Molecular Phylogenetics
Maximum Likelihood & Bayesian Statistics
Optimality criterion:
a rule used to decide which of two trees is best. Four optimality criteria are currently widely used:

**Maximum parsimony** – the tree requiring the fewest character state changes is the better of the two trees.

**Maximum likelihood** – the tree maximizing the likelihood under the assumed evolutionary model is better.

**Minimum evolution** – the tree having the smallest sum of branch lengths (estimated using ordinary least squares) is better.

**Least squares** – the tree showing the best fit between estimated pairwise distances and the corresponding pairwise distances obtained by summing paths through the tree is better.
Distance Methods:

*Measures of “Goodness of fit”:*

The goodness of fit $F$ between observed distances $d_{ij}$ and tree distances $p_{ij}$, for each pair of sequence $i$ and $j$, is given by:

$$F_{\alpha} = \sum_{1 \leq i \leq j \leq n} |d_{ij} - p_{ij}|^{\alpha}$$

If $\alpha = 1$, the criterion is Farris’s $f$ statistic. If $\alpha = 2$, the criterion is the least-squares-fit.

*Minimum Evolution:*

An unrooted metric tree for $n$ sequences has $(2n-3)$ branches, so that the total length of the tree, $L$, is:

$$L = \sum_{i=1}^{2n-3} e_i$$

The minimum evolution tree is the tree which minimizes $L$. 

Maximum Likelihood:

The likelihood measure the goodness of fit between a model and the data. The goal is to find the tree having the highest score, or the maximum likelihood

\[
L(M|D) = \Pr(D|M)
\]

The likelihood score is the probability of the observed data given a particular tree topology, set of branch lengths, and substitution model (it is NOT the probability that the tree is correct!, but rather that the tree has given the data we collected)

In other terms, the likelihood of the model given the data equals the probability of the data given the model

In molecular phylogenetics, D is the set of sequences and M is a tree topology
Maximum Likelihood of a single sequence:

Imaging a sequence comprising 894 sites: 272 A, 297 C, 95 G, and 230 T.

Under the Jukes and Cantor model, the Likelihood (probability of the data given the model) = $0.25^{894}$

Under any model allowing unequal base frequencies (F81, HKY, etc..),

$L =$ the Likelihood (probability of the data given the model) =

$L = \pi_A^{nA} \cdot \pi_C^{nC} \cdot \pi_G^{nG} \cdot \pi_T^{nT} = 0.30425^{272} \cdot 0.33221^{297} \cdot 0.10626^{95} \cdot 0.25727^{230}$

$\ln L = 272 \ln(0.30425) + 297 \ln(0.33221) + 95 \ln(0.10626) + 230 \ln(0.25727)$

$= -1176.17675$

$\ln L$ under the JC69 model = -1239.34715
Maximum Likelihood of a tree of 2 nodes and 1 branch:

Imaging a pair of sequences with 816 identical sites and 78 sites showing differences.

The likelihood of this tree, under the JC69 model, is given by:

\[
L = \left[ \frac{1}{16} \left( 1 + 3e^{-4\alpha} \right) \right]^{816} \left[ \frac{1}{16} \left( 1 - e^{-4\alpha} \right) \right]^{78}
\]

\[
\ln L = 816 \ln \left[ \frac{1}{16} \left( 1 + 3e^{-4\alpha} \right) \right] + 78 \ln \left[ \frac{1}{16} \left( 1 - e^{-4\alpha} \right) \right]
\]

The idea is to try plugging in values for \( \alpha \) t until you cannot find any value making the likelihood go any higher.

Each parameter in a particular model will have a likelihood curve and the values putting you at the top of the curve are the Maximum Likelihood Estimates of those parameters.
A likelihood ratio test returns a significant result if the data are better explained by the general model (for which the maximum of the likelihood function is $L_1$) than by the constrained model that represents the null hypothesis (for which the likelihood maximum is $L_0$). A ratio $L_1 / L_0$ much greater than 1 indicates significance and the likelihood ratio test statistic, defined as $LR = -2(\ln L_0 - \ln L_1)$, is large and positive.

LR is distributed as a chi-squared random variable with degrees of freedom equal to the difference in the number of free (estimated) parameters between the two models if the null model is perfectly nested within the unconstrained model.
Bayesian statistics in phylogenetics:

Bayes' Rule is based on the definition of conditional probability:

\[ \Pr(A, B) = \Pr(A) \Pr(B | A) = \Pr(B) \Pr(A | B) \]

Joint probability of A and B  \hspace{1cm} \text{conditional probability of B given A}

Posterior probability, or probability of the hypothesis (or parameter value) given the data

Likelihood, or probability of the data given the hypothesis (or parameter value)

Prior probability, or unconditional probability of the hypothesis (or parameter value) given the data

Unconditional probability of the data, which serves as a normalizing constant which ensures that the sum of the posterior probability is 1.

The posterior for an unrooted phylogenetic tree involves at least the topology and \(2N-3\) branch length parameters (\(N=\#\) of tip nodes)
Bootstrapping:

a statistical technique, first applied to phylogenetics by Felsenstein, in which new data sets are created by sampling randomly (and with replacement) from the original characters.

These new data sets (called bootstrap data sets) are of the same size as the original. A desired quantity is computed for each bootstrap data set and the resulting distribution is used to estimate the dispersion that would be expected if the same number of new independent data sets had been collected.

Bootstrapping assumes that the original characters were sampled independently.

Likelihood:

a quantity that is proportional to the probability of the data (or probability density, if the data are continuous-valued), given specific values for all parameters in the model. The likelihood function provides a means to estimate the parameters of the model.

Parameter values that are associated with the global maximum of the likelihood function are termed maximum likelihood estimates (MLEs)