

# Lecture Outline: Phylogeny Reconstruction using Maximum Likelihood

## 1. Principle of Maximum Likelihood

- (a) In a typical statistical model, given parameters  $\theta$ , the probability of observing data  $X$  is  $f(X | \theta)$ . (Both  $X$  and  $\theta$  can be multi-dimensional.)
- (b) Keeping  $\theta$  fixed, and treating  $f$  as a function of  $X$ , the total probability is one.
- (c) When the roles of  $X$  and  $\theta$  are reversed so that  $X$  is treated as fixed and  $\theta$  is treated as unknown, the function  $f$  is called the *likelihood function*.
- (d) This is often written as  $L(\theta) = f(X | \theta)$ .
- (e) The *principle of maximum likelihood* is to estimate  $\theta$  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

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

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

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

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  $\theta$  of being a head.
- (b) Consider tossing the coin 100 times. The probability of each single sequence with exactly  $x$  heads is  $f(x | \theta) = p^x (1 - p)^{100-x}$ .
- (c) Say we observe the sequence

*HHTHTHHT...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

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

which has the form

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

where

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

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

(e) Plugging in and solving for  $d$  gives

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

(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.