulations [12] have demonstrated superiority of the BME approach, not only in speed, but also in the quality of output trees. Topologies output by FASTME using the balanced minimum evolution scheme have been shown to be superior to those produced by BIONJ, WEIGHBOR, and standard WLS (e.g. FITCH or PAUP*), even though FASTME requires considerably less time to build them.

The balanced minimum evolution scheme assigns edge lengths according to a particular WLS scheme that appears to be biologically realistic. In this scheme, variances of distance estimates are proportional to the exponent of topological distances. Since variances have been shown to be proportional to the exponent of evolutionary distances in the Jukes and Cantor [30] and related models of evolution [7], this model seems reasonable as one expects topological distances to be linearly related to evolutionary distances in most data sets.

The study of cyclic permutations by Semple and Steel [44] provides a new proof of the validity of Pauplin's tree length formula [38], and also leads to a connection between the balanced edge length scheme and Neighbor Joining. This connection, and the WLS interpretation of the balanced scheme, may explain why NJ's performance has traditionally been viewed as quite good, in spite of the fact that NJ had been thought to not optimize any global criterion. The fact that FASTME itself more exhaustively optimizes the same WLS criterion may explain the superiority of the balanced approach over other distance algorithms.

There are several mathematical problems remaining to explore in studying balanced minimum evolution. The "safety radius" of an algorithm has been defined [2] to be the number ρ such that, if the ratio of the maximum measurement error over minimum edge length is less than ρ, then the algorithm will be guaranteed to return the proper tree. Although we have no reason to believe BME has a small safety radius, the exact value of its radius has yet to be determined. Also, though the BME approach has been proven to be consistent, the consistency and safety radius of the BME heuristic algorithms (e.g. FASTBNNI and the greedy construction of Section 1.5.5) have to be determined. Finally, there remains the question of generalizing the balanced approach—in what settings would this be meaningful and useful?

Acknowledgements

O.G. was supported by ACI IMPBIO (Ministère de la Recherche, France) and EPML 64 (CNRS-STIC). The authors thank Katharina Huber and Mike Steel for their helpful comments during the writing of this chapter.

References

[1] Agarwala, R., Bafna, V., Farach, M., Paterson, M., and Thorup, M. (1999). On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing*, **28**(3), 1073–1085.

[2] Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, **25**(2–3), 251–278.

[3] Bandelt, H. and Dress, A. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, **1**, 242–252.

[4] Barthélemy, J.-P. and Guénoche, A. (1991). *Trees and Proximity Representations*. Wiley, New York.

[5] Bruno, W.J., Socci, N.D., and Halpern, A.L. (2000). Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, **17**(1), 189–197.

[6] Bryant, D. and Waddell, P. (1998). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution*, **15**, 1346–1359.

[7] Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, **8**, 868–883.

[8] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences* (ed. F.R. Hodson *et al.*), pp. 387–395. Edinburgh University Press, Edinburgh.

[9] Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis, models and estimation procedures. *Evolution*, **32**, 550–570.

[10] Day, W.H.E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, **49**, 461–467.

[11] Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, **9**, 687–705.

[12] Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, **21**, 587–598.

[13] Desper, R. and Vingron, M. (2002). Tree fitting: Topological recognition from ordinary least-squares edge length estimates. *Journal of Classification*, **19**, 87–112.

[14] Farach, M., Kannan, S., and Warnow, T. (1995). A robust model for finding optimal evolutionary trees. *Algorithmica*, **13**, 155–179.

[15] Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **22**, 240–249.

[16] Felsenstein, J. (1984). Distance methods for inferring phylogenies: A justification. *Evolution*, **38**, 16–24.

[17] Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics*, **5**, 164–166.

[18] Felsenstein, J. (1997). An alternating least-squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, **46**, 101–111.

[19] Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, **155**, 279–284.

[20] Gascuel, O. (1994). A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology and Evolution*, **11**, 961–961.

[21] Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**(7), 685–695.

[22] Gascuel, O. (1997). Concerning the NJ algorithm and its unweighted version, UNJ. In *Mathematical Hierarchies and Biology* (ed. B. Mirkin, F. McMorris, F. Roberts, and A. Rzetsky), pp. 149–170. American Mathematical Society, Providence, RI.

[23] Gascuel, O. (2000). Data model and classification by trees: The minimum variance reduction (MVR) method. *Journal of Classification*, **19**(1), 67–69.

[24] Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular Biology and Evolution*, **17**(3), 401–405.

[25] Gascuel, O., Bryant, D., and Denis, F. (2001). Strengths and limitations of the minimum evolution principle. *Systematic Biology*, **50**(5), 621–627.

[26] Gascuel, O. and Levy, D. (1996). A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *Journal of Classification*, **13**, 129–155.

[27] Gascuel, O. and McKenzie, A. (2004). Performance analysis of hierarchical clustering algorithms. *Journal of Classification*, **21**, 3–18.

[28] Harding, E.F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, **3**, 44–77.

[29] Hubert, L.J. and Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology*, **48**, 281–317.

[30] Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. Munro), pp. 21–132. Academic Press, New York.

[31] Kidd, K.K. and Sgaramella-Zonta, L.A. (1971). Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics*, **23**, 235–252.

[32] Kuhner, M.K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal rates. *Molecular Biology and Evolution*, **11**(3), 459–468.

[33] Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. *Molecular Biology and Evolution*, **13**(4), 584–593.

[34] Lawson, C.M. and Hanson, R.J. (1974). *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, NJ.

[35] Makarenkov, V. and Leclerc, B. (1999). An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification*, **16**, 3–26.

[36] Nei, M. and Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution*, **6**, 290–300.

[37] Nei, M., Stephens, J.C., and Saitou, N. (1985). Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular date from humans and apes. *Molecular Biology and Evolution*, **2**(1), 66–85.

[38] Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, **51**, 41–47.

[39] Rzhetsky, A. and Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, **10**(5), 1073–1095.

[40] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

[41] Sanderson, M.J., Donoghue, M.J., Piel, W., and Eriksson, T. (1994). TreeBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany*, **81**(6), 183.

[42] Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, **42**, 319–345.

[43] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press, New York.

[44] Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, **32**, 669–680.

[45] Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*, pp. 230–234. W.K. Freeman and Company, San Francisco, CA.

[46] Studier, J.A. and Keppler, K.J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, **5**(6), 729–731.

[47] Susko, E. (2003). Confidence regions and hypothesis tests for topologies using generalized least squares. *Molecular Biology and Evolution*, **20**(6), 862–868.

[48] Swofford, D. (1996). PAUP—Phylogenetic Analysis Using Parsimony (and other methods), version 4.0.

[49] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogenetic inference. In *Molecular Systematics* (ed. D. Hillis, C. Moritz, and B. Mable), Chapter 11, pp. 407–514. Sinauer, Sunderland, MA.

[50] Vach, W. (1989). Least squares approximation of addititve trees. In *Conceptual and Numerical Analysis of Data* (ed. O. Opitz), pp. 230–238. Springer-Verlag, Berlin.

[51] Yule, G.U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. *Philosophical Transactions of the Royal Society of London, Series B*, **213**, 21–87.

[52] Zaretskii, K. (1965). Constructing a tree on the basis of a set of distances between the hanging vertices. In Russian, *Uspeh Mathematicheskikh Nauk*, **20**, 90–92.