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

Trees Recall the following formula:

- There are $1 \times 3 \times \cdots \times (2n-5) \equiv (2n-5)!! \equiv u(n)$ unrooted binary tree topologies with n leaves ($n > 2$).
- There are $1 \times 3 \times \cdots \times (2n-3) \equiv (2n-3)!! \equiv r(n)$ rooted binary tree topologies with n leaves ($n > 2$).
- There are $\frac{n!(n-1)!}{2^{n-1}} \equiv \ell(n)$ labeled histories for n taxa.

1. Create your own fully resolved unrooted tree with $n = 8$ taxa.
   (a) How many rooted tree topologies are consistent with your unrooted tree?
   (b) How many internal nodes are in your unrooted tree?
   (c) How many edges are in your unrooted tree?
   (d) Find a formula that counts the number of internal nodes and edges in a general fully resolved unrooted tree with n leaves (taxa).
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.
   (a) How many different labeled histories are there for the tree you sketched?
   (b) Briefly explain why there are more labeled histories than rooted tree topologies in general for n taxa where $n \geq 4$.
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?
   (b) What is the probability that A, B, D, and E form a clade?
   (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).
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.

Parsimony

```
                        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?

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

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

Molecular Evolution

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

Recall these facts about the Exponential diatribution.

- Single parameter λ is called the *rate*.
- Density is $f(t) = \lambda \mathrm{e}^{-\lambda t}, \quad \text{for } t \geq 0.$
- Density satisfies $\displaystyle\int_0^\infty f(t)\mathrm{d}t = 1.$
- Cumulative distribution function is $\mathsf{P}\{T \leq t\} = F(t) = \displaystyle\int_0^t f(s)\mathrm{d}s = 1 - \mathrm{e}^{-\lambda t}.$
- Tail probability (probability of no event in time t) is $\mathrm{e}^{-\lambda t}$.
- Mean is $1/\lambda$.

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.

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.

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?

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

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