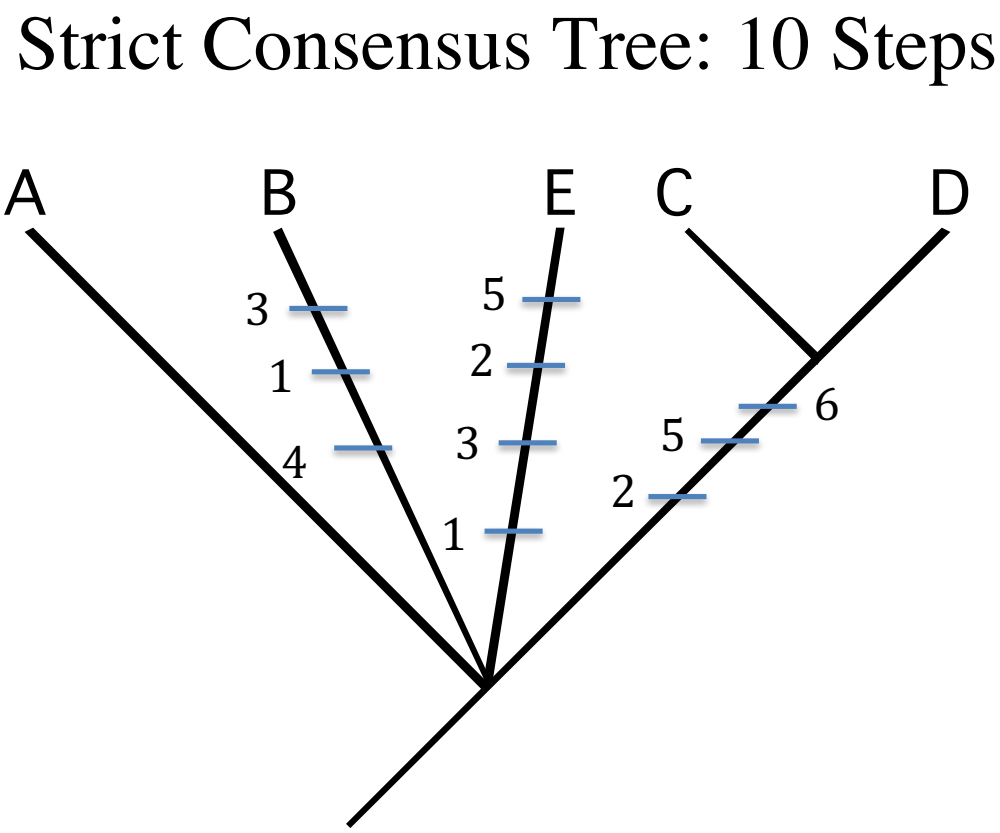
Problem Set 2

1) Please indicate the parameters of an HKY-SSR_5 model.

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

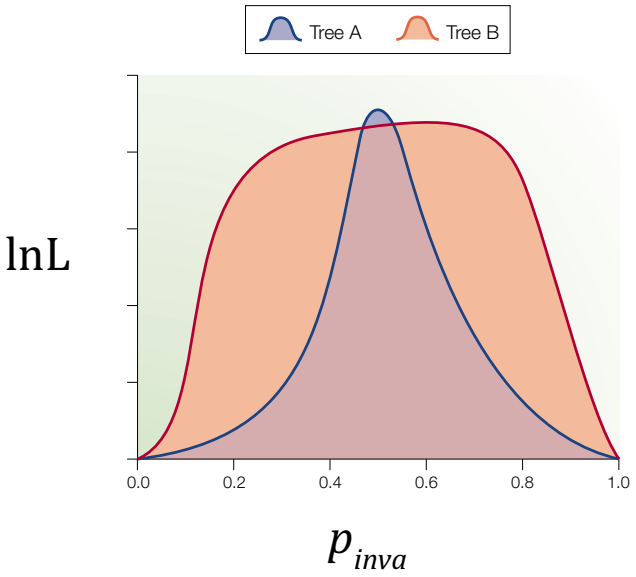
2) Use the character-by-taxon matrix below to demonstrate why one should never map characters onto a strict consensus tree.

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.



4) Please write the formula for the AIC, and provide a simple justification and its theoretical justification.

$AIC_i = -2\ln L_i + 2d_i$ - This measures the fit, via the likelihood score, and penalizes for over-parameterzation, because d_i is the number of parameters in model i.

It also measures the expected loss of information incurred by using incorrect model i *relative to using the unknown true model (i.e., the K-L distance).*

5) What it the difference between consistency and efficiency?

Consistency is an infiniate data property in which an estimate converges to its true value as the sample size increases.
Efficiency measures the amount of data required to converge to a good estimate.

6) What's the difference between a KH test that uses the normal approximation and a KH test with a RELL bootstrap?

In the original implementation, the difference in observed lnL of two trees is assessed for significance by using the distribution of SSLs to calculate the standard deviation and assumes that a t-test can detect a significantly large value.

In the RELL bootstrap approach, the null distribution is generated by resampling the differences in SSLs.

7) 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 rate.

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

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