CHAPTER 5

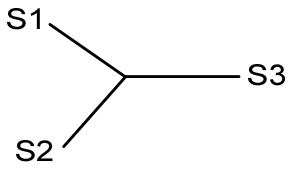
Constructing Phylogenetic Trees

5.1. Phylogenetic Trees

Warning: Trees are not drawn to scale.

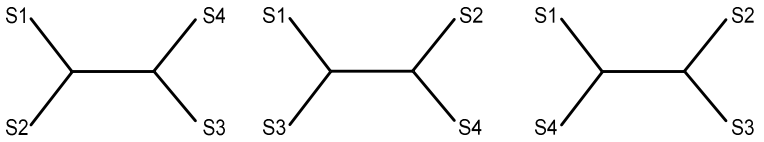
5.1.1. a. $\{T_2, T_3\}$
b. $\{T_2, T_3, T_5\}$
c. $\{T_1, T_6\}$, $\{T_2, T_3, T_5\}$
d. $\{T_1, T_2, T_3, T_4, T_5, T_6\}$
e. T_4, T_6

5.1.2. a.

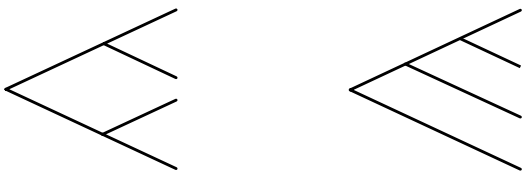


b. In the tree in part (a), the root can be placed along the edge joining the internal node to S1, S2, or S3.

5.1.3. a.



b. In each of the three trees in part (a), the root can be located on any of the five edges.
Equivalently, for the tree below on the left there are three distinct labelings (from top to bottom) of the leaves: {S1, S2, S3, S4}, {S1, S3, S2, S4}, {S1, S4, S2, S3}, and for the tree on the right there are twelve distinct labelings: {S1, S2, S3, S4}, {S1, S2, S4, S3}, {S1, S3, S2, S4}, {S1, S3, S4, S2}, {S1, S4, S2, S3}, {S1, S4, S3, S2}, {S2, S3, S1, S4}, {S2, S3, S4, S1}, {S2, S4, S1, S3}, {S2, S4, S3, S1}, {S3, S4, S1, S2}, {S3, S4, S2, S1}.



5.1.4.

n	3	4	5	6	7	8	9	10
$\frac{(2n-5)!}{2^{(n-3)}(n-3)!}$	1	3	15	105	945	10395	135135	2027025

5.1.5.

n	2	3	4	5	6	7	8	9	10
$\frac{(2n-3)!}{2^{(n-2)}(n-2)!}$	1	3	15	105	945	10395	135135	2027025	34459425

5.1.6. a. When we add a new edge to an existing tree, the edge count increases by two: one for the new edge, and one more since an existing edge is split into two edges where the new one is attached.

b. By part (a), each time we add an edge the edge count increases by two. Thus, we see the pattern:

n	2	3	4	5	$\cdots$	n
e	1	3	5	7	$\cdots$	$2n-3$

Alternatively, $e = 1+2(n-2)$ counts 1 edge for the first two terminal vertices, plus 2 edges for each of the other $(n-2)$ terminal vertices successively attached to the tree.

c. An unrooted tree with n terminal vertices has $2n-3$ edges. To create an unrooted tree with $n+1$ terminal vertices from such a tree, a new edge with the new terminal vertex can be attached to any of the $2n-3$ existing edges. Thus, if there are m unrooted trees with n terminal vertices, there are $m(2n-3)$ unrooted trees with $n+1$ terminal vertices.

d. Iterating the result of part (c) gives the formula:

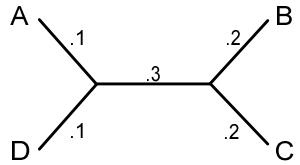
n	no. unrooted trees
2	1
3	$1(2\cdot 2-3)=1$
4	$1(2\cdot 3-3)=1\cdot 3$
5	$1\cdot 3\cdot(2\cdot 4-3)=1\cdot 3\cdot 5$
$\vdots$	$\vdots$
n	$(1)(3)(5)\cdots(2n-5)$

e. The denominator contains as factors all the even numbers between 2 and $(2n-6)$, canceling out the even numbers in $(2n-5)!$

f. Imagine a rooted tree with n terminal vertices. Attaching a new edge at the root location creates an unrooted tree with $n+1$ terminal vertices.

5.1.7. The most accurate estimate, produced by writing a brief computer program to find the product, is 4.89×10^{296}.

5.1.8. a. Approaches may vary, but the only tree fitting the data is:



b. There is no way to determine the root, without making some additional assumptions. If you assume a molecular clock, then the root belongs on the central branch, .2 away from the node joining A and D.

5.2. Tree Construction: Distance Methods – Basics

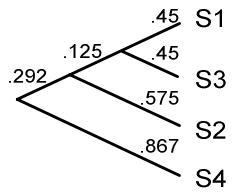
Warning: Trees are not drawn to scale.

5.2.1.

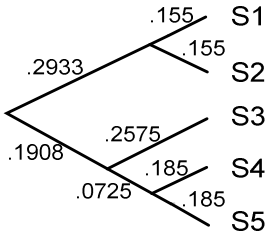
	S1	S2	S3	S4
S1		.425	.27	.55
S2			.425	.55
S3				.55

While the distance between the first two taxa to be joined, $d(\text{S1}, \text{S3})$, agrees exactly with the original distance table, the other distances are only close to the original distances. The duplication of some table entries reflects the molecular clock hypothesis, since certain subsets of taxa will be equidistant from a common ancestor.

5.2.2.



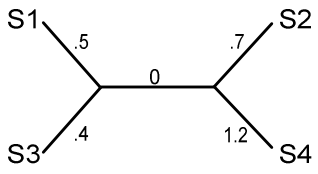
5.2.3.



Topologically, the rooted UPGMA tree is the same as the unrooted FM tree. However, the metric distances are not the same; you can see the molecular clock hypothesis at work in the UPGMA tree.

5.2.4. a. There are several algebraic approaches: Either *ad hoc* algebra or methodical elimination of variables can be used, or matrix algebra. The nicest solution (since it makes the formulas memorable) is a geometric one: $d_{AB}+d_{AC}$ includes the edge x twice, and the edges y and z once, so subtracting d_{BC} gives $2x$, etc.
b. $x = .555$, $y = .079$, $z = .772$

5.2.5.



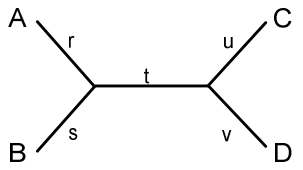
Topologically, the trees are the same as unrooted trees. They are not the same metrically. Note, for instance, FM assigns a branch length of 0 to the internal edge, while UPGMA assigns .125.

5.2.6. a. In order for a molecular clock hypothesis to hold, all the terminal vertices would have to be equidistant from the root. This is impossible. The root cannot be placed at the internal node since the edge lengths are different. Moreover, the root cannot be placed on any of the three edges since no two of the edges have the same length.

b. Since the two shortest edge lengths are equal to .1, it is possible to assume a molecular clock. The root would have to be placed on the edge of length .2 at a distance of .05 from the internal node. Then all terminal vertices are .15 from the root.

c. Here two of the edge lengths are equal, but their length .2 is larger than the length of the third edge. This means it is not possible to locate the root on either of the longer edges nor the shorter edge and achieve equal distances from the root. Of course, the internal node could not serve as a root either, if a molecular clock is to be assumed.

5.2.7. a.



b. $d_{AB} = r+s$, $d_{AC} = r+t+u$, $d_{AD} = r+t+v$, $d_{BC} = s+t+u$, $d_{BD} = s+t+v$, $d_{CD} = u+v$; As this is a system of six equations in only five unknowns, in general there will not be a solution.

c. Answers may vary; one possibility follows. For the distances $d_{AB} = .2$, $d_{AC} = .3$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system does not have a solution, whereas for the distances $d_{AB} = .17$, $d_{AC} = .32$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system has a solution.

5.2.8. a. For calculating these measures of errors, the length b was assigned to zero.

	s_{FM}	s_F	s_{TNT}
FM tree	.4699	.6370	.2592
UPGMA tree	.4933	.8968	.3515

The FM tree is a better fit to the data according to all three of these measures.

b. All of these formulas give 0 if a tree exactly fits the data.

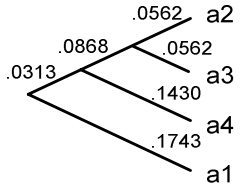
s_F simply sums the absolute value of the difference between the tree lengths and the original distance data. The absolute value prevents negative differences from canceling with positive ones. All deviations of tree lengths from the data are treated identically.

s_{TNT} sums the squares of the differences between tree lengths and distance data (again preventing cancelation), then takes the square root. This is reminiscent of the formula for standard deviation. This measure penalizes large differences more than s_F does, while weighing small differences less: If $|d_{ij} - e_{ij}| < 1$, then $(d_{ij} - e_{ij})^2$ is even smaller, while if $|d_{ij} - e_{ij}| > 1$, then $(d_{ij} - e_{ij})^2$ is larger.

s_{FM} measures the differences between tree lengths and data as proportions. Other than that, it is similar to s_{TNT} in that in penalizes large differences

in the proportions much more than small ones. When tree edge lengths vary greatly in size, s_{FM} will, unlike the other two measures, not allow greater proportional errors in the short edges than the long ones.

5.2.9. a.

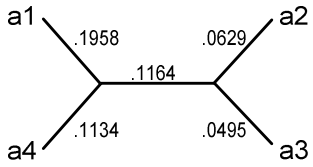


b.

	a1	a2	a3	a4
a1		.3486	.3486	.3486
a2			.1124	.2860
a3				.2860

From the table, the distance between the first pair of taxa joined, a2 and a3, agrees with the original distance data (up to rounding error). The other distances approximate the original distances, and in fact represent averages, with duplication occurring because of the molecular clock hypothesis. Since we know the sequences were created assuming a molecular clock hypothesis, however, we should assume that this tree more accurately reflects the relationships between the sequences than a FM tree does.

5.2.10. a.



b.

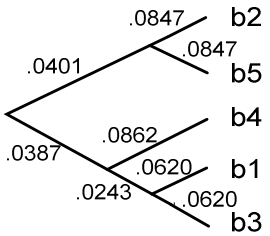
	a1	a2	a3	a4
a1		.3751	.3617	.3092
a2			.1124	.2927
a3				.2793

The distances in this table match up reasonably well with the original distances. In particular, since FM joins a pair of taxa at each step, you find several matches (within rounding error) between the distance tables.

Even though the FM tree produced here appears to match the distance table better than the UPGMA tree of the last problem, since these sequences were created with a molecular clock hypothesis, the tree FM produces should not be preferred to the UPGMA tree. This is an example of *overfitting* the data, by using a more general approach than actually best describes the simulated evolution.

This is analogous to an issue raised in Chapter 8 while studying the method of least squares: while it is possible to fit a degree five polynomial exactly to six data points, if the points are approximately linearly related, this may disguise the true trend of the data.

5.2.11. a.

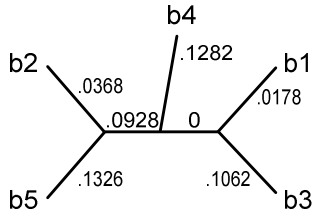


b.

	b1	b2	b3	b4	b5
b1		.2497	.1240	.1724	.2497
b2			.2497	.2497	.1694
b3				.1724	.2497
b4					.2497

The distances computed from the UPGMA tree are not in very close agreement with the data. Since we know that the data was produced without a molecular clock, we should not expect UPGMA to be able to fit the data well. UPGMA makes an assumption that we know is incorrect.

5.2.12. a.



b.

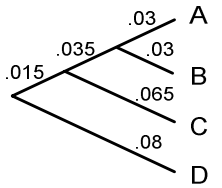
	b1	b2	b3	b4	b5
b1		.1474	.1240	.1460	.2432
b2			.2358	.2578	.1694
b3				.2344	.3316
b4					.3536

The tree distances agree reasonably well with the distance data. If we believe a molecular clock hypothesis is invalid, then the FM tree might be a better reconstruction of evolutionary relationships than the UPGMA tree. (Recall however, that they differ only metrically, not topologically.)

5.2.13. a. There is no way that all four taxa can be equidistant from a root: Since A and C are not equidistant from the internal node to which they are both joined, if a molecular clock is assumed, the root would have to be on the edge leading to taxon C. If this were the case, then it would be impossible for B and D to be equidistant from the root.

b.

	A	B	C	D
A		.06	.12	.14
B			.14	.12
C				.22



Note: If the labels C and D are exchanged in this tree, then the resulting tree would also be a UPGMA tree, since there were two minima in the collapsed data table.

c. Notice that the UPGMA tree constructed has the wrong topological structure. Since A and B are closest in distance, they are joined first by UPGMA, even though this results in the wrong topology. Neighbor joining, introduced in the next section, will not make this mistake.

d. FM creates the same topological tree as UPGMA, so it too will construct an incorrect tree topology.

5.3. Tree Construction: Distance Methods – Neighbor Joining

Warning: Trees are not drawn to scale.

5.3.1. a. Let G the group of all taxa other than Si and Sj. Then

$$\begin{aligned} d(\mathrm{S}i, V) &= \frac{1}{2}(d(\mathrm{S}i, G) + d(\mathrm{S}i, \mathrm{S}j) - d(\mathrm{S}j, G)) \\ &= \frac{(\sum_{k\neq j} d(\mathrm{S}i, \mathrm{S}k))}{2(N-2)} + \frac{d(\mathrm{S}i, \mathrm{S}j)}{2} - \frac{(\sum_{l\neq j} d(\mathrm{S}j, \mathrm{S}l))}{2(N-2)} \\ &= \frac{d(\mathrm{S}i, \mathrm{S}j)}{2} + \frac{R_i - R_j}{2(N-2)}. \end{aligned}$$

Notice the $d(\mathrm{S}i, \mathrm{S}j)$ terms cancel in the expression $R_i - R_j$ to make the last equality hold.

b. This follows from determining edge lengths for the three-taxa tree with Sk, Si, Sj as terminal nodes and V as internal vertex.

5.3.2. a. $R_1 = 1.52$, $R_2 = 2.52$, $R_3 = 1.48$, $R_4 = 1.86$, and M is given by the table:

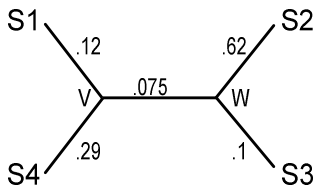
	S2	S3	S4
S1	−2.38	−2.44	−2.56
S2		−2.56	−2.44
S3			−2.38

b. $d(\mathrm{S1}, V) = .12$, $d(\mathrm{S4}, V) = .29$

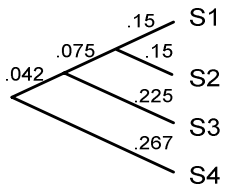
c. $d(\mathrm{S2}, V) = .695$, $d(\mathrm{S3}, V) = .175$

d. $d(\mathrm{S2}, W) = .62$, $d(\mathrm{S3}, W) = .1$, $d(V, W) = .075$

e.



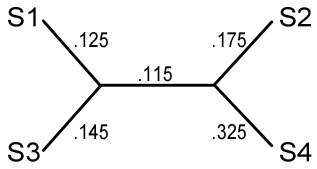
5.3.3. a.



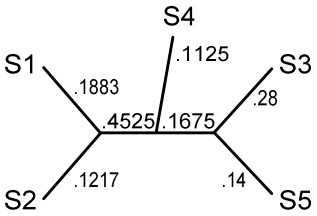
The UPGMA tree does not recover the correct topology. Note: Another UPGMA tree has taxa S3 and S4 interchanged above.

b. Neighbor Joining does recover the correct metric and topological tree.

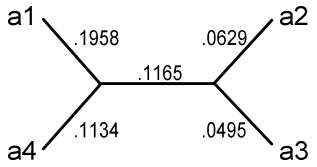
5.3.4. a. The Neighbor Joining tree (shown below) has the same unrooted topological structure as the UPGMA one, but a different metric structure.



b. The Neighbor Joining tree (shown below) differs both topologically and metricly from the FM one.

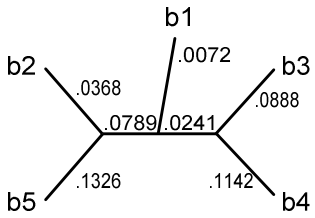


5.3.5. a.



The unrooted topological structure of the UPGMA, FM, and NJ trees are all the same. All trees show that `a1` is furthest from the neighbors `a2` and `a3`. The metric features of the FM and NJ trees are essentially the same, and differ from the UPGMA tree. (However, since this data was simulated with a molecular clock, we might still prefer the UPGMA tree to the others.)

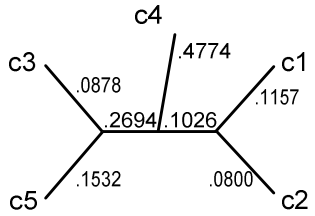
b.



The NJ tree is topologically and metrically different from the UPGMA and FM trees. In particular, NJ chose `b3` and `b4` to be neighbors, even though `b3` is closer in distance to `b1` than `b4`. FM could not do this. Notice that NJ, like FM, created branches of quite different lengths. As we know this data was not simulated with a molecular clock, we should prefer the NJ tree to either of the others.

5.3.6. a. A frequency table for every pair of sequences shows that transitions are more common than transversions. In addition, it appears that all transitions occur at about the same rate, and all transversions occur at about the same rate.

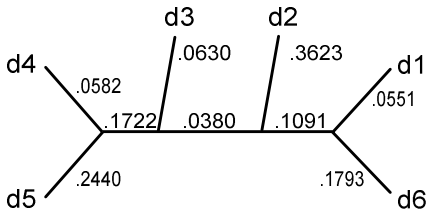
b.



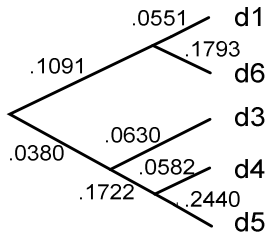
c. The tree does not appear to support a molecular clock hypothesis. Assuming a molecular clock, the best location for the root is either along the edge joining `c4` to the main tree or along the edge of length .2694. However, with either choice there is still much variation in distances between the root and taxa.

5.3.7. a. Frequency tables for pairs of sequences show transitions are more common than transversions. In addition, it appears that all transitions occur at about the same rate, and all transversions occur at about the same rate. Thus using the Kimura 2-parameter distance seems a reasonable choice.

b.



c. The outgroup appears to be d2.

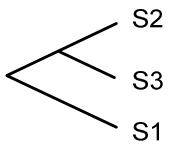


5.4. Tree Construction: Maximum Parsimony

5.4.1. a. Both trees have parsimony score 3.
b. The most parsimonious trees have score 2.
c. Since there are only four bases, we can always find a tree that requires three substitutions. For example, if we create one joined cluster of taxa with A's, another cluster with G's, another with C's, and a last with T's, then join up these clusters, the resulting tree will have parsimony score 3.

5.4.2. a. The tree on the left has parsimony score 7; the tree on the right has parsimony score 8.
b. The third unrooted tree has parsimony score 10. Therefore, the tree pictured on the left is the most parsimonious unrooted tree.

5.4.3. a. Sites 3, 6, 8, and 11
b. S1 and S4 are neighbors and S2 and S3 are neighbors. The parsimony score for the rooted tree relating them is 5.
c.



5.4.4. Suppose there are n sequences. If, at a particular site, $n-1$ sequences are in agreement and the remaining sequence disagrees with these, then the mutation count must be increased by one. If there are n_1 such sites, then the mutation count must be augmented by n_1. If, at a particular site, $n-2$ of the sequences are in agreement and the two remaining sequences disagree with each other and all the other sequences, then the count is increased by two. If there are a total of n_2 such sites, then the mutation count is augmented by $2n_2$. Similarly, for $3n_3$.

5.4.5. Both trees require three mutations.

5.4.6. In order for a notion of informative sites to make sense, there must be at least four sequences being compared, since an informative site is one for which at least two bases occur twice each. Since parsimony scores measure the fitness of unrooted trees and there is only one unrooted tree relating three taxa, there is no need for informative sites when we want to compare three taxa.

5.4.7. a. There are n sites and one of 4 bases must occur at each site. For the first sequence, there are 4 possibilities for the base occurring there; for the second sequence there are also 4 possibilities. This gives a total of $4^2 = 16$ possible

patterns for two sequences. If we consider a third sequence, since there are 4 possibilities for the base occurring at the site, there are four times as many patterns or $4^3 = 4(16)$ total patterns for three sequences. In general, for n sequences, there are 4^n possible patterns.

b. There are n possible ways to select the sequence that does not agree with the other $n-1$ sequences. There are 4 bases which can occur at this particular sequence and 3 remaining choices for the base that occurs in the $n-1$ remaining sequences. This gives a total of $(4)(3)n$ possibilities.

c. There are $\frac{n(n-1)}{2}$ ways to select the two sequences that disagree with the other $n-2$ sequences (n choices for the first sequence, $n-1$ for the second, divide by 2! since order does not matter). There are 4 ways to choose the base appearing at the first of these, 3 ways to choose the base occurring at the second of these, and 2 ways to choose the base occurring at the other $n-2$ sequences. Thus, there are $(4)(3)n(n-1)$ possible ways to obtain this pattern.

d. This is similar to part (c). There are $\frac{n(n-1)(n-2)}{3!}$ ways to select the three sequences that disagree with the other $n-3$ sequences. There are 4 ways to choose the base appearing at the first of these, 3 ways to choose the base occurring at the second of these, 2 ways to choose the base occurring at the third of these, and 1 way to choose the base occurring at the other $n-3$ sequences. Thus, there are $(4)n(n-1)(n-2)$ possible ways to obtain this pattern.

e. The number of informative patterns is 4^n - $((4)(3)n + (4)(3)n(n-1) + (4)n(n-1)(n-2))$. Since 4^n grows much more rapidly than n^3, most patterns are informative.

5.4.8. The parsimony score would be $\sum f_{pattern} p_{pattern}$, where the sum is taken over all patterns.

5.4.9. a. Informative patterns for four taxa contain two bases, each occurring twice. There are 3 ways to make such patterns without regard to base choices.

b. 25

5.4.10. a. 8.44% or 38/450

b. There are three unrooted trees relating four taxa.

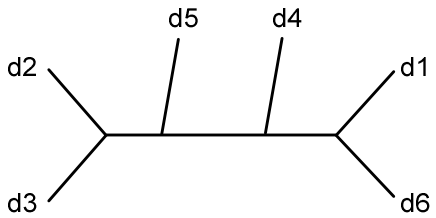
c. Using the first ten informative sites, `a1` and `a4` are neighbors, as are `a2` and `a3`. This branching structure is in agreement with both the UPGMA and the NJ trees.

5.4.11. a. 30.8% or 244/792

b. 105

c. For the tree topology reported in Problem 5.3.7b, the parsimony score is 18.

d. Answers may vary. Interchanging the taxa `d2` and `d3` on the tree of part (c) results in a parsimony score of 17. For the tree below, the parsimony score is 20.



e. You have calculated parsimony scores for only five out of 105 possible trees (4.76%), using only ten out of 244 informative sites (4.10%). It is hard to have