Lecture Outline: Phylogeny Reconstruction using Maximum Likelihood

1. **Principle of Maximum Likelihood**

   (a) In a typical statistical model, given parameters θ, the probability of observing data X is $f(X \mid \theta)$. (Both X and θ can be multi-dimensional.)

   (b) Keeping θ fixed, and treating f as a function of X, the total probability is one.

   (c) When the roles of X and θ are reversed so that X is treated as fixed and θ is treated as unknown, the function f is called the *likelihood function*.

   (d) This is often written as $L(\theta) = f(X \mid \theta)$.

   (e) The *principle of maximum likelihood* is to estimate θ with the value $\hat{\theta}$ that maximizes $L(\theta)$.

   (f) In practice, it is common to maximize the log-likelihood, $\ell(\theta) = \ln L(\theta)$.

   (g) This is because X often takes the form of an independent sample so that

   $$L(X) = \prod_{i=1}^{n} L(X_i) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

   and the usual trick of maximizing a function by taking derivatives and setting equal to zero is easier on the log scale since

   $$\ln \left(\prod_{i=1}^{n} f(X_i \mid \theta) \right) = \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

   and derivatives can be moved inside of sums, but not products.

   (h) Also, likelihoods tend to be very small as they are the product of many probabilities, so they are easier to represent on a log scale.

2. **Simple Example**

   (a) A coin has a probability θ of being a head.

   (b) Consider tossing the coin 100 times. The probability of each single sequence with exactly x heads is $f(x \mid \theta) = p^x(1-p)^{100-x}$.

   (c) Say we observe the sequence

   $$HHTHTHHT \ldots TTH$$

   where heads appear 57 times.

   (d) The maximum likelihood estimate is the value $\hat{\theta}$ that maximizes the function

   $$L(\theta) = \theta^{57}(1-\theta)^{43},$$

   or, equivalently that maximizes

   $$\ell(\theta) = 57(\ln \theta) + 43(\ln(1-\theta)) \; .$$

   Simple calculus and common sense lead to the estimate $\hat{\theta} = 0.57$.

3. **Maximum-likelihood edge lengths**

   (a) The same principle can be used to find the maximum likelihood estimate of the distance d between two aligned DNA sequences.

   (b) For the Jukes-Cantor model, say that a pair of sequences have x sites with observed differences and $n - x$ sites with the same base.

(c) The probability of any given sequence pair is

$$L(d) = \left(\frac{1}{4}\right)^n \times \left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^x \times \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{n-x}$$

which has the form

$$L(\theta) = C \times \theta^x (1-3\theta)^{n-x}$$

where

$$\theta = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}d} \, .$$

(d) Solving the calculus problem yields $\hat{\theta} = \frac{x}{3n}$.

(e) Plugging in and solving for d gives

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{x}{n}\right)$$

(f) More complicated models have more complicated formula for estimating the distance (and other parameters).

4. **Computing Likelihood on a Tree**

(a) The likelihood of a site pattern on a tree is the sum of the probabilities over all of the possible unknown states at internal nodes in the tree.

(b) Draw a picture and do an example.

5. **Felsenstein's Pruning Algorithm**

(a) Felsenstein's pruning algorithm is a clever trick to speed the likelihood calculation.

(b) The number of terms in the sum grows like 4^n as the number of sites n increases, but the amount of necessary work only increase as a linear function of n.

(c) Demonstrate the algorithm.

6. **Tree Search**

(a) Evaluating the likelihood of a tree topology is harder than evaluating the likelihood score, as it requires estimating the edge lengths (and likelihood model parameters).

(b) In principle, each tree topology has its own maximum likelihood value.

(c) The maximum likelihood tree is that topology and estimated parameter values that has the highest maximum likelihood score.

(d) Seeking the maximum likelihood tree requires a heuristic search.

7. **Model Selection**

(a) It is possible to select a best model from a nested set of models by carrying out a series of likelihood ratio tests, or by using a criterion such as AIC or BIC.

(b) See example.

8. **The Bootstrap**

- The bootstrap was introduced to the world by Brad Efron, chair of the Department of Statistics at Stanford University, in 1979.
- The bootstrap is one of the most widely used new method in statistics that was invented within the past 50 years.
- In a special issue of *Statistical Science* that celebrates the 25th anniversary of the bootstrap, Brad Efron uses its application to phylogenetics as one of a small number of examples to illustrate its use and importance.

The general principle:

(a) We have a sample $x_1, \ldots, x_n$ drawn from a distribution F from which we wish to estimate a parameter θ using a statistic $\hat{\theta} = T(x_1, \ldots, x_n)$.

(We might think of θ as being the median of the distribution, for example, and $\hat{\theta} = T(x_1, \ldots, x_n)$ as the sample median.)

(b) If we wanted to compute the standard error of the estimate, we would ideally compute the standard deviation of $T(X_1, \ldots, X_n)$ where $X_i \sim \text{iid } F$.

(c) We could estimate this to any desired degree of accuracy by generating a large enough number (say B) of random samples $X_1, \ldots, X_n$, computing $T_i = T(X_1, \ldots, X_n)$ for the ith such sample, and then computing the standard deviation of these estimates.

$$\sqrt{\frac{\sum_{i=1}^{B}(T_i - \theta)^2}{B}}$$

(d) Unfortunately, we cannot take multiple samples from F.

(e) However, our original sample $x_1, \ldots, x_n$ is an estimate of the distribution F.

(f) Instead of taking samples from F, we could sample from the estimated distribution $\hat{F}$ by sampling from our original sample *with replacement*.

(g) We sample n values $x_1^*, \ldots, x_n^*$ with replacement from $x_1, \ldots, x_n$. It is very likely that some of the original x values will be sampled multiple times and others will not be sampled at all.

(h) For each sample, compute the estimate of θ using the original statistic. The ith estimate is $T_i^* = T(x_1^*, \ldots, x_n^*)$.

(i) Repeat this B times and compute the standard deviation of the bootstrap estimates around the estimate from the original sample.

$$\sqrt{\frac{\sum_{i=1}^{B}(T_i^* - \hat{\theta})^2}{B}}$$

(j) If the sampling distribution of the bootstrap sample estimates around the estimate $\hat{\theta}$ is similar to the sampling distribution of the estimates $\hat{\theta}$ around the true value θ, then the boostrap standard error will be a good estimate of the real standard error.

9. Applying the Bootstrap in Phylogenetics

(a) The bootstrap was introduced in phylogenetics by Joe Felsenstein in 1985.

(b) The bootstrap may be applied to any tree reconstruction method that produces a single estimated tree from a sample of data (such as maximum likelihood, neighbor-joining, UPGMA, and maximum parsimony).

(c) Assume that we have n sites in an alignment of DNA sequences.

(d) Treat the sites as an independent sample.

(e) Estimate the tree using your method of choice from the original alignment.

(f) Create B bootstrap samples, each of which consists of n sites selected uniformly at random with replacement from the original alignment of n sites.

(g) For each bootstrap data set, use the original estimation procedure to estimate the phylogeny, creating B *bootstrap trees*.

(h) For each clade in the original estimate, report the proportion of bootstrap trees that contain the clade.

(i) Alternatively, use a consensus method to summarize the sample of bootstrap trees.

10. Consensus Trees

(a) A *strict consensus tree* shows only those clades that appear in every sampled tree.

(b) A *majority rule consensus tree* shows all clades that appear in more than half the sample of trees. (Notice that two clades that each appear in more than half the sampled trees must appear in at least one tree together, implying that they are compatible with one another.)

(c) A *priority consensus tree* adds clades to the majority rule consensus tree in order of decreasing frequency in the sample provided that these clades do not conflict with a clade with higher frequency.

11. Dynamic Exploration of Tree Samples

Show off Mark Derthick's **Summary Tree Explorer**.

Software is free and available at `http://cityscape.inf.cs.cmu.edu/phylogeny/`.

12. Interpretation of Bootstrap Proportions

What does a bootstrap porportion mean? Let me count the ways.

(a) *Confidence* that the clade is in the true tree.

(b) Bayesian posterior probability that the clade is in the true tree.

(c) One minus p-value for a formal hypothesis test that the clade is in the true tree.

(d) Rough measure of method robustness.

(e) Measure of repeatability of the inferences for the method at hand.

(f) Others?