Genetics 629 **Solution to Homework #2** Exam on October 8, 2009

Distance Methods

	A	B	C	D
A	0	10	6	4
B	10	0	12	8
C	6	12	0	6
D	4	8	6	0

1. Apply the UPGMA algorithm to the matrix to find the UPGMA tree.

   Solution: (((A:2,D:2):1,C:3):2,B:5);

2. Assume that the first step of the neighbor-joining algorithm joins taxa A and C, that the edge to A has length 0.5, the edge to C has length 5.5, and that the new distance matrix relating the remaining groups is as follows.

	A/C	B	D
A/C	0	8	2
B	8	0	8
D	2	8	0

   Find the neighbor-joining tree.

   Solution: ((A:0.5,C:5.5):1,B:7,D:1);

3. For both the UPGMA tree and the neighbor-joining tree, find the pairwise distance matrix that agrees with the tree perfectly. Compare these with the original distance matrix. Which of the two estimates a tree where the distances are closest to the original?

   Solution: For UPGMA:

	A	B	C	D
A	0	10	6	4
B	10	0	10	10
C	6	10	0	6
D	4	10	6	0

   For Neighbor-joining:

	A	B	C	D
A	0	8.5	6	2.5
B	8.5	0	13.5	8
C	6	13.5	0	7.5
D	2.5	8	7.5	0

   All of the neighbor-joining distances are within 1.5 of the original while UPGMA are with 2.0 of the original.

Maximum Likelihood

4. Answer true or false: *There is an algorithm similar to that for parsimony which allows rapid calculation of the log-likelihood of a tree topology.*

   Solution: False: maximum likelihood estimation requires solving an optimization problem over branch lengths and other parameters.

5. Answer true or false: *For a specified likelihood model, there is an algorithm which allows rapid calculation of the log-likelihood of a tree (where both the topology and branch lengths are specified).*

   Solution: True: Felsenstein's pruning algorithm allows likelihood calculations on fixed trees to be fast.

6. Answer true or false: *All likelihood models, such as Jukes-Cantor and the General Time Reversible model lead to the same maximum likelihood tree topology, but the branch lengths may differ.*

   Solution: False: they might for some data sets, but will not in general.

7. Answer true or false: *For a specified maximum likelihood model, the maximum likelihod tree is the one that makes the probability of the observed data as high as possible.*

   Solution: True: this is the definition of maximum likelihood estimation.

8. Recall that the AIC score is $-2(\text{log-likelihood}) + 2(\text{\# of parameters})$ and that the AIC criterion is to choose the model that minimizes AIC. Which model would be prefered from this list according to AIC?

Model	$\ln L$	# of parameters
1	-3068	25
2	-2953	25
3	-2953	26
4	-2935	28
5	-2680	28
6	-2616	29

   Solution: Model 6 has an AIC score of $(-2)(-2616)+2(29) = 5290$ which is the lowest among these choices.

Bootstrap

9. Suppose a molecular alignment includes 300 sites for seven taxa. The maximum-likelihood tree from this data contains a clade with taxa 1–3. Explain how the bootstrap could be used to generate 1000 bootstrap data sets to quantify the support for this clade.

   Solution: Sample 300 sites with replacement from the 300 sites. For this new bootstrap data set, estimate the maximum likelihood tree and determine whether or not it contains clade 1–3. Repeat this process 1000 times (sampling with replacement from the original data each time). The bootstrap support for the clade is the proportion of the 1000 bootstrap trees that contain clade 1–3.

10. What is the probability that the first site is not included in the first bootstrap data set?

    Solution: $(299/300)^{300} = 0.3673$.

11. For data sets with n columns where $n = 100, 200, 300, \ldots, 1000$, find the probability that the first site is not part of the first bootstrap data set. What happens to these numbers as n increases?

    Solution: Compute $(1 - 1/n)^n$ for each n and find these probabilities: 0.3660, 0.3670, 0.3673, 0.3674, 0.3675, 0.3676, 0.3676, 0.3676, 0.3677, and 0.3677. The numbers tend to a constant (which is $\exp(-1)$.

12. For a large data set, estimate to the nearest 10% the proportion of bootstrap data sets you would expect to include any single specific site.

    Solution: Since each site will not be in about 37% of the data sets, each site will be included in about 63% (close to 60%) of the bootstrap data sets.

13. Assume a molecular data set for 4 taxa and the maximum parsimony method of tree estimation. Suppose that 5 sites favor the tree with an A/B clade, 3 sites favor the tree with an A/C clade, 2 sites favor a tree with an A/D clade, and the remaining 90 sites are not parsimony informative.

    (a) What is the probability that there are no parsimony informative sites in a bootstrap data set?
    Solution: $(0.9)^{100} = 2.7 \times 10^{-5}$.

    (b) For each bootstrap data set, how would you determine the parsimony tree?
    Solution: The parsimony tree would be the one with the largest number of informative sites that favor the particular tree topology.

    (c) What potential difficulty would need to be addressed to find the bootstrap support for the most parsimonious tree?
    Solution: There needs to be a way to handle ties.

Bayesian Inference

14. Answer true or false: *Similar to maximum parsimony and maximum likelihood, Bayesian inference requires a search through tree-space for the best tree.*

    Solution: False. Bayesian inference produces a probability distribution on possible trees, not a single best tree. This distribution is typically summarized by sampling from it using MCMC and summarizing the sample.

15. Answer true or false: *Similar to maximum parsimony and maximum likelihood, the bootstrap is often used to assess uncertainty in the Bayesian tree.*

    Solution: False. The Bayesian approach estimates trees and assesses uncertainty in a single unified approach. There is no selection of a single tree that optimizes some criterion.

16. Answer true or false: *The Bayesian approach includes both likelihood models of molecular evolution and prior probability distributions for the tree and model parameters.*

    Solution: True.

17. In an inference problem with four taxa and three possible unrooted tree topologies, assume that it is possible to calculate the likelihood of data for each tree topology by averaging over some prior distribution on branch lengths and parameters of the likelihood model. The following table shows these probabilities on the natural log scale.

Tree	$\ln L$
1	-5000
2	-5002
3	-5010

    In other words, the probability of the observed data (averaging over uncertainty in other parameters) is $\exp(-5000)$ for the first tree topology and so on.

(a) If a person places a uniform prior probability distribution on the three tree topologies (and uses the other unspecified prior distributions on branch lengths and other model parameters), what will the Bayesian posterior distribution for these three trees be? *(Hint: the problem is similar to the box and balls problem from lecture — drawing a tree diagram can help.)*

Solution: The posterior probabilities are proportional to the prior probabilities times the likelihoods; $(1/3)\exp(-5000)$, $(1/3)\exp(-5002)$, and $(1/3)\exp(-5010)$; and are equal to these divided by their sum. These numbers are all very small, so it helps to factor out a common factor, such as $(1/3)\exp(-5000)$, so that the posterior probabilities are proportional to 1, $\exp(-2)$, and $\exp(-10)$ which sum to 1.135381. The probabilities are 0.881, 0.119, and 0.000 to three digits of accuracy.

(b) If the prior distribution on tree topologies has instead been $(0.02, 0.49, 0.49)$, which tree would have had the highest posterior probability (and what would it have been)?

Solution: Solve like the previous problem, but use a different prior distribution. The posterior distribution is proportional to $(0.02)\exp(-5000)$, $(0.49)\exp(-5002)$, and $(0.49)\exp(-5010)$ and works out to 0.2317, 0.7681, and 0.0002. With strong prior probability against the first tree, the second has a larger posterior probability even though the first has higher likelihood.