CPSC 536A Course Notes for 8 Feb 2001
More Maximum Likelihood

Lecture by Anne Condon
Notes by Brian Winters

1 The Model

The underlying model $\mathcal{M}$ (e.g. Jukes-Cantor) gives the expected number of nucleotide changes in a single evolutionary step. The model must assume that each site evolves separately. Specifically, the model must define:

- $P(y|x,t)$ – the probability nucleotide x changes to y in evolutionary time t. (Note that the only nucleotides upon which it is dependent are x and y.)
- Initial equilibrium distance of bases at root.

2 The Likelihood Tree

Given a tree $\mathcal{T}$ (Figure 1), determine its likelihood. Each leaf of the tree has a sequence associated with it. Assume the sequence is of length one to start, for simplicity (Equation 1). We wish to find:

$$\begin{aligned} \mathcal{L}(\mathcal{T}) &= Pr(\mathcal{D}|\mathcal{T},\mathcal{M}) && (1) \\ &= \prod_{i=1}^{n} Pr(\mathcal{D}_i|\mathcal{M},\mathcal{T}) && (2) \end{aligned}$$

Where $\mathcal{L}$ is likelihood and $\mathcal{D}$ is the sequence data. (For b's of length greater than one, $\mathcal{L}$ is the product of each independent site, as shown in Equation 2.)

For our example (Figure 1):

$$\mathcal{L}(\mathcal{T}) = \sum_{b_r \in \{A,C,G,T\}} \sum_{b_v \in \{A,C,G,T\}} P(r=b_r)P(C|b_r,t_{rs})P(b_v|b_r,t_{rv})P(A|b_v,t_{vu})P(C|b_v,t_{vw}) \quad (3)$$

We sum over all possible labels of the internal nodes. Note that b_s, b_u, and b_w are fixed because they are leaves.

The goal of maximum likelihood tree reconstruction is to find the tree (topology and evolutionary distances) that has the highest likelihood.

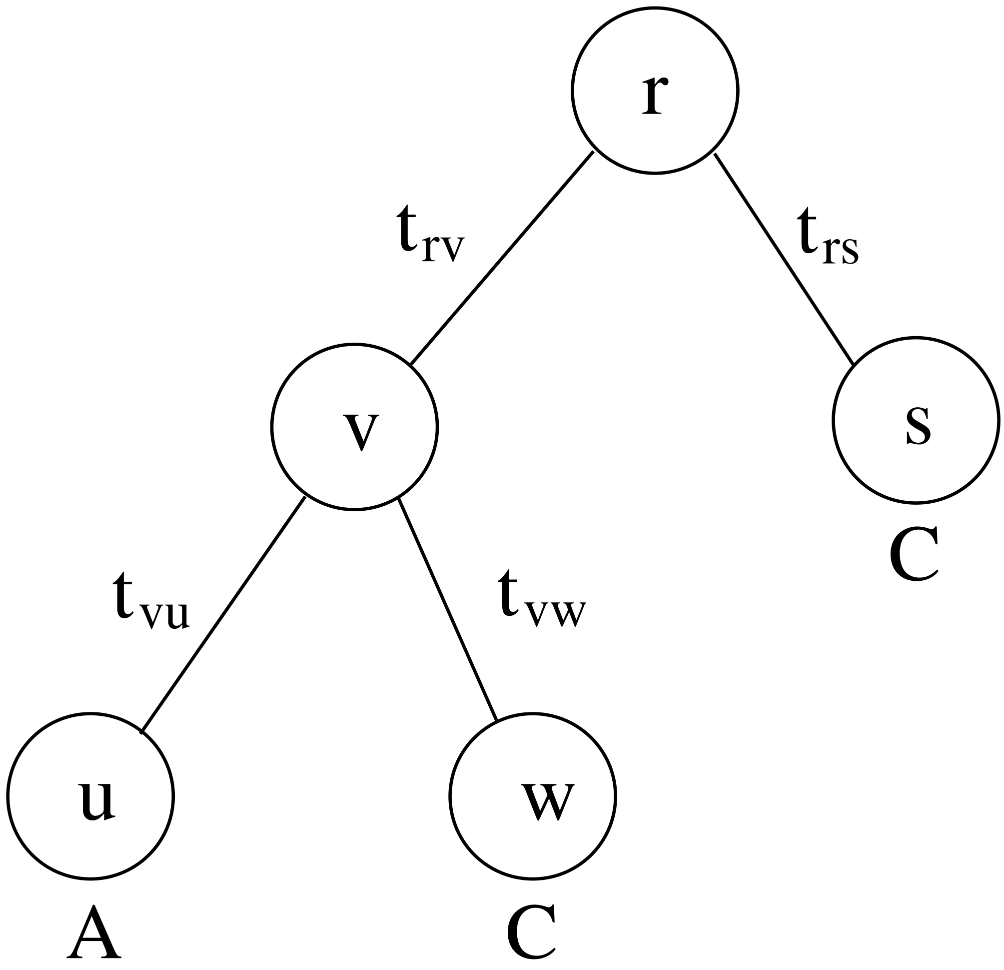

Figure 1: One possible tree for $\mathcal{D} = \{A, C, C\}$, where $b_s = C$, $b_u = A$, and $b_w = C$. Edge weights (e.g. t_{vu}) denote evolutionary distance.

3 Discussion of the material so far

Several questions were asked (and answered) at this point.

- From last lecture, the notation $P(\ldots|\ldots)$ is a Poisson distribution.
- The evolutionary distances, e.g. t_{rv}, are already fixed when we get to this point.
- *What if the internal node (e.g. v) is the same as an adjacent leaf (e.g. $w = C$)? Isn't the associated $t = 0$?* No, because "no change" is a valid evolutionary step after time t.

4 Overview of Maximum Likelihood

Likelihood of a tree somehow measures the suitability of that tree as a solution. What we would like to be able to do (subtasks):

A. Given a topology $\mathcal{T}$, evolutionary distances t, and data $\mathcal{D}$, calculate likelihood $\mathcal{L}$.

B. Given $\mathcal{T}$ and $\mathcal{D}$, find the evolutionary distances that maximize $\mathcal{L}$.

C. Given $\mathcal{D}$, find a topology and distances that maximize $\mathcal{L}$.

Subtask A has an efficient solution. The definition of likelihood allows for "optimal subproblem" style solutions. There are no efficient algorithms for Subtasks B or C, but there are effective solutions for Subtask B. Subtask C is hard.

5 Calculating $\mathcal{L}$

Let's look at our previous solution to what we've now defined as Subtask A. Rearrange Equation 3 to sum over the possibilities at the root of the tree, and then have two components representing the left and right subtrees:

$$\sum_{b_r \in \{A,C,G,T\}} P(r = b_r)P(C|b_r,t_{rs}) \sum_{b_v \in \{A,C,G,T\}} P(b_v|b_r,t_{rv})P(A|b_v,t_{uv})P(C|b_v,t_{vw}) \tag{4}$$

$P(C|b_r,t_{rs})$ covers the root. $P(b_v|b_r,t_{rv})$ is the probability that v is labeled b_v given that root r is labeled b_r. $P(A|b_v,t_{uv})P(C|b_v,t_{vw})$ is the likelihood of the tree rooted at v, given that b_v labels the root.

Selecting one part of Equation 4, the subproblem $P(C|b_r,t_{rs})$ is equivalent to:

$$\sum_{b_s \in \{A,C,G,T\}} P(b_s|b_r,t_{rs})\mathcal{L}(\mathcal{T}_s|b_s)$$

Where the likelihood of the tree rooted at s given that b_s labels the root is:

$$\mathcal{L}(\mathcal{T}_s|b_s) = \begin{cases} 1 & \text{if } b_s = C \\ 0 & \text{otherwise} \end{cases}$$

More generally, let u be an internal node with children v and w, and x be an external node (leaf) with label l. Given a tree $\mathcal{T}_r$ rooted at r:

$$\begin{aligned}
\mathcal{L}(\mathcal{T}_r) &= \sum_{b_r \in \{A,C,G,T\}} P(r = b_r)\mathcal{L}(\mathcal{T}_r|b_r) && (5) \\
\mathcal{L}(\mathcal{T}_u|b_u) &= \left(\sum_{b_v \in \{A,C,G,T\}} P(b_v|b_u, t_{uv})\mathcal{L}(\mathcal{T}_v, b_v) \right) \left(\sum_{b_w \in \{A,C,G,T\}} P(b_w|b_u, t_{uw})\mathcal{L}(\mathcal{T}_w, b_w) \right) && (6) \\
\mathcal{L}(\mathcal{T}_x|b_x) &= \begin{cases} 1 & \text{if } b_x = l \\ 0 & \text{otherwise} \end{cases} && (7)
\end{aligned}$$

Evaluation of $\mathcal{L}$ begins with Equation 5. Equation 6 is the recursive case and Equation 7 is the base case.

6 Transparencies

1. Tools they had to tackle the problem came from paleontology, phenotypic data, etc.

   > The limbs divided into great branches...were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species... *Darwin, 1859*

   The problem with these tools is that you can only get so far... until molecular biology.

2. Note that the prokaryotes are not all on one branch, with Archaea more closely related to Eukarya than Bacteria. This tree was constructed based on genes encoding for rRNA.

4. On SSU rRNA:

   > It is abundant... it has slow- and fast-evolving portions ("hour hands and minute hands"), and it has a universally conserved structure. Two other factors... are its obviously ancient an dessential fundamental function... and its interaction with many other coevolved cellular RNAs and proteins... least liable of all genes to experience interspecific lateral gene transfer. *Doolittle, 1999*

   Used to produce an unrooted tree instead of a rooted tree. SSU (small sub-unit) rRNA is universal to all life.

5. 16S, which is commonly used in phylogenetic studies, in part because it leads to a rooted tree.

6. Some parts of the molecule are universal. Notice that the similarities of Archaea and Eukarya really start to show up.

   Just rRNA sequence studies on their own are not sufficient to root the tree. The tree is rooted by looking at genes which encode for "elongation factors," which play a roll in replication (they extend the growing replica by one base).

   Study looked at two different kinds which are more or less common to everything.

   The process produces an unrooted tree, but you can infer the root based on where the three main trees join.

7. Scientific controversy: As more and more phylogenetic studies are done, the data is conflicting. High up the trees are similar, but deep down in the tress there is lots of divergence. This suggests that lateral transfer is important.

8. An example of a tree produced with neighbor joining (NJ). This work tested the molecular clock hypothesis — the hypothesis is generally supported, so they decided to use NJ for their tree. In this example they used a chimp to help root the tree.

9. This slide contains a table of some significant sequences used to produce the tree in the previous slide.

10. Similar work looking at the evolution of mammals. Prior to phylogenetics, evolutionary trees for mammals were constructed primarily based on teeth.