Chapter 5

Distance Methods

The next class of methods for inferring phylogenetic trees that we will discuss are *distance methods*.

Roughly put, a *distance* is some numerical measure of similarity of two taxa. A distance of 0 should indicate the two taxa are the same, and a larger number signifies the degree of their 'differentness.' Although a distance could be based on something other than a comparison of genetic sequences, in our examples that will be its origin. Given a collection of taxa, we somehow compute distances between each pair. We then attempt to find a metric tree so that distances between taxa on it matches our collection of distances. For a variety of reasons, it is seldom possible to do this exactly, so a key issue is how we deal with inexact fit.

We primarily focus on fast algorithmic approaches to using distances to build a single tree, though we will mention some other approaches that instead choose a 'best' tree according to some criterion.

5.1 Dissimilarity Measures

Before beginning, we need to clarify our terminology. In the introductory paragraph above, we actually used the word 'distance' in two different ways. First, it was the measure of differentness between taxa, and second it was a measurement along branches of a metric tree that described the relationships of the taxa. The first of these should be thought of as coming from data, and the second from some processing of the data that yields an inferred metric tree. When necessary to avoid confusion between these two distinct meanings, we'll call the first a *dissimilarity*, and the second a *tree metric* (as defined in Section 2.3). However it is very common for the word 'distance' to be used for both, so we will also use that term when it either seems unlikely to be confusing, or its use in some special terminology is standard practice.

Definition. A *dissimilarity* map for the set X of taxa is a function $\delta : X \times X \to \mathbb{R}$ such that $\delta(x,x) = 0$ and $\delta(x,y) = \delta(y,x)$.

In our applications, we'll always use dissimilarities with non-negative values: $\delta(x, y) \geq 0$.

If we already knew a metric tree relating the taxa in X, then the tree metric arising from it as defined in Chapter 2 would give a dissimilarity measure on X. Indeed, Proposition 5 shows it has the required properties, and additional ones as well.

But since our goal is to find such a tree, we instead look for a dissimilarity measure that can be calculated from data. The simplest such natural one is the *Hamming metric*, which simply counts the proportion of sites that are different in two sequences. If we are given a sequence of characters $\chi_1, \chi_2, \dots, \chi_n$ on X, set

$$\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\chi_i(x), \chi_i(y)},$$

where

$$\delta_{a,b} = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}.$$

For instance, given sequences such as

$$x : \texttt{AACTAGATCCTGTATCGA}$$
$$y : \texttt{ACCTAGGTCCTGTTTCGC}$$

we count 4 differences among 18 sites, so $\delta(x, y) = 4/18$. In biological literature, the Hamming metric is more usually called the *p-distance* (because its value is often denoted by p) or the *uncorrected distance.* While it is a metric by the mathematical definition of that term, it usually is not a tree metric as defined in Section 2.3.

Of course one can easily imagine variations on the Hamming metric, for instance that might assign different weightings to transitions or transversions. We won't pursue such possibilities, though, since in a subsequent chapter we'll construct much more sophisticated dissimilarity maps using a probabilistic model of molecular evolution. We introduce this particular dissimilarity map here only to provide a concrete example to keep in mind as we explore how a dissimilarity can lead to trees.

Dissimilarities can be defined in very different ways than by comparing sites in sequences. For instance, before sequencing was affordable, DNA hybridization was sometimes used to measure dissimilarity. DNA from two taxa was heated, mixed, and cooled, so that the two strands of the helices would separate, and then combine with similar strands from the other taxon, to form less stable molecules due to mismatches in base pairings. The melting temperature (on an appropriate scale) of the hybridized DNA could then be used as a measure of dissimilarity. One could also imagine using measures of dissimilarity that have nothing to do with DNA directly, but capture other features of the taxa.

We'd like to think of dissimilarity values as representing at least approximate distances between taxa along some metric tree, even if we don't think they will exactly match any tree metric. That is, we hope there is some metric tree (T, w), with positive edge lengths and tree metric d, so that the dissimilarity map δ has values close to the restriction of the tree metric d when applied to pairs of taxa. Of course there's nothing in our definition of a dissimilarity map to ensure this, or even to suggest that the dissimilarity values are even close to such distances measured along a metric tree. For now, we simply hope for the best, and forge ahead.

Suppose then that we have 4 taxa, and using some dissimilarity map we've computed a table of dissimilarities such as:

	S1	S2	S3	S4
S1		.45	.27	.53
S2			.40	.50
S3				.62

Table 5.1: Dissimilarities between taxa

How might we find a metric tree (T, w) for which these data are at least approximately the same as distances computed by the tree metric? More informally, if we know how far apart we want every pair of leaves to be, how can we come up with a tree topology and edge lengths to roughly ensure that?

If there were only 2 taxa, and so one dissimilarity value, this would be simple; draw a one edge unrooted tree and make the edge length match the dissimilarity value.

With 3 taxa, the situation is not much harder: draw the only possible binary unrooted tree relating the taxa. Then with the edge lengths called x, y, z, the 3 leaf-to-leaf distances on the tree are $x + y$, $y + z$, and $x + z$. If we set these equal to the three dissimilarity values, we have 3 linear equations in 3 unknowns, and can easily solve for x, y, and z.

However, if there are 4 taxa, this approach encounters two problems: First, we have 3 different unrooted trees to consider, and since we don't know which one is going to be best, we have to try them all. Second, there are 5 edge lengths to be determined on each of these trees. Since there are 6 dissimilarity values, if we attempt to solve equations to find the edge lengths, we obtain 6 linear equations in 5 unknowns. Such a system typically has no solution, indicating that there is no tree metric that will exactly match the dissimilarity values.

With larger sets of taxa, the problem only grows worse. There are more and more trees to consider, and we find we always have more equations than unknowns. Although computing a least-squares approximate solution would be a possible way of dealing with the overdetermined systems of equations, having to consider all trees will still be as difficult as it was with parsimony.

5.2 An Algorithmic Construction: UPGMA

Instead of trying to consider all possible trees, an alternative is to use a dissimilarity measure to *build* a tree. That is, by considering the dissimilarity values, we guess or infer certain features of the tree, trying to then combine these features until we reach a single tree. In this process we think of the dissimilarities as approximations of an unknown tree metric. We use the numbers in the dissimilarity table to assign plausible edge lengths on the metric tree we build.

When pressed to follow this outline, most students come up with some variant of an approach that is called the *average distance method*, or, more formally, the *unweighted pair-group method with arithmetic means (UPGMA)*. Rather than present the algorithm formally, we develop it through the example data in Table 5.1 above.

The first natural step is to assume that the two taxa which are shown as closest by the dissimilarity map are probably closest in the tree. With the data table above, we pick the two closest taxa, S1 and S3, and join them to a common ancestral vertex by edges. Drawing Figure 5.1, since S1 and S3 should be .27 apart, we decide to split this, making each edge $.27/2 = .135$ long.

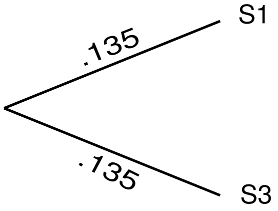

Figure 5.1: UPGMA, step 1

Since S1 and S3 have been joined, we now collapse them into a single combined group S1-S3. To say how far this group is from another taxon, we simply average the distances from S1 and S3 to that taxon. For example, the distance between S1-S3 and S2 is $(.45 + .40)/2 = .425$, and the distance between S1-S3 and S4 is $(.53 + .62)/2 = .575$. Our dissimilarity table thus collapses to Table 5.2.

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

Table 5.2: Distances between groups; UPGMA, step 1

Now we simply repeat the process, using the distances in the collapsed table. Since the closest taxa and/or groups in the new table are S1-S3 and S2, which are .425 apart, we draw Figure 5.2.

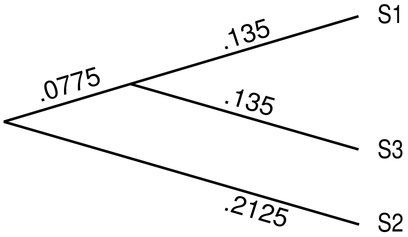

Figure 5.2: UPGMA; step 2

Note the edge to S2 must have length .425/2 = .2125. However the other new edge must have length (.425/2) − .135 = .0775, since we already have the edges of length .135 to account for some of the distance between S2 and the other taxa.

Again combining taxa, we form a group S1-S2-S3, and compute its distance from S4 by averaging *the original distances from S4 to each of S1, S2, and S3.* This gives us (.53+.5+.62)/3 = .55. Note that this is *not* the same as averaging the distance from S4 to S1-S3 and to S2. That average would downweight the contribution from S1 and S3, and thus not treat all our dissimilarities in the same way.

Since a new collapsed distance table would have only the one entry 0.55, there's no need to give it. We draw Figure 5.3, estimating that S4 is .55/2 = .275 from the root. The final edge has length .0625, since that places the other taxa .275 from the root as well.

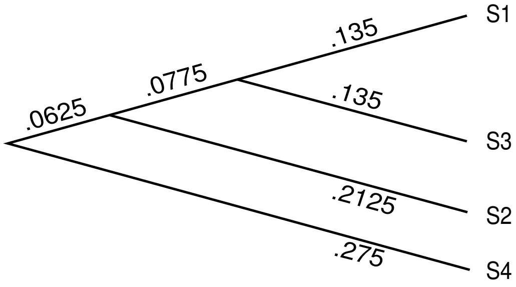

Figure 5.3: UPGMA; step 3

As suspected, the tree in Figure 5.3 constructed from the dissimilarity data does not exactly fit it. The tree metric from S3 to S4, for instance, gives .55, while according to the dissimilarity it should be .62. Nonetheless, the tree metric distances are at least fairly close to the dissimilarity distances.

If we had more taxa to relate, we'd have to do more steps to perform in the UPGMA algorithm, but there would be no new ideas involved. At each step, we join the two closest taxa or groups together, always placing them equidistant from a common ancestor. We then collapse the joined taxa or groups into a single group, using averaging to compute a distance from that new group to the taxa and groups still to be joined. The one point to be particularly careful about is that when the distances between two groups are computed, we average *all the original distances* from members of one group to another — if one group has n members and another has m members, we have to average nm distances. Each step of the algorithm reduces the size of the distance table by one group/taxon, so that after enough steps, all of the taxa are joined into a single tree.

A few comments on the algorithm are in order here.

First, UPGMA requires little work to give us a tree — at least in comparison to using a parsimony approach where we search through all possible trees, performing an algorithm on each. UPGMA requires searching only for the smallest entry in a succession of tables, and some rather simple algebraic computations. This can be viewed as a strength, as the algorithm is very fast[1], and we get a single answer. But it also can be viewed as a weakness, since it is not completely clear why we should consider the output tree 'good.' We have not explicitly formulated a criterion for judging trees and then found the tree that is optimal. Instead we've designed a process that uses many small steps, each of which may be reasonable on its own, to simply give us a tree.

Second, UPGMA implicitly assumes that we are seeking a rooted ultrametric tree. In this example, when we placed S1 and S3 at the ends of equal length branches, we assumed that the amount of mutation each underwent from their common ancestor was equal. UPGMA *always* places *all* the taxa at the same distance from the root. While this feature of UPGMA might be desirable if we believe a molecular clock underlies our data, in other situations it could be problematic.

5.3 Unequal Branch Lengths

It is not always desirable to impose a molecular clock hypothesis, as use of UPGMA requires. One way of dealing with this arose in a suggested algorithm of Fitch and Margoliash, which builds on the basic approach of UPGMA, but attempts to drop the molecular clock assumption through an additional step. Though the Fitch-Margoliash algorithm is probably never used in current data analysis, understanding it is useful for developing ideas.

Before giving the algorithm, we make a few mathematical observations. First, if we attempt to put 3 taxa on an unrooted tree, then there is only one topology that needs to be considered. Furthermore, for 3 taxa we can assign lengths to the edges to *exactly* fit any given dissimilarities (provided we possibly accept negative lengths). To see this consider the tree in Figure 5.4. If we have

[1] For n taxa, UPGMA takes $\mathcal{O}(n^3)$ steps, though clever programming can reduce this somewhat.

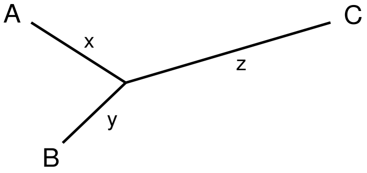

Figure 5.4: The unrooted 3-taxon tree

some dissimilarity data δ_{AB}, δ_{AC}, and δ_{BC}, then

$$\begin{array}{ccccccc} x & + & y & & & = & \delta_{AB}, \\ x & + & & & z & = & \delta_{AC}, \\ & & y & + & z & = & \delta_{BC}. \end{array} \tag{5.1}$$

Solving these equations leads to

$$\begin{aligned} x &= (\delta_{AB} + \delta_{AC} - \delta_{BC})/2, \\ y &= (\delta_{AB} + \delta_{BC} - \delta_{AC})/2, \\ z &= (\delta_{AC} + \delta_{BC} - \delta_{AB})/2. \end{aligned} \tag{5.2}$$

We'll refer to the formulas in equations (5.2) as the *three-point formulas* for fitting taxa to a tree. Unfortunately, with more than 3 taxa, exactly fitting dissimilarities to a tree is usually not possible (see Exercise 8). However, the Fitch-Margoliash algorithm uses the 3-taxon case to handle more taxa.

Now we explain the operation of the algorithm with an example. We'll use the dissimilarity data in Table 5.3.

	S1	S2	S3	S4	S5
S1		.31	1.01	.75	1.03
S2			1.00	.69	.90
S3				.61	.42
S4					.37

Table 5.3: Dissimilarities between taxa

We begin by choosing the closest pair of taxa to join, just as we did with UPGMA. Looking at our distance table, S1 and S2 are the first pair to join. In order to join them *without* placing them at an equal distance from a common ancestor, we temporarily reduce to the 3 taxa case by combining *all other* taxa into a group. For our data, we thus introduce the group S3-S4-S5. We find the distance from each of S1 and S2 to the group by averaging their distances to each group member. The distance from S1 to S3-S4-S5 is thus $d(\text{S1}, \text{S3-S4-S5}) = (1.01 + .75 + 1.03)/3 = .93$, while the distance from S2 to

	S1	S2	S3-S4-S5
S1		.31	.93
S2			.863

Table 5.4: Distances between groups; FM algorithm, step 1a

S3-S4-S5 is $d(\text{S2}, \text{S3-S4-S5}) = (1.00 + .69 + .90)/3 = .863$. This gives us Table 5.4.

With only three taxa in this table, we can exactly fit the data to the tree using the three-point formulas to get Figure 5.5. The key point here is that the three-point formulas, unlike UPGMA, can produce unequal distances of taxa from a common ancestor.

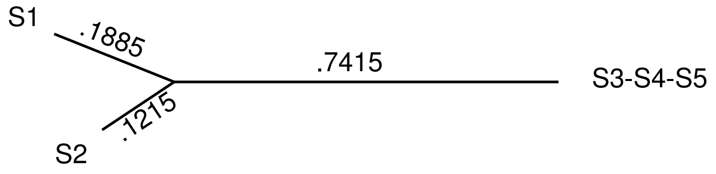

Figure 5.5: FM Algorithm; step 1

We now keep only the edges ending at S1 and S2 in Figure 5.5, and return to our original data. Remember, the group S3-S4-S5 was only needed temporarily so we could use the three-point formulas; we didn't intend to join those taxa together yet. Since we have joined S1 and S2, however, we combine them into a group for the rest of the algorithm, just as we would have done with UPGMA. This gives us Table 5.5.

	S1-S2	S3	S4	S5
S1-S2		1.005	.72	.965
S3			.61	.42
S4				.37

Table 5.5: Distances between groups; FM algorithm, step 1b

We again look for the closest pair (now S4 and S5), and join them in a similar manner. We combine everything but S4 and S5 into a single temporary group S1-S2-S3 and compute $d(\text{S4}, \text{S1-S2-S3}) = (.75 + .69 + .61)/3 = .683$ and $d(\text{S5}, \text{S1-S2-S3}) = (1.03 + .90 + .42)/3 = .783$. This gives us Table 5.6. Applying the three-point formulas to Table 5.6 produces Figure 5.6.

We keep the edges joining S4 and S5 in Figure 5.6, discarding the edge leading to the temporary group S1-S2-S3. Thus we now have two joined groups, S1-S2 and S4-S5. To compute a new table containing these two groups we've

	S1-S2-S3	S4	S5
S1-S2-S3		.683	.783
S4			.37

Table 5.6: Distances between groups; FM algorithm, step 2a

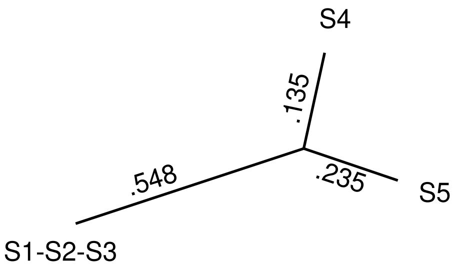

Figure 5.6: FM Algorithm; step 2

found, we average $d(\text{S1-S2}, \text{S4-S5}) = (.75 + 1.03 + .69 + .90)/4 = .8425$ and $d(\text{S3}, \text{S4-S5}) = (.61 + .42)/2 = .515$. We've already computed $d(\text{S1-S2}, \text{S3})$ so we produce Table 5.7. At this point we can fit a tree exactly to the table by a final application of the three-point formulas, yielding Figure 5.7.

	S1-S2	S3	S4-S5
S1-S2		1.005	.8425
S3			.515

Table 5.7: Distances between groups; FM algorithm, step 2b

Now we replace the groups in this last diagram with the branching patterns we've already found for them. This gives Figure 5.8.

Our final step is to fill in the remaining lengths a and b, using the lengths in Figure 5.7. Since S1 and S2 are on average $(.1885 + .1215)/2 = .155$ from the vertex joining them and S4 and S5 are on average $(.135 + .235)/2 = .185$ from the vertex joining them, we compute $a = .66625 - .155 = .51125$ and $b = .17625 - .185 = -.00875$ to assign lengths to the remaining sides.

Notice that one edge has turned out to have negative length. Since that can't really be meaningful, many practitioners would choose to simply reassign the length as 0. If this happens, however, we should at least check that the negative length was close to 0. If not, we should doubt that our data really are described well by the tree we produced.

A dissappointing feature of the Fitch-Margoliash algorithm is that from any dissimilarity data it always produces the same unrooted topological tree as UPGMA. The reason for this is that each time we decide which taxa or groups

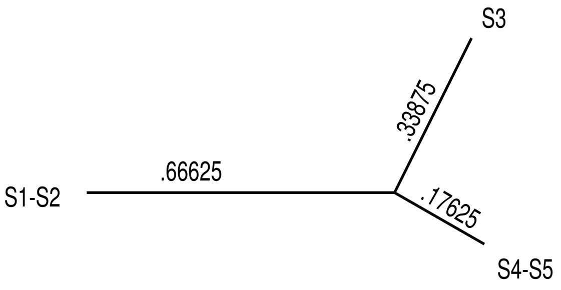

Figure 5.7: FM Algorithm; step 3

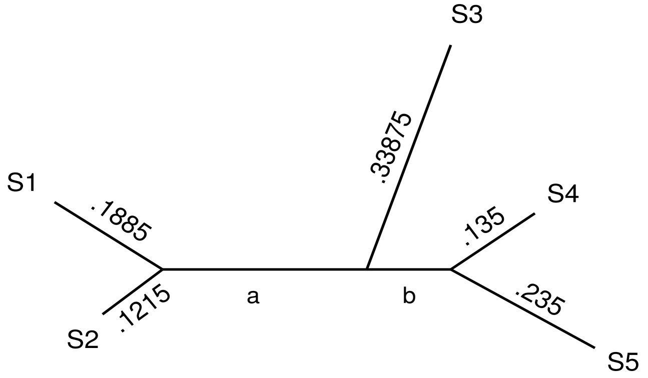

Figure 5.8: FM Algorithm; completion

to join, both methods consider exactly the same collapsed data table and both choose the pair corresponding to the smallest entry in the table. It is therefore only the metric features of the resulting trees that will differ.

To be fair, Fitch and Margoliash actually proposed their algorithm not as an end in itself, but a heuristic method for producing a tree likely to have a certain optimality property (see Exercise 10). We are viewing it here as a step toward the Neighbor Joining algorithm which will be introduced shortly. Familiarity with UPGMA and the Fitch-Margoliash algorithm will aid us in understanding that more elaborate method.

5.4 The Four-point Condition

If we are interested in potentially non-ultrametric trees, as is often the case because of the implausibility of a molecular clock hypothesis, there is a fundamental flaw in using UPGMA to construct trees. Moreover, this flaw is not corrected by the use of the three-point formulas alone.

To better understand the problem, consider the metric quartet tree in Figure

5.9, which we imagine representing the true relationships between the taxa. Here x and y represent specific lengths, with x much smaller than y. Then perfect dissimilarity data (*i.e.*, dissimilarities determined by the tree metric) would give us the distances in Table 5.8.

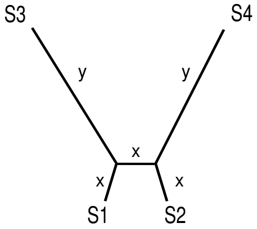

Figure 5.9: The true relationships of taxa S1, S2, S3, S4

	S1	S2	S3	S4
S1		$3x$	$x+y$	$2x+y$
S2			$2x+y$	$x+y$
S3				$x+2y$

Table 5.8: Distances between taxa in Figure 5.9

But if y is much bigger than x, (in fact, $y > 2x$ is good enough) then the closest taxa by distance are S1 and S2. Now the first step of UPGMA will be to look for the smallest dissimilarity in the table, and join those two taxa. This means we'll choose $S1$ and $S2$ as the most closely related, and relate them by a tree that already has a topological mistake. No matter what we do to compute edge lengths, or how the subsequent steps proceed, we will not recover the original tree topology.

The essential problem here is a conflict between closeness of taxa as measured by the tree metric and closeness in the graph theoretic sense as measured by the number of edges on the path connecting them. These are very different notions. It is only reasonable to expect the first to be obvious from data, while the second is the one more relevant to determining the topology of the tree. This is the issue we need to explore theoretically, so that in the next section we can give a practical algorithm to address the problem.

Definition. Two leaves of a tree that are graph-theoretic distance 2 apart are said to form a *cherry*, or are said to be *neighbors.*

Focusing on quartet trees for simplicity, the problem with UPGMA is that it can incorrectly identify a cherry. Once this first cherry is chosen, the full topological quartet tree is determined. So to improve on UPGMA, the key insight is to that we must find a better way to pick a cherry.

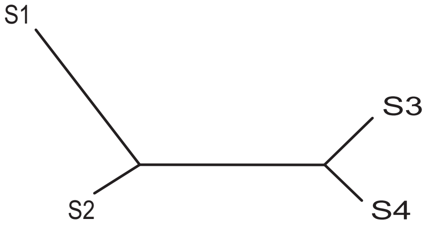

Figure 5.10: A quartet tree with cherries S1, S2 and S3, S4

Consider the quartet tree in Figure 5.10, viewed as a metric tree with positive edge lengths. Letting $d_{ij} = d(\mathrm{S}i, \mathrm{S}j)$ denote the tree metric between leaves, we see that

$$d_{12} + d_{34} < d_{13} + d_{24},$$

since the quantity on the left includes only the lengths of the four edges leading from the leaves of the tree, while the quantity on the right includes all of those and, in addition, twice the central edge length. Notice also that

$$d_{13} + d_{24} = d_{14} + d_{23}$$

by similar reasoning.

On the other hand, if we consider a different quartet tree, ((S1, S3), (S2, S4)), we find

$$d_{13} + d_{24} < d_{12} + d_{34} = d_{14} + d_{23}.$$

Notice the inequalities and equalities in this statement are all incompatible with those for the first tree. For instance $d_{13}+d_{24} < d_{12}+d_{34}$ for ((S1, S3), (S2, S4)), while $d_{13} + d_{24} > d_{12} + d_{34}$ for ((S1, S2), (S3, S4)) . These simple observations lead to a way of using metric information to identify cherries. We summarize this as a theorem.

Theorem 13. A metric quartet tree relating taxa a, b, c, d which has positive edge lengths has cherries a, b and c, d if, and only if, any (and hence all) of the three inequalities/equalities in the tree metric that follow hold:

$$d(a, b) + d(c, d) < d(a, c) + d(b, d) = d(a, d) + d(b, c).$$

It should already be clear this observation will be useful, since identifying quartet trees enables us to identify larger tree topologies: For a positive-edge-length binary phylogenetic X-tree T, we get induced positive-edge-length trees for every 4-element subset of X, and hence we can identify the appropriate quartet tree for each such subset. Then, by results in Chapter 4, these quartet trees determine the topology of the tree T.

However, we can do even better in this vein, giving a condition on a dissimilarity map that enables us to relate it to a tree metric.

Definition. A dissimilarity map δ on X satisfies the *four-point condition* if for every choice of four taxa $x, y, z, w \in X$ (including non-distinct ones),

$$\delta(x, y) + \delta(z, w) \leq \max\{\delta(x, z) + \delta(y, w), \delta(x, w) + \delta(y, z)\}. \tag{5.3}$$

Note the non-strict inequality in this definition allows us to formulate the following for trees that are not necessarily binary.

Theorem 14. Given any metric tree with positive edge lengths, its tree metric is a dissimilarity map that satisfies the four-point condition.

Proof. We leave the proof as Exercise 17. □

The converse to this theorem is more remarkable.

Theorem 15. Suppose $|X| \geq 3$, and δ is a dissimilarity on X with $\delta(x, y) \neq 0$ whenever $x \neq y$. Then if δ satisfies the four-point condition, there is a unique metric X-tree with positive edge lengths whose tree metric agrees with δ.

Proof. The case of $|X| = 3$ follows by taking $z = w$ in the 4-point condition as stated above, and applying the three-point formulas with a little care to show edge lengths are positive (see Exercise 18). The case $|X| = 4$ is Exercise 19.

For larger $|X| = N$, we proceed by induction. However, we first need the notion of a *generalized cherry* on an X-tree. For any taxon labeling an internal vertex of the tree, temporarily attach a new edge at that vertex and move the label to its other end, so all labels are now on unique leaves of a phylogenetic X-tree. Then we refer to any cherry on this modified tree as a generalized cherry on the original tree.

Now suppose $X = \{\mathrm{S}1, \mathrm{S}2, \dots, \mathrm{S}N\}$. Using the inductive hypothesis, let T' be the unique metric X'-tree with positive edge lengths relating $X' = \{\mathrm{S}1, \mathrm{S}2, \dots, \mathrm{S}(N-1)\}$ whose tree metric restricts to δ on X'. Now choose some generalized cherry on T', and, assume the taxa in the cherry are S1, S2 by renaming them if necessary.

For all $j \neq 1, 2, N$ consider the 4-leaf metric trees on the taxa $\{\mathrm{S}1, \mathrm{S}2, \mathrm{S}j, \mathrm{S}N\}$ whose tree metrics agree with δ. (These trees exists by the 4-taxon case previously shown.) We claim that at least one of the pairs $\mathrm{S}1, \mathrm{S}N$, or $\mathrm{S}2, \mathrm{S}N$, or $\mathrm{S}1, \mathrm{S}2$ forms a generalized cherry in all of these 4-leaf trees. To see why this is so, suppose $\mathrm{S}1, \mathrm{S}N$ form a generalized cherry in the tree for $\{\mathrm{S}1, \mathrm{S}2, \mathrm{S}k, \mathrm{S}N\}$ for some k. On this 4-leaf tree if the vertex where paths to S1, S2 and SN meet is at a metric distance a from S1, and the vertex where paths from S1, S2, Sk meet is at a distance b from S1, then $a \leq b$. But since S1 and S2 form a generalized cherry in T', the vertex where paths from S1, S2, Sk meet on T' is the same as that where paths from S1, S2, and Sj meet for all $j < N$. Now a can be computed from the distances between $\mathrm{S}1, \mathrm{S}2$ and SN, and b from the distances between $\mathrm{S}1, \mathrm{S}2, \mathrm{S}j$ for any $2 < j < N$, always giving the same value regardless of j, by the inductive hypothesis. But then the inequality $a \leq b$ shows that S1 and SN must form a generalized cherry in the trees for all $\{\mathrm{S}1, \mathrm{S}2, \mathrm{S}j, \mathrm{S}N\}$. Similarly if $\mathrm{S}2, \mathrm{S}N$ form a generalized cherry for one of the 4-leaf trees, they form a generalized cherry for all. If there are no 4-leaf trees where either $\mathrm{S}1, \mathrm{S}N$ or $\mathrm{S}2, \mathrm{S}N$ form a generalized cherry, then $\mathrm{S}1, \mathrm{S}2$ must form a generalized cherry in all.

Using the claim of the last paragraph, by interchanging the names of $\mathrm{S}1, \mathrm{S}2, \mathrm{S}N$ if necessary, we may assume $\mathrm{S}1, \mathrm{S}N$ always form a generalized cherry for all the 4-leaf trees relating $\mathrm{S}1, \mathrm{S}2, \mathrm{S}j, \mathrm{S}N$. Again let T' be the tree for $\{\mathrm{S}1, \mathrm{S}2, \dots, \mathrm{S}(N-1)\}$. (So now S1 and S2 may not form a generalized cherry in T' due to interchanging taxa names.) We leave to the reader the final step of giving the unique way to 'attach' $\mathrm{S}N$ to T' consistent with all dissimilarity values (Exercise 20). □

5.5 The Neighbor Joining Algorithm

In practice, UPGMA with its ultrametric assumption is only used on biological data in very special circumstances. Much preferred is a more elaborate algorithm, called *Neighbor Joining*, which is built on the four-point condition.

However, it is important that Neighbor Joining not require that the four-point condition be *exactly* met for the dissimilarity data to which it is applied, since dissimilarities computed from data should be expected at best to be only roughly consistent with a metric tree. We therefore want to perform various averaging processes as we go along in order to smooth out some of the errors in fit.

To motivate the algorithm, we imagine a binary positive-edge-length tree in which taxa S1 and S2 form a cherry joined at vertex v, with v somehow joined to the remaining taxa S3, S4,..., SN, as in Figure 5.11.

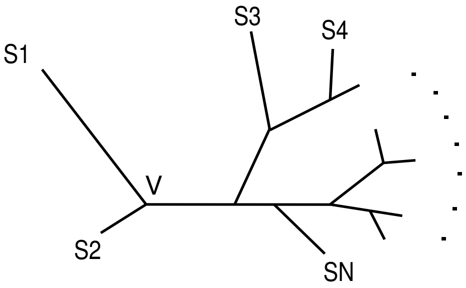

Figure 5.11: Tree with S1 and S2 forming a cherry

If our dissimilarity data agreed exactly with a metric for this tree then for every $i, j = 3, 4, \dots, N$, we'd find from the four-point condition that

$$d_{12} + d_{ij} < d_{1i} + d_{2j}. \tag{5.4}$$

For fixed i, there are $N-3$ possible choices of j with $3 \le j \le N$ and $j \ne i$. If we sum the inequalities (5.4) for these j we get

$$(N-3)d_{12} + \sum_{\substack{j=3 \\ j\neq i}}^{N} d_{ij} < (N-3)d_{1i} + \sum_{\substack{j=3 \\ j\neq i}}^{N} d_{2j}. \tag{5.5}$$

To simplify this, define the total dissimilarity between taxon Si and all other taxa as

$$R_i = \sum_{j=1}^{N} d_{ij}.$$

Then adding $d_{i1} + d_{i2} + d_{12}$ to each side of inequality (5.5) allows us to write it in the simpler form

$$(N-2)d_{12} + R_i < (N-2)d_{1i} + R_2.$$

Subtracting $R_1 + R_2 + R_i$ from each side of this then gives a more symmetric statement,

$$(N-2)d_{12} - R_1 - R_2 < (N-2)d_{1i} - R_1 - R_i.$$

If we apply the same argument to Sn and Sm, rather than S1 and S2, we are led to define

$$M_{nm} = (N-2)d_{nm} - R_n - R_m. \tag{5.6}$$

Then if Sn and Sm form a cherry, we'll have that

$$M_{nm} < M_{nk}$$

for all $k \neq m$.

This gives us the criterion used for Neighbor Joining: From the dissimilarity data, compute a new table of values for M_{ij} using equation (5.6). Then choose to join the pair Si, Sj of taxa with the smallest value of M_{ij}.

The argument above shows that if Si and Sj form a cherry in a metric tree producing the dissimilarity data, then the value M_{ij} will be the smallest of the values in the ith row and jth column of the table for M. However, this is not enough to justify the Neighbor Joining criterion. An additional argument is still needed to show the smallest entry in the *entire table* for M truly identifies a cherry. Though this claim is plausible, we outline a proof of it in Exercise 24.

We now describe the full Neighbor Joining algorithm.

Algorithm.

1. Given dissimilarity data for N taxa, compute a new table of values of M using equation (5.6). Choose the smallest value in the table for M to determine which taxa to join. (This value may be, and usually is, negative, so 'smallest' means the negative number with the greatest absolute value.)

2. If Si and Sj are to be joined at a new vertex v, temporarily collapse all other taxa into a single group G, and determine the lengths of the edges from

Si and Sj to v by using using the three-point formulas for Si, Sj, and G as in the algorithm of Fitch and Margoliash.

3. Determine distances/dissimilarities from each of the taxa Sk in G to v by applying the three-point formulas to the distance data for the three taxa Si, Sj and Sk. Now include v in the table of dissimilarity data, and drop Si and Sj.

4. The distance table now includes $N-1$ taxa. If there are only 3 taxa, use the three-point formulas to finish. Otherwise go back to step 1.

Exercise 24 completes the proof of the following theorem, by showing that for dissimilarity data consistent with a tree metric, the Neighbor Joining algorithm really does join taxa to correctly recover the metric tree.

Theorem 16. Suppose a dissimilarity map on X is the restriction of a tree metric for a binary metric phylogenetic X-tree T with all positive edge lengths. Then the Neighbor Joining algorithm will reconstruct T and its edge lengths.

As you can see already, Neighbor Joining is not pleasant to do by hand. Even though each step is relatively straightforward, it's easy to get lost in the process with so much arithmetic to do. In the exercises you'll find an example partially worked that you should complete to be sure you understand the steps. After that, we suggest you use a computer program to avoid mistakes (or, even better, write your own program).

The accuracy of various tree construction methods – the ones outlined so far in these notes and many others – has been tested primarily through simulating DNA mutation according to certain specified models of mutation along phylogenetic trees and then applying the methods to see how often they recover the tree that was the basis for the simulation. These tests have lead researchers to be considerably more confident of the results given by Neighbor Joining than by UPGMA. While UPGMA may be reliable, or even preferred, under some special circumstances, Neighbor Joining works well on a broader range of data. Since it appears that a molecular clock hypothesis is often violated by real data, Neighbor Joining is by far the most commonly used distance method for tree construction in phylogenetics.

5.6 Additional Comments

An important point to remember is that so far we have simply hoped the dissimilarity measure we began with was reasonably close to a tree metric, and so could be lead to an appropriate metric tree. But we have given no argument that the Hamming distance, or any other dissimilarity measure, should be approximately 'tree-like.' In fact, under plausible models for the evolution of DNA sequences the Hamming distance need not be close to tree-like, unless the total amount of mutation between taxa is small. We'll overcome this problem in the next chapter, by developing such models and then using them to introduce corrections.

Although Neighbor Joining and UPGMA lack an explicit criterion for how they determine the 'best' tree to fit data, there are distance methods which are based on such criteria. For instance, the Minimum Evolution approach considers each possible topological tree in turn, use least-squares minimization to determine edge lengths that best fit the dissimilarity data, and then chooses from these metric trees the one with the minimum total of edge lengths. To follow this scheme, however, one is forced to consider all possible topological trees, and this prevents the method from being comparable in speed to the algorithmic approaches discussed here. A variant of this, Balanced Minimum Evolution, introduces weights in the least-squares minimization in a specific way. Neighbor Joining has been shown to be a greedy algorithm to approximate the BME tree, which avoids searching over all trees.

A valid criticism of all distance methods is that they do not use the full information in the data, since they are based on only pairwise comparisons of the taxa. By not comparing all taxa at once, some potential information is lost. It's not too hard to see that it is impossible to reconstruct sequence data, or even a rough description of it, from the collection of pairwise Hamming distances between them. Thus we have lost something by boiling the sequences down so crudely into a few numbers. Indeed, this is probably the main reason distance methods should be viewed as a less than optimal approach, to be used primarily for quick initial explorations of data, or on very large data sets when other methods are too slow to be feasible. Often software for more elaborate inference methods will begin by constructing an initial 'pretty-good' tree by a distance method, before searching for a better tree that is similar to it. However, for a very large number of taxa it may not be possible to effectively search among similar trees in an acceptable amount of time.

Finally, theoretical work on distance methods is continuing. There is a variant of Neighbor Joining, called Bio-NJ, that gains improved performance by taking into account that the dissimilarity between taxa a, b will be statistically correlated with that between c, d if the path from a to b shares some edges with that from c to d. Other recent theoretical work has indicated that if this correlation were more fully exploited, then distance methods could be (in a precise technical sense) similar in effectiveness to more elaborate inference through a maximum likelihood framework.

5.7 Exercises

1. For the tree in Figure 5.3 constructed by UPGMA, compute a table of distances between taxa along the tree. How does this compare to the original dissimilarities of Table 5.1?

2. Suppose four sequences S1, S2, S3, and S4 of DNA are separated by dissimilarities as in Table 5.9. Construct a rooted tree showing the relationships

between S1, S2, S3, and S4 by UPGMA.

	S1	S2	S3	S4
S1		1.2	.9	1.7
S2			1.1	1.9
S3				1.6

Table 5.9: Dissimilarity data for Problems 2 and 5

3. Perform UPGMA on the data in Table 5.3 that was used in the text in the example of the FM algorithm. Does UPGMA produce the same tree as the FM algorithm topologically? metrically?

4. The fact that dissimilarity data relating three taxa can be exactly fit by appropriate edge lengths on the single unrooted topological 3-taxon tree is used in the Neighbor Joining algorithm.

   a. Derive the three-point formulas of Equation 5.1.

   b. If the dissimilarities are $\delta_{AB} = .634$, $\delta_{AC} = 1.327$, and $\delta_{BC} = .851$, what are the lengths x, y, and z?

5. Use the FM algorithm to construct an unrooted tree for the data in Table 5.9 that were also used in Problem 2. How different is the result?

6. A desirable feature of a dissimilarity map on sequences is that it be *additive*, in the sense that if S0 is an ancestor of S1, which is in turn an ancestor of S2, then
$$d(\mathrm{S0}, \mathrm{S2}) = d(\mathrm{S0}, \mathrm{S1}) + d(\mathrm{S1}, \mathrm{S2}).$$

   a. Explain why an additive dissimilarity map is desirable if we are trying to use dissimilarities to construct metric trees.

   b. Give an example of sequences to illustrate that the Hamming dissimilarity might not be additive.

   c. If mutations are rare, why might the Hamming dissimilarity be approximately additive?

7. While any dissimilarity values between 3 taxa can be fit to a metric tree (possibly with negative edge lengths), that's not the case if we want to fit an ultrametric tree.

   a. If the three dissimilarities are 0.3, 0.4, and 0.5, find an unrooted tree that fits them, and explain why no choice of a root location can make this tree ultrametric.

   b. If the three dissimilarities are 0.2, 0.3, 0.3 find an unrooted tree that fits them, and locate a root to make it ultrametric.

c. If the three dissimilarities are 0.3, 0.3, and 0.4, find an unrooted tree that fits them, and explain why no choice of a root location can make this tree ultrametric.

8. While distance data for 3 taxa can be exactly fit to an unrooted tree, if there are 4 (or more) taxa, this is usually not possible.

   a. For the tree $((a,b),(c,d))$, denoting distances between taxa with notation like d_{ab}, write down equations for each of the 6 such distances in terms of the 5 edge lengths. Explain why if you use dissimilarity values in place of the distances these equations are not likely to have an exact solution.

   b. Give a concrete example of 6 dissimilarity values so that the equations in part (a) cannot be solved exactly. Give another example of values where the equations can be solved.

9. Suppose you have a dissimilarity values for n taxa, and wish to see if it could exactly fit a tree metric on a particular binary tree T. How many equations and how many unknowns would be in the resulting system of equations you would need to solve? Show for $n \geq 4$ there are more equations than unknowns in this system, and thus it is unlikely to have a solution.

10. A number of different measures of goodness of fit between dissimilarity data and metric trees have been proposed. Let δ_{ij} denote the dissimilarity between taxa i and j, and e_{ij} denote the tree metric distance from i to j. A few of the these measures are:

$$\begin{aligned} s_{FM} &= \left(\sum_{i,j} \left(\frac{\delta_{ij} - e_{ij}}{\delta_{ij}} \right)^2 \right)^{\frac{1}{2}}, \\ s_F &= \sum_{i,j} \left| \delta_{ij} - e_{ij} \right|, \\ s_{TNT} &= \left(\sum_{i,j} (\delta_{ij} - e_{ij})^2 \right)^{\frac{1}{2}}. \end{aligned}$$

   In all these measures, the sums include terms for each distinct pair of taxa, i and j.

   a. Compute these measures for the tree constructed in the text using the FM algorithm, as well as the tree constructed from the same data using UPGMA in Problem 3. According to each of these measures, which of the two trees is a better fit to the data?

   b. Explain why these formulas are reasonable ones to use to measure goodness of fit. Explain how the differences between the formulas make them more or less sensitive to different types of errors.

Note: Fitch and Margoliash proposed choosing the optimal metric tree to fit data as the one that minimized s_{FM}. The FM algorithm was introduced in an attempt to get an approximately optimal tree.

11. Suppose the unrooted metric tree in Figure 5.12 correctly describes the evolution of taxa A, B, C, and D.

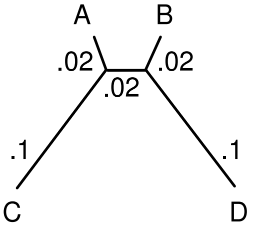

Figure 5.12: Tree for problem 11

a. Explain why, regardless of the location of the root, a molecular clock could not have operated.

b. Give a dissimilarity table by calculating tree metric distances between each pair of the four taxa. Perform UPGMA on that data.

c. UPGMA did not reconstruct the correct tree. Where did it go wrong? What was it about this metric tree that led it astray?

d. Explain why the FM algorithm will also not reconstruct the correct tree.

12. Show that every unrooted binary phylogenetic tree with at least three taxa has at least two cherries. (A unrooted binary tree with n leaves that has only two cherries is sometimes called a *caterpillar*. Despite the unfortunate mixing of metaphors, why is this terminology reasonable?)

13. For the quartet tree $((a,d),(b,c))$ that was not explicitly treated in Section 5.4, write the inequalities and equalities that hold expressing the four-point condition.

14. Show that if a dissimilarity map on X satisfies the four-point condition, then it satisfies the triangle inequality on X. (Rather than deducing this from Theorem 15, show it directly by taking several of the taxa to be the same in the four-point condition.)

15. If a dissimilarity map arises from a tree metric on X, then it is a metric on X. Show that the converse is false by giving an example of a metric on a set X that is not a tree metric.

16. Show that the four-point condition is equivalent to the following statement: For every choice of $x, y, z, w \in X$, of the three quantities

$$\delta(x,y)+\delta(z,w),\ \delta(x,z)+\delta(y,w),\ \delta(x,w)+\delta(y,z)$$

the two (or three) largest are equal.

17. Prove Theorem 14, by first observing that it's enough to prove it for 4-leaf trees. For 4-leaf trees, be sure you consider both binary and non-binary trees, and cases where x, y, z, w are all distinct and when they are not.

18. Prove Theorem 15 for the case $|X| = 3$. (Be sure you show all edge lengths are non-negative by using the four-point condition.)

19. Prove Theorem 15 for the case $|X| = 4$.

20. Give the final step of the proof of Theorem 15.

21. Before working through an example of Neighbor Joining, it's helpful to derive formulas for Steps 2 and 3 of the algorithm. Suppose we've chosen to join Si and Sj in Step 1.

    a. Show that for Step 2, the distances of Si and Sj to the internal vertex v can be computed by

$$d(\mathrm{S}i, v) = \frac{\delta(\mathrm{S}i, \mathrm{S}j)}{2} + \frac{R_i - R_j}{2(N-2)},$$
$$d(\mathrm{S}j, v) = \frac{\delta(\mathrm{S}i, \mathrm{S}j)}{2} + \frac{R_j - R_i}{2(N-2)}.$$

    Then show the second of these formulas can be replaced by

$$d(\mathrm{S}j, v) = \delta(\mathrm{S}i, \mathrm{S}j) - d(\mathrm{S}i, v).$$

    b. Show that for Step 3, the distances of Sk to v, for $k \neq i, j$ can be computed by

$$d(\mathrm{S}k, v) = \frac{\delta(\mathrm{S}i, \mathrm{S}k) + \delta(\mathrm{S}j, \mathrm{S}k) - \delta(\mathrm{S}i, \mathrm{S}j)}{2}.$$

22. Consider the distance data of Table 5.10. Use the Neighbor Joining algorithm to construct a tree as follows:

	S1	S2	S3	S4
S1		.83	.28	.41
S2			.72	.97
S3				.48

Table 5.10: Taxon distances for Problem 22

    a. Compute R_1, R_2, R_3 and R_4 and then a table of values for M for the taxa S1, S2, S3, and S4. To get you started

$$R_1 = .83 + .28 + .41 = 1.52 \text{ and } R_2 = .83 + .72 + .97 = 2.52$$

so

$$M(\mathrm{S1}, \mathrm{S2}) = (4 - 2).83 - 1.52 - 2.52 = -2.38.$$

b. If you did part (a) correctly, you should have a tie for the smallest value of M. One of these smallest values is $M(\mathrm{S1}, \mathrm{S4}) = -2.56$, so let's join S1 and S4 first.

For the new vertex v where S1 and S4 join, compute $d(\mathrm{S1}, v)$ and $d(\mathrm{S4}, v)$ by the formulas in part (a) of the previous problem.

c. Compute $d(\mathrm{S2}, v)$ and $d(\mathrm{S3}, v)$ by the formulas in part (b) of the previous problem.

Put your answers into the new distance Table 5.11.

	v	S2	S3
v		—	—
S2			.72

Table 5.11: Group distances for Problem 22

d. Since there are only 3 taxa left, use the three-point formulas to fit v, S2, and S3 to a tree.

e. Draw your final tree by attaching S1 and S4 to v with the distances given in part (b).

23. Consider the distance data in Table 5.12, which is exactly fit by the tree of Figure 5.9, with $x = .1$ and $y = .3$.

	S1	S2	S3	S4
S1		.3	.4	.5
S2			.5	.4
S3				.7

Table 5.12: Taxon distances for Problem 23

a. Use UPGMA to reconstruct a tree from these data. Is it correct?

b. Use Neighbor Joining to reconstruct a tree from these data. Is it correct?

24. Complete the proof of Theorem 16 by showing that the criterion used by Neighbor Joining to pick cherries from dissimilarity data arising from a positive edge length binary metric tree will pick only true cherries. Do this by following the outline below of the proof of Studier and Keppler.

a. Show $M_{mn} - M_{ij} = \sum_{k \neq i,j,m,n} ((d_{ik} + d_{jk} - d_{ij}) - (d_{mk} + d_{nk} - d_{mn}))$.

Now suppose M_{ij} is minimal but that i and j do not form a cherry.

b. Explain why neither i nor j are in any cherry.

Pick some cherry Sm, Sn, and consider the 4-leaf subtree Q of T joining Si, Sj, Sm, Sn inside T. Denote the internal vertices of it with degree 3 by u (joined to i, j) and v (joined to m, n). For each additional taxon Sk, show the following:

c. If the path in T from Sk to Q joins Q at a vertex x along the path from i to j, then $(d_{ik} + d_{jk} - d_{ij}) - (d_{mk} + d_{nk} - d_{mn}) = -2d_{ux} - 2d_{uv}$.

d. If the path in T from Sk to Q joins Q at a vertex x along the path from u to v, then $(d_{ik} + d_{jk} - d_{ij}) - (d_{mk} + d_{nk} - d_{mn}) = -4d_{vx} + 2d_{uv}$.

e. Conclude from the fact that left hand side of the equality in (a) is nonnegative that at least as many of the Sk are described by (d) as by (c); and thus there are strictly more leaves that are joined to the path from i to j at u than at any other vertex.

f. Explain why since i and j are not in any cherries there must be a cherry r, s which is joined to the path from i to j at some vertex other than u.

g. Applying the argument of parts (c), (d), (e) to r, s instead of m, n, arrive at a contradiction.