5. Maximum Likelihood

EEOB563

Spring 2021

"The likelihood supplies a natural order of preferences among the possibilities under consideration." R. A. Fisher, 1956

1 Definition

Given two possible explanations for a particular outcome, it is logical to choose one that makes the observed outcome more likely ...

Or (more formally) given some data D and hypothesis H, the likelihood of the hypothesis is given by $L_H = Pr(D|H)$.

2 Simple examples

2.1 Three dice example

You have three dice: one with 6, one with 8, and one with 12 faces. A person rolls two of them and obtains the score 14. Which pair of dice is most likely to yield this result? Given the definition above, we have to calculate the probability of obtaining the score 14 for each combination of dies. We first calculate the number of combinations that give us this score for each pair of dice (1, 5, 7, respectively) and multiply by the probability of any single outcome (or divide by the total number of outcomes) (1/48, 1/72, 1/96, respectively). Our likelihoods are 0.0208, 0.0694, and 0.0729. The last is the maximum likelihood value. *Note, that unlike probabilities, likelihoods don't sum to one!*

6 face die

8 face die

12 face die

2.2 Unfair coin example (from the book)

Our experiment is the toss of a (possibly unfair) coin. Our model is simply that the toss will produce 'heads' with probability p and 'tails' with probability $1 - p$. The parameter here is p, which might have any numerical value from 0 to 1. We define the continuous function $L(p) = L(p|data) = P(data|p)$. How do we calculate the latter probability? $P(data|p) = p^m(1-p)^{n-m}$.

We want to find p corresponding to the maximum value of this function. Using product and power rules of calculus,

$$L'(p) = mp^{m-1}(1-p)^{n-m} - (n-m)p^m(1-p)^{n-m-1}$$

Equate this to zero and get $p = m/n$

3 Using likelihood in molecular phylogenetics

In the context of molecular phylogenetics D is the set of sequences being compared, H is a phylogenetic tree + the model of sequence evolution. The model itself has two parts: the composition and the process.

We want to find the likelihood of obtaining the observed data given the tree. The tree that makes the data the most probable is the Maximum Likelihood estimate of the phylogeny.

3.1 The likelihood of a sequence

What is the probability of a single nucleotide "A"? It depends on our model! If our model is $f(A) = 1/4$, the likelihood of an "A" string is 0.25. However, if our model is $f(A) = 1$, the likelihood of an "A" string is 1.

What is the likelihood of a string of N nucleotides? We assume their independence. Then

$$L = L_1 * L_2 * L_3 * ... * L_N$$

This number gets very small very quickly. Hence we convert L to lnL:

$$lnL = lnL_1 + lnL_2 + lnL_3 + ... + L_N$$

Because each $L \leq 1$, $lnL \leq 0$ and $\sum lnL$ is a negative number

3.2 The likelihood of a simple tree

For homologous sequences (sequences whose evolutionary relationship can be represented by a tree), we need to consider the process part or the probability of each nucleotide remaining the same or changing to a different nucleotide along each branch of a tree.

For example, consider a simple tree relating two sequences X and Y, X ———— Y, where X = CC Y = CT

Given what we already know, we can calculate the probability of a given distance d between X and Y under the Jukes Cantor model. We will call this probability the likelihood of d or $L(d)$

$$L(d) = P(X)P(X \to Y) = P(Y)P(Y \to X)$$

or the probability of observing one sequence times the probability of that sequence mutating into the other. We will again assume that our sites are independently and identically distributed. This allows us to break up L into the product of the likelihoods at each site:

$$L(d) = L_1(d)L_2(d) = P(C)P(C \to C)P(C)P(C \to T)$$

What is the probability of change (or no change) from one nucleotide to another? This depends on our model of evolution. Recall, we have already calculated such probabilities for the Jukes-Cantor model:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

Also recall that the evolutionary distance between two sequences according to this model $d = 3\alpha t$. Hence, we can replace $-4\alpha t$ in the equations above by $-\frac{4}{3}d$

Given a distance d we could then plug it into this equation and compute the likelihood. In maximum likelihood estimation, however, we assume d is unknown and seek the value that maximizes L(d), the maximum likelihood estimate of d. We can use calculus to maximize L(d) (take the derivative with respect to d, set it equal to 0, solve for d):

Omitting the calculations,

$$\hat{d} = -\frac{3}{4}ln(1 - \frac{4}{3} * \frac{1}{2})$$

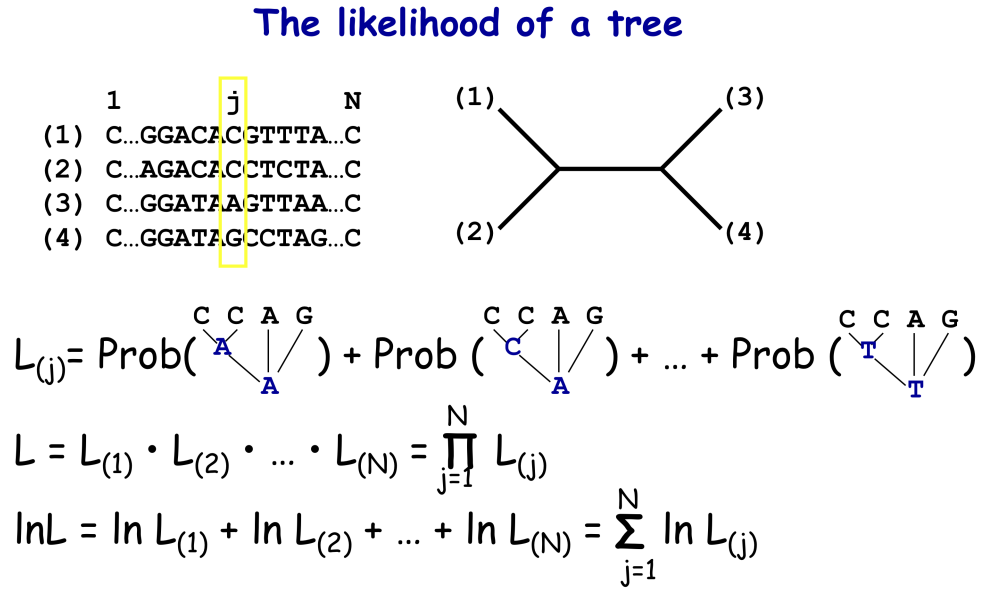
But this is just the Jukes Cantor distance betrween X and Y for $p = 0.5$, because proportion of different nucleotides is $1/2$ in our example above.

In other words, **the Jukes Cantor distance is the maximum likelihood estimate of d under the Jukes Cantor model!**

Note, that the likelihood is not the probability that the tree is the true tree! It's the probability that the tree (and other parameters of the model) gave rise to the data.

3.3 The likelihood of a more complex tree

If we have a more complex tree (like the one we saw in class), we will need to consider all possible character states at the internal nodes. Note also, that we will have to do it even for the "invariable" characters:



3.4 The likelihood calculation in real life

Efficient ML computation for phylogenetics should perform the following steps:

1. Count the number of occurrences of each pattern of bases in the aligned sequences (this will allow us to avoid repetition in calculating likelihood functions)
2. Consider all/many possible trees that might related the taxa;
3. For each such tree, construct the likelihood function of base frequencies, the relative rates and all edge lengths;
4. For each tree's likelihood function constructed in step 3 find the maximum value, and values of all the parameters that produce this maximum;
5. Choose the tree and the numerical parameters for it that had the largest maximum, and report this as ML tree.