Genetics 629 **Solutions to Homework #1** Exam on September 22, 2009

1. Create your own fully resolved unrooted tree with $n = 8$ taxa.

   There will be many different answers depending on your choice. I will answer for this unrooted tree:
   `(((1,2),3),(4,5),((6,7),8));` .

   (a) How many rooted tree topologies are consistent with your unrooted tree?
   Solution: There are 13 edges on the tree (in general, $2n - 3$), so there are 13 rooted tree topologies.

   (b) How many internal nodes are in your unrooted tree?
   Solution: There are 6 internal nodes ($n - 2$ in general).

   (c) How many edges are in your unrooted tree?
   Solution: There are 13 edges, as above.

   (d) Find a formula that counts the number of internal nodes and edges in a general fully resolved unrooted tree with n leaves (taxa).
   Solution: $\#(\text{internal nodes}) = n - 2$, $\#(\text{edges}) = 2n - 3$.

2. Select a specific rooting of your tree from the previous problem and sketch it such that all leaves are the same distance from the root.

   I chose to root the tree so that the triple ((6,7),8) was an outgroup. The rooted tree has this topology:
   `((((1,2),3),(4,5)),((6,7),8));` .

   (a) How many different labeled histories are there for the tree you sketched?
   Solution: Not a fair question for a large semi-balanced tree .... The general algorithm is when combining subtrees with k_1 and k_2 taxa which include m_1 and m_2 labeled histories respectively, the new tree has $m = m_1 \times m_2 \times (k_1 + k_2 - 2)!/((k_1 - 1)!(k_2 - 1)!)$ labeled histories (and $k = k_1 + k_2$ taxa). The reason is that there are m_i ways to pick the relative orders of the internal nodes in each respective subtree (for $i = 1, 2$) and once these are selected, there are $k_1 + k_2 - 2$ internal nodes to place in order and we need to choose which $k_1 - 1$ are from the first subtree (leaving the others for the second subtree).
   For my choice, there are 45 possible labeled histories. Here is how I did it. For each subtree I can find the pair (m, k) using the above approach.

$$
\begin{aligned}
1 &: (1,1) \\
2 &: (1,1) \\
(1,2) &: (1 \times 1 \times (0!/(0!0!)), 1+1) = (1,2) \\
((1,2),3) &: (1 \times 1 \times (1!/(0!1!)), 1+2) = (1,3) \\
4 &: (1,1) \\
5 &: (1,1) \\
(4,5) &: (1 \times 1 \times (0!/(0!0!)), 1+1) = (1,2) \\
(((1,2),3),(4,5)) &: (1 \times 1 \times (3!/(2!1!)), 3+2) = (3,5) \\
6 &: (1,1) \\
7 &: (1,1) \\
(6,7) &: (1 \times 1 \times (0!/(0!0!)), 1+1) = (1,2) \\
((6,7),8) &: (1 \times 1 \times (1!/(0!1!)), 1+2) = (1,3) \\
((((1,2),3),(4,5)),((6,7),8)) &: (3 \times 1 \times (6!/(4!2!)), 5+3) = (45,8)
\end{aligned}
$$

You **do not** need to be able to do a calculation like this. Just know that some topologies have multiple possible labeled histories and that balanced tree topologies have more than imbalanced ones.

(b) Briefly explain why there are more labeled histories than rooted tree topologies in general for n taxa where $n \geq 4$.

Solution: Some topologies have more than one possible labeled history, so the total is larger.

3. Consider a uniform probability distribution on rooted tree topologies with the species A, B, C, D, E, F, G, and H.

(a) What is the probability that the A, B, and C form a clade?
Solution: If we let $r(n)$ be the number of rooted trees of size n, then the answer is

$$\frac{r(3) \times r(6)}{r(8)} = \frac{(1 \times 3)(1 \times 3 \times 5 \times 7 \times 9)}{1 \times 3 \times 5 \times 7 \times 9 \times 11 \times 13} = \frac{3}{143} \doteq 0.021$$

There are $r(3)$ ways to build a rooted tree for the clade of three taxa and $r(5+1)$ ways to build a rooted tree for the remaining 5 taxa plus the root of the subtree for the clade, and the product counts the number of ways to combine these differently.

(b) What is the probability that A, B, D, and E form a clade?
Solution: As above,

$$\frac{r(4) \times r(5)}{r(8)} = \frac{(1 \times 3 \times 5)(1 \times 3 \times 5 \times 7)}{1 \times 3 \times 5 \times 7 \times 9 \times 11 \times 13} = \frac{5}{429} \doteq 0.012$$

(c) Repeat the previous two problems if we assume that H is an outgroup (so that A–G have a common ancestor that is not an ancestor of H).
Solution: If H is an outgroup, there are $r(7)$ possible rooted trees for the remainder. The calculations are then very similar. For clade A,B,C, the probability is

$$\frac{r(3) \times r(5)}{r(7)} = \frac{(1 \times 3)(1 \times 3 \times 5 \times 7)}{1 \times 3 \times 5 \times 7 \times 9 \times 11} = \frac{1}{33} \doteq 0.030$$

and the probability for clade A,B,D,E is

$$\frac{r(4) \times r(4)}{r(7)} = \frac{(1 \times 3 \times 5)(1 \times 3 \times 5)}{1 \times 3 \times 5 \times 7 \times 9 \times 11} = \frac{5}{231} \doteq 0.022$$

4. Consider a set of four taxa A, B, C, D and a uniform probability distribution on rooted tree topologies with n taxa.

(a) Find the probability that the set of four taxa form a clade for $n = 5, 10, 15, 20$ taxa.

(b) Comment on any patterns in these probabilities.

Solution: The same general formula holds, but this time $k = 4$ is fixed and n varies. The probabilities as n increases are as follows.

n	Probability
5	$r(2)r(4)/r(5) = (1)(5!!)/7!! = 1/7 \doteq 0.14286$
10	$r(7)r(4)/r(10) = (11!!)(5!!)/17!! = 1/221 \doteq 0.00452$
15	$r(12)r(4)/r(15) = (21!!)(5!!)/27!! = 1/1035 \doteq 0.00097$
20	$r(17)r(4)/r(20) = (31!!)(5!!)/37!! = 1/2849 \doteq 0.00035$

A general solution to this particular problem is $15/((2n-7)(2n-5)(2n-3))$ which can be found by simplifying $r(n-3)/r(n)$. The probabilities decrease rapidly (on the order of n^{-3}) when the tree size increases. So, for increasingly large trees, the probability of any given clade of four taxa gets smaller and smaller.

```
                        10           20           30
                         +            +            +
alligator  GTG AAC TTC CAC --- CGT TGA CTC TTC TCT
goose      GTG ACC TTC ATC AAC CGA TGA CTA TTT TCT
swan       GTG ACC TTC ATC AAC CGA TGA CTA TTT TCC
finch      ATG ACA TAC ATT AAC CGA TGA TTA TTC TCA
osprey     ATG ACA TTC ATC AAC CGA TGA CTA TTC TCA
```

5. The data set above is the first 30 bases of the *cytochrome oxidase I* mitochondrial gene from alligator and four species of birds. (The sequences are separated by space every three bases to help readability.) How many of these sites are unvaried? How many of these sites are parsimony informative?

6. Assume that goose and swan form a clade and that alligator is the outgroup. There are then three possible phylogenetic trees to relate the five species. For each of these possible trees, compute the parsimony score on the basis of the displayed data. Which tree is the maximum parsimony estimate?

Solution: Fifteen sites are unvaried, (18 if you add in the three with a gap in alligator). Of the 12 complete sites with some variation, only four are parsimony informative: 1, 6, 27, and 30.

Sites 1, 6, and 30 require one fewer substitution for the tree with a (finch,osprey) clade than the other two possibilities and site 27 requires only one for each of these trees with the goose/swan clade. So, the most parsimonious informative tree is

```
(alligator, ( (finch,osprey) , (goose,swan) ) );
```

It has a score of 12 (5 from the parsimony informative sites) while the other two trees score 15 each. If I counted right.

7. Complete this phrase: A site will be parsimony informative if and only if ....

Solution: there are two pairs of taxa that share a base with each other that is different from the base for the other pair.

8. In a few sentences, describe a situation where the method of maximum parsimony may be likely to choose the incorrect tree topology.

Solution: If the true tree has long branches that are not adjacent (such as two fast-evolving species that are not sister taxa), then parsimony may be prone to grouping these fast evolvers together in a clade, which would be incorrect.

$$Q = \{q_{ij}\} = \begin{pmatrix} -1.03 & 0.26 & 0.52 & 0.25 \\ 0.36 & -1.49 & 0.13 & 1.00 \\ 1.44 & 0.26 & -1.95 & 0.25 \\ 0.36 & 1.04 & 0.13 & -1.53 \end{pmatrix}$$

9. Consider the evolution of a single site of a DNA sequence according to a continuous-time Markov chain modulated by the rate matrix Q shown above where the bases are in order A, C, G, T. Assume that the base at time 0 is chosen at random from A, C, G, T with probabilities 0.36, 0.26, 0.13, 0.25 respectively.

   Compute the probability density of this sequence of events. At time $t = 0$, the site is an A. At time 1.2, there is a substitution from an A to a T. At time 2.1 (after a further time of 0.9), there is a substitution from a T to a C. There are no further substitutions during the next 0.4 time units before time 2.5.

   Solution:
   $$0.36 \times \left(1.03\mathrm{e}^{-1.03(1.2)}\frac{0.25}{1.03}\right) \times \left(1.53\mathrm{e}^{-1.53(0.9)}\frac{1.04}{1.53}\right) \times \left(\mathrm{e}^{-1.49(0.4)}\right) \doteq 0.0033781$$

10. Verify that $\pi = c(0.36, 0.26, 0.13, 0.25)$ is the stationary distribution of the continuous-time Markov chain with rate matrix Q.

    Solution: Show that $\pi^T Q = (0, 0, 0, 0)^T$. This implies that the expected rate of substitutions into and out of each state is equal.

    Here are the four calculations:

    $$\begin{aligned} 0.36(-1.03) + 0.26(0.36) + 0.13(1.44) + 0.25(0.36) &= 0 \\ 0.36(0.26) - 0.26(1.49) + 0.13(0.26) + 0.25(1.04) &= 0 \\ 0.36(0.52) + 0.26(0.13) - 0.13(1.95) + 0.25(0.13) &= 0 \\ 0.36(0.25) + 0.26(1.00) + 0.13(0.25) - 0.25(1.53) &= 0 \end{aligned}$$

11. What number would you need to multiply to every entry in the Q matrix to change the scale so that one unit of time would represent one substitution per site?

    Solution: The average rate of substitution is

    $$0.36(1.03) + 0.26(1.49) + 0.13(1.95) + 0.25(1.53) = 1.3942$$

    Multiplying Q by the reciprocal of this, $1/1.3942 = 0.7172572$ rescales time so that one unit of time represents one expected substitution.

12. Which base will have the longest average dwell-time before a substitution?

    Solution: The base with the longest average dwell-time will be the one with the lowest rate — A with a rate of 1.03.

13. If the base is A and there is a substitution, what is the probability that the new base will be a G?

    Solution: The probability of a switch to G given you begin at A is $0.52/1.03 \doteq 0.505$.