Problem Set 2

1) Please indicate the parameters of an HKY-SSR$_5$ model.

$\pi_{A(1)}$ $\pi_{A(2)}$ $\pi_{A(3)}$ $\pi_{A(4)}$ $\pi_{A(5)}$
$\pi_{C(1)}$ $\pi_{C(2)}$ $\pi_{C(3)}$ $\pi_{C(4)}$ $\pi_{C(5)}$
$\pi_{G(1)}$ $\pi_{G(2)}$ $\pi_{G(3)}$ $\pi_{G(4)}$ $\pi_{G(5)}$
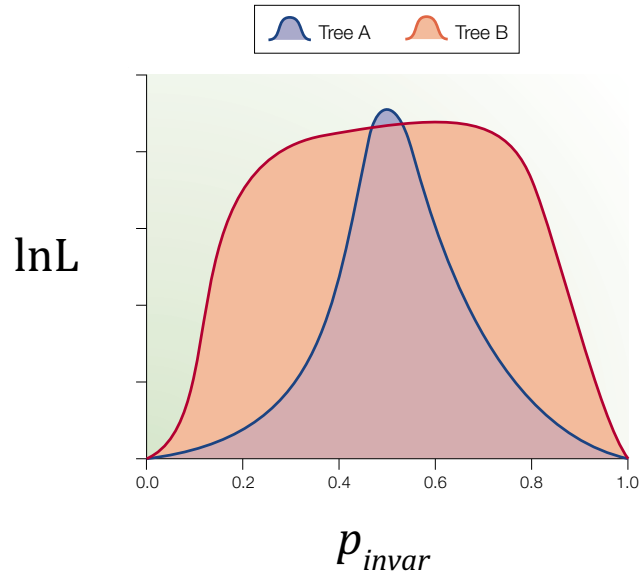$\pi_{T(1)}$ $\pi_{T(2)}$ $\pi_{T(3)}$ $\pi_{T(4)}$ $\pi_{T(5)}$
$\kappa_{(1)}$ $\kappa_{(2)}$ $\kappa_{(3)}$ $\kappa_{(4)}$ $\kappa_{(5)}$

2) Use the character-by-taxon matrix below to demonstrate why one should never map characters onto a strict consensus tree.

A	0 0 0 1 0 0
B	1 0 1 0 0 0
C	0 1 0 1 1 1
D	0 1 0 1 1 1
E	1 1 1 1 1 0

Because there are 10 steps if one characters onto the consensus tree and only 8 on either of the MP trees, branch lengths are biased upwards.

3) Please draw a likelihood surface for two trees where one has a higher joint likelihood with respect to p_{invar} and the other has a higher marginal likelihood.

lnL

p_{invar}

Tree A   Tree B

ation and its

' score, and
er of

y using
(i.e., the K-L

erges to its true

to a good estimate.

6) Why should the GTR+CAT model used in RAxML only be used for very large data
   sets?

This approach uses a parsimony tree to estimate the relative rate of evolution of each site and then lumps sites into categories based on similarity of the rate estimates. Because these are single-site estimates of rates, very many taxa are necessary to have sufficient data to derive sufficiently accurate estimates of relative rates. Use of too few taxa results in poor estimates of site-by-site rates and therefore inappropriate categorization of sites into rate classes.

7) Please write Bayes' Theorem for the posterior distribution of trees.

$$P(\tau_i \mid D) = \frac{P(\tau_i)P(D \mid \tau_i)}{\sum_{i=1}^{s} P(\tau_i)P(D \mid \tau_i)}$$