4.5. Phylogenetic Distances

4.5.1. .1367

4.5.2. a. The Jukes-Cantor distance is .1102158097.

b. The Kimura 2-parameter distance is .1102165081.

c. Since the two distance calculations agree to several decimal spaces, we might hypothesize (if the problem had not already told us) that the data is fit reasonably well by the Jukes-Cantor model. Notice that the Kimura 2-parameter distance reports more mutations (including hidden mutations) than the Jukes-Cantor distance.

4.5.3. a. .2224580274

b. .2308224444

c. The Kimura 2-parameter distance is probably a better choice (assuming we did not already know that the sequences were created with the Kimura 2-parameter model). The frequency table shows a definite pattern of more transitions than transversions. Notice too that the distances differ in the second decimal position.

4.5.4. Jukes-Cantor simulation: $d_{K3} = .1104707856$ and $d_{LD} = .1105916542$. Notice these are about the same as the Jukes-Cantor distance, since that model is a special case of the more general ones.

Kimura 2-parameter simulation: $d_{K3} = .2308544863$ and $d_{LD} = .2337622488$. Notice these are about the same as the Kimura 2-parameter distance, since that model is a special case of the more general ones.

4.5.5. Graph the Jukes-Cantor distance on a graphing calculator or computer.

a. If the sequences are identical, then $p = 0$. This means the Jukes-Cantor distance is $-.75\log(1) = 0$.

b., c. Mathematically, if two sequences differ in more than $3/4$ of the sites, then $p > 3/4$. Then the Jukes-Cantor distance formula requires taking the logarithm of a negative number, which is impossible. This is not a limitation with real data. If we took two sequences that were in no way related, we would expect that about $1/4$ of the sites agree and about $3/4$ of the sites disagree, since with a uniform distribution of bases about 25% of the time the two sequences should agree if everything is chosen at random. For related sequences the formulas for the Jukes-Cantor model derived in the last section show p is at most $3/4$, and in practice p is usually much less than $3/4$. Notice that the Jukes-Cantor distance gets huge as the values of p get close to .75. This is desirable, since distances should be large when comparing sequences that appear almost unrelated.

4.5.6.

$$p = \frac{3}{4} - \frac{3}{4}(1 - \frac{4}{3}\alpha)^t \implies (1 - \frac{4}{3}p) = (1 - \frac{4}{3}\alpha)^t \implies$$
$$\ln(1 - \frac{4}{3}p) = t\ln(1 - \frac{4}{3}\alpha) \implies t = \frac{\ln(1 - \frac{4}{3}p)}{\ln(1 - \frac{4}{3}\alpha)}.$$

4.5.7. Substituting $(1-q)$ for p yields $d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = -\frac{3}{4}\ln(1 - \frac{4}{3}(1-q)) = -\frac{3}{4}\ln(\frac{4}{3}q - \frac{1}{3}) = -\frac{3}{4}\ln(\frac{4q-1}{3})$.

4.5.8. Some numerical comparisons are given in the table below. The graphs of $y = \ln(1+x)$ and $y = x$ are very close when x is around 0. In fact, they are tangent to one another at the point $(0, 0)$.

x	$-.1$	$-.01$	$-.001$	$-.0001$
$\ln(1+x)$	$-.1053605157$	$-.0100503359$	$-.0010005003$	$-.0001000050$
x	$.0001$	$.001$	$.01$	$.1$
$\ln(1+x)$	$.0000999950$	$.0009995003$	$.0099503309$	$.0953101798$

4.5.9. Since $f'(x) = \frac{1}{1+x}$, the slope of the tangent line is $f'(0) = 1$. The tangent line passes through the point $(0, 0)$. Using the point-slope formula, the equation of the tangent line to $f(x) = \ln(1+x)$ is $g(x) = x$. Thus, for values of x near 0, $f(x) \approx g(x) = x$.

4.5.10. $d(\mathrm{S}_0, \mathrm{S}_1) + d(\mathrm{S}_0, \mathrm{S}_2) = d(\mathrm{S}_1, \mathrm{S}_0) + d(\mathrm{S}_0, \mathrm{S}_2)$ by symmetry. By additivity, this equals $d(\mathrm{S}_1, \mathrm{S}_2)$.

4.5.11. Two transitions at a particular site will result in a return to the original base and thus a hidden mutation. For example, $A \to G \to A$ is a hidden mutation. Two transversions may produce a hidden mutation, but often don't (e.g., $A \to C \to G$). If transitions are more likely than transversions, then hidden mutations are more likely.

4.5.12. a. If there are a lot of point mutations at a site, then hidden mutations are more likely. Thus p, the proportion of observed point mutations, is an underestimate of the true proportion of point mutations.

b. When p is small, few point mutations are observed. Since little mutation is observed, it's reasonable to assume little occurred, and therefore that few mutations have been hidden. Thus p should be a good estimate of the proportion of point mutations.

4.5.13. The Kimura 3-parameter distance is given by $d_{K3} = -\frac{1}{4}(\ln(1-2\beta-2\gamma) - \ln(1-2\beta-2\delta) - \ln(1-2\beta-2\gamma))$. Substituting $\alpha/3$ for β, γ, and δ gives

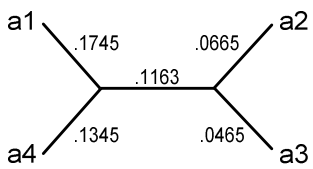
$$\begin{aligned} d &= -\frac{1}{4}(\ln(1-2\alpha/3-2\alpha/3) - \ln(1-2\alpha/3-2\alpha/3) - \ln(1-2\alpha/3-2\alpha/3)) \\ &= \frac{1}{4}(3\ln(1-4\alpha/3)) = d_{JC}. \end{aligned}$$

4.5.14. The distance from the Jukes-Cantor formula is not equal to .4 and may occasionally even be quite far away. Lots of factors are responsible for the discrepancy: the length of the sequences is relatively short; simulated data is always an imperfect reflection of the underlying model; the larger p is, the greater effect a small variation in it has on the reconstructed value of αt; etc.

4.5.15. a. The Jukes-Cantor distances are given in the table below.

	a1	a2	a3	a4
a1		.3721	.3648	.3091
a2			.1125	.2958
a3				.2763

b. Answers will vary. One possibility is



This tree was constructed by observing that a2 and a3 were closest and so perhaps should have an immediate common ancestor. (This also means a2 and