

## Lecture 5 – Optimality Criteria: Maximum Likelihood & Minimum Evolution

### I. Maximum Likelihood – Chapter 16

**A. Introduction.** - ML estimation, in general, is a standard and incredibly useful statistical procedure. There is an excellent introduction to it in the text, on pages 248-251 in the text.

1. We're interested in calculating the Likelihood of the data given a particular hypothesis.

$$L = \Pr(\text{data} \mid \text{hypothesis})$$

We evaluate several hypotheses, and select the one that maximizes the probability of having generated the observed data.

2. Calculation of these probabilities requires an **explicit model**. This has limited its application to molecular data, although there have been recent advances in applying likelihood to morphology.

So the application of ML as an optimality criterion in phylogeny estimation is as such:

$$L_{(\tau)} = \Pr(D \mid \tau, m)$$

This is simply the probability of the data (the set of aligned sequences), given the tree (and some assumed model of character evolution).

### B. The likelihood function.

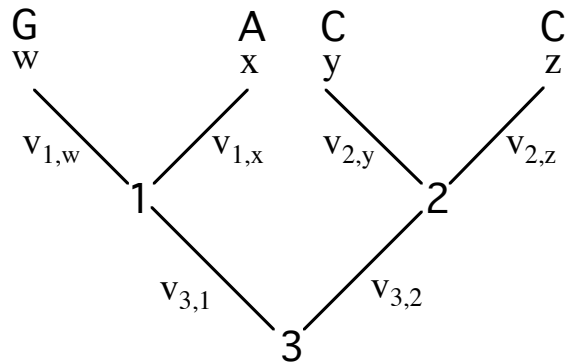
1. **Single-site likelihoods.** Just as in parsimony, we make the assumption that characters are independent so that we may treat likelihoods for each site separately:

$$L_{(\tau)} = \prod_{i=1}^s \Pr(D^i \mid \tau) = \prod_{i=1}^s L^i(\tau),$$

where  $s$  is the number of sites (characters) and  $\Pr(D^i \mid \tau)$  is the probability of site  $i$  (character  $i$ ) given tree  $\tau$ . This is the **single-site likelihood**.

So, just as the parsimony score for a tree across an entire data set is the sum of the character lengths, the likelihood of a tree across an entire data set is the product of the single-site likelihoods.

**2. SSL on a tree** - Let's take the same (arbitrarily rooted) tree we used Tuesday to demonstrate how single-site likelihoods are calculated.



Now, notice that we're paying attention to the branches, which are labeled  $v_{x,y}$ , and we're interested in their lengths. The units are expected number of substitutions per site and they are a function of rate of evolution times the duration of branch (rate x time).

So, to calculate the single-site likelihood for this site, we sum the probabilities for all possible character-state reconstructions. Since there are  $n - 1 = 3$  internal nodes (for a rooted tree) and 4 possible character states at each node, there are  $4^{n-1} = 4^3 = 64$  possible reconstructions. So:

$$L^i(\tau) = \text{Prob} \left( \begin{array}{c} G & A & C & C \\ & \swarrow & \searrow & \\ & A & A & \\ & & \swarrow & \searrow \\ & & A & \\ & & & C \end{array} \right) + \text{Prob} \left( \begin{array}{c} G & A & C & C \\ & \swarrow & \searrow & \\ & A & A & \\ & & \swarrow & \searrow \\ & & C & \end{array} \right) \\ + \text{Prob} \left( \begin{array}{c} G & A & C & C \\ & \swarrow & \searrow & \\ & A & A & \\ & & \swarrow & \searrow \\ & & G & \end{array} \right) \bullet \bullet \bullet + \text{Prob} \left( \begin{array}{c} G & A & C & C \\ & \swarrow & \searrow & \\ & T & T & \\ & & \swarrow & \searrow \\ & & T & \\ & & & C \end{array} \right)$$

Or,

$$L^i(\tau) = \sum_r^{4^{n-1}} \text{Pr}(R_r^i | \tau)$$

Of course many reconstructions are extremely unlikely, so they'll contribute very little to the single-site likelihood, but we still consider them as possibilities.

So now the issue is how one calculates the probabilities of a particular reconstruction.

Let's assume that in reconstruction  $r$ ,  $m$  is the state at node 3,  
 $k$  is the state at node 1, and  
 $l$  is the state at node 2

We know (from our data) that nodes  $w$ ,  $x$ ,  $y$ , &  $z$  have states  $A$ ,  $G$ ,  $C$  &  $C$ , respectively

$$\Pr(\mathbf{R}_r | \tau) = \pi_m \times P_{m,k}(v_{3,1}) \times P_{k,G}(v_{1,w}) \times P_{k,A}(v_{1,x}) \times P_{m,l}(v_{3,2}) \\ \times P_{l,C}(v_{2,y}) \times P_{l,C}(v_{2,z})$$

where:

$\pi_m$  is the frequency of the nucleotide  $a$  at node 3. This provides an estimate of the probability of observing state  $m$  at the root node.

$P_{i,j}$  is the probability of substitution between states  $i$  &  $j$ . This is derived from the model of sequence evolution that we assume (much more on those in a few weeks). This is in some sense analogous to the step matrix that determines costs for transformations between states in the Sankoff algorithm in parsimony.

So, this can be calculated from all  $4^{n-1}$  reconstructions and summed to calculate the single-site likelihoods.

**3. Increasing computational efficiency.** However, there are more efficient ways to calculate the same value.

**a. Felsenstein's pruning algorithm** (p. 253). We can use tree structure to atomize the computation.

$$L_{(\tau)}^i = \sum_m \pi_m \times \left( \sum_k P_{m,k}(v_{3,1}) P_{k,G}(v_{1,w}) P_{k,A}(v_{1,x}) \right) \times \left( \sum_l P_{m,l}(v_{3,2}) P_{l,C}(v_{2,y}) P_{l,C}(v_{2,z}) \right)$$

Each summation is across all four nucleotides.



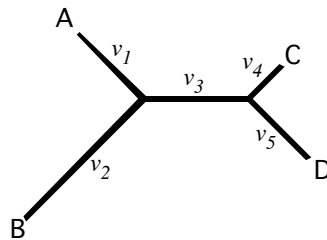
determining the best order in which to take sites, but that's more than recovered by reducing the redundancy of computation and time-savings range from 15% to 50%.

## II. Minimum Evolution (Chapter 11).

If data are either inherently distance-based, or are converted to a series of pairwise distances, one may choose to evaluate trees under the Minimum-Evolution criterion.

This matrix has  $(n^2-n)/2$  cells (pairwise comparisons).

Consider the following unrooted 4-taxon tree:



Note that, again, we're paying attention to branch lengths.

The optimality criterion is either the sum of the absolute value of the branch lengths (in the original formulation (Kidd & Sgaramella-Zonta, 1971. Amer. J. Hum. Gen. 23:235).

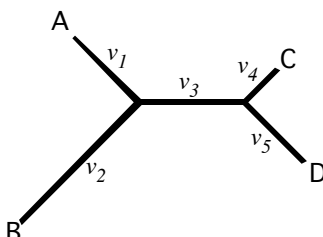
$$ME_{(\tau)} = \sum_{k=1}^{2n-3} |v_k|$$

Or, in subsequent formulations (Rzhetsky and Nei, 1992. Mol. Biol. Evol., 9:945), the sum of branch lengths.

$$ME_{(\tau)} = \sum_{k=1}^{2n-3} v_k$$

So, given a matrix of pairwise distances and a tree, how does one find the branch lengths?

Let's think for a minute about the data (pairwise distance matrix). If we can erect a distance that captures the true amount of evolution that has occurred, we will confer the property of tree additivity on our distances, such that on the tree:



$$\begin{aligned}
 d_{AB} &= p_{AB} = v_1 + v_2, \\
 d_{AC} &= p_{AC} = v_1 + v_3 + v_4, \\
 d_{AD} &= p_{AD} = v_1 + v_3 + v_5, \\
 d_{BC} &= p_{BC} = v_2 + v_3 + v_4, \\
 d_{BD} &= p_{BD} = v_2 + v_3 + v_5, \\
 d_{CD} &= p_{CD} = v_4 + v_5
 \end{aligned}$$

Where  $p_{xy}$  is the **patristic distance** between taxa  $x$  &  $y$ , or the distance calculated from tree. There are other properties of additive distances that we won't worry about (e.g., the four point metric condition).

The point is, if we could come up with error-free distance estimates, and therefore perfectly additive distances, there would be a perfect fit between the observed distance matrix and the patristic distance matrix.

So for any combination of branch lengths, we can measure the distortion associated with fitting the distance matrix to the tree:

$$E_{(\tau)} = \sum_{i=1}^n \sum_{j=1}^n w_{ij} |d_{ij} - p_{ij}|^\alpha$$

$\alpha$  is usually 2 (it may be 1 in some cases – if you suspect some distances are really poor estimates but don't know which ones), so this is usually a least-squares method.

$w_{ij}$  allows weighting of the pairwise error terms.

$w_{ij} = 1$  assumes that the errors are identical across all  $d_{ij}$ .

$w_{ij} = 1/d_{ij}$  assumes that errors are proportional to  $d_{ij}$ .

$w_{ij} = 1/d_{ij}^2$  assumes that errors are proportional to the square root of  $d_{ij}$ .

$w_{ij} = 1/\sigma_{ij}^2$  weights by the expected variance in the  $d_{ij}$  (which may not be known).

So, optimum branch lengths for a particular tree are found by the least squares criterion, but remember that the Minimum-Evolution criterion operates on the sum of branch lengths.

This use of two different optimality criteria can cause problems searching for a tree.

Another point is that the optimization of branch lengths may actually result in negative branches, which have no biological meaning.

Probably the best thing to deal with them is to set them to zero when summing branch lengths to estimate the ME score.

### **III. Relationships among optimality criteria.**

Lots of the controversy in phylogenetics revolves around choice of optimality criteria, and lots of work (including some of my own) has been focussed on performance of the various criteria.

Certainly these differences are important, but often lost in the debate are the various fundamental conceptual similarities that are common to certain sets of them.

Both MP and ML are character based (whereas ME is not).

Both ME and MP minimize the amount of evolution (i.e., sum of branch lengths).

Both ME and ML rely on an explicit model of sequence evolution.

So certainly we should expect there to be a broad range of conditions across which the methods perform similarly, and that certainly has been shown to be the case.

Nevertheless, the differences among them are real, and it's incredibly informative to study the conditions under which the methods perform differently.

We'll do that later.