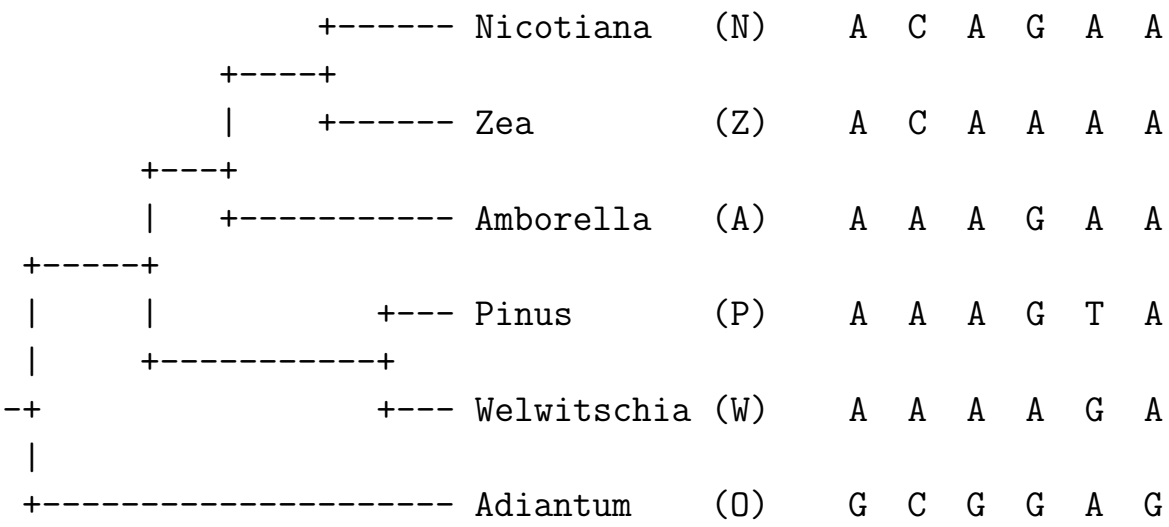


The first several problems refer to the phylogeny of several plant genera shown above. Letters in parentheses are one-letter abbreviations you may use below when asked to sketch trees (O is for outgroup). The data are the first six varied sites from the *rbcL* gene.

Recall that there are $1 \times 3 \times \cdots \times (2n-3) \equiv (2n-3)!!$ rooted binary tree topologies with n leaves ($n \geq 2$).

1. (3 points) Complete the following table showing the number of possible fully resolved rooted tree topologies for up to six taxa.

   Solution:

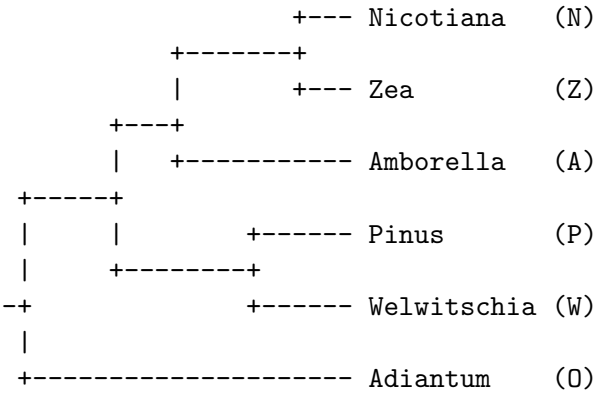
# of taxa	1	2	3	4	5	6
# of rooted trees	1	1	3	15	105	945

2. (4 points) Sketch a tree with a different tree topology for these six taxa that has fewer labeled histories than the one shown.

   Solution: There are many possible solutions. The simplest are those where every split separates a single taxon as each of these is consistent with only one labeled history.

3. (4 points) Sketch a tree with the same tree topology as above that has a different labeled history than the one shown above.

   Solution: All we need to do is to change the order of two splits. Here is one possible solution.



4. (5 points) The tree has a clade comprising of *Nicotiana*, *Zea*, and *Amborella*. What proportion of all fully resolved rooted trees with these taxa contain the same clade?

   Solution:

$$\frac{r(3)r(4)}{r(6)} = \frac{45}{945} = \frac{1}{21}$$

   There are $r(3)$ subtrees for the clade. There are $r(4)$ ways to make a rooted tree for the remaining three taxa and node which is the most recent common ancestor of the clade.

5. (5 points) What proportion of all fully resolved rooted trees with these taxa contain the N/Z/A clade and have *Adiantum* as an outgroup?

   Solution:

$$\frac{r(3)r(3)}{r(6)} = \frac{9}{945} = \frac{1}{105}$$

   As before, there are $r(3)$ subtrees for the clade. There are also $r(3)$ subtrees for the two non-outgroup taxa and the most recent common ancestor of the clade. Thus, there are 9 such trees which include node O as the outgroup and the clade. The question asked for the proportion of *all trees* with the clade *and* the outgroup which is why the denominator is still 945. Had the question been about the probability of the clade among those with the outgroup, the denominator would have been 105 instead.

6. (4 points) The figure shows an alignment of six varied sites. Which of these (from sites 1–6) are parsimony informative?

   Solution: Sites 2 and 4 are the only ones with at least two bases appearing at least twice.

7. (5 points) Find the parsimony score for the displayed tree for the second site.

   Solution: By Fitch's algorithm or some trial and error, you can find that two substitutions are needed for this tree and site.

8. (4 points) Complete the following sentence:

   Solution: *If a site is not parsimony informative and it contains x distinct bases, then it adds $x-1$ to the parsimony score for all possible trees.*

   If there are x distinct bases, one would be present at the root and the additional $x-1$ bases could each be added by a single substitution.

The following problems refer to this Q-matrix which parameterizes a continuous-time Markov chain for molecular evolution of a single site. The order of the states is A, C, G, T.

$$Q = \begin{pmatrix} -12.5 & 3.0 & 8.0 & 1.5 \\ 1.5 & -8.5 & 4.0 & 3.0 \\ 3.0 & 3.0 & -7.5 & 1.5 \\ 1.5 & 6.0 & 4.0 & -11.5 \end{pmatrix}$$

9. (4 points) If a base is currently a G and there is a substitution, what is the probability that the new base will be a C? Report your answer as a fraction.

   Solution:

$$\frac{3.0}{7.5}$$

10. (4 points) The stationary distribution of the Markov chain satisfies $\pi_A + \pi_C + \pi_G + \pi_T = 1$. Write down one more equation using values from Q that the stationary distribution must satisfy.

    Solution: The total rate to leave A must equal the rate going in.

$$12.5\pi_A = 1.5\pi_C + 3.0\pi_G + 1.5\pi_T$$

    Other choices are possible using the other columns of Q for the coefficients.

11. (4 points) The displayed Q matrix has not been scaled. If we choose an initial base from the stationary distribution and observe the process for one unit of time, would we expect to see more than or fewer than one substitution? Briefly explain (one sentence or phrase).

    Solution: More. The average number of substitutions per unit time is 7.5 or more from every base. *(This is in contrast to the expected time to wait before a substitution which is less than one.)*

12. (4 points) The stationary distribution is $\pi = (\pi_A, \pi_C, \pi_G, \pi_T) = (0.15, 0.3, 0.4, 0.15)$. Write down a numerical estimate of the probability transition matrix $P(t) = \mathrm{e}^{Qt}$ for a large time such as $t = 1000$.

    Solution: Each column will have all elements approximately equal to the corresponding value from the stationary distribution. This can be interpreted, for example, that the chance of being in state A after a long time is about 0.15 regardless of where the process begins.

$$P(t) \doteq \begin{pmatrix} 0.15 & 0.3 & 0.4 & 0.15 \\ 0.15 & 0.3 & 0.4 & 0.15 \\ 0.15 & 0.3 & 0.4 & 0.15 \\ 0.15 & 0.3 & 0.4 & 0.15 \end{pmatrix}$$