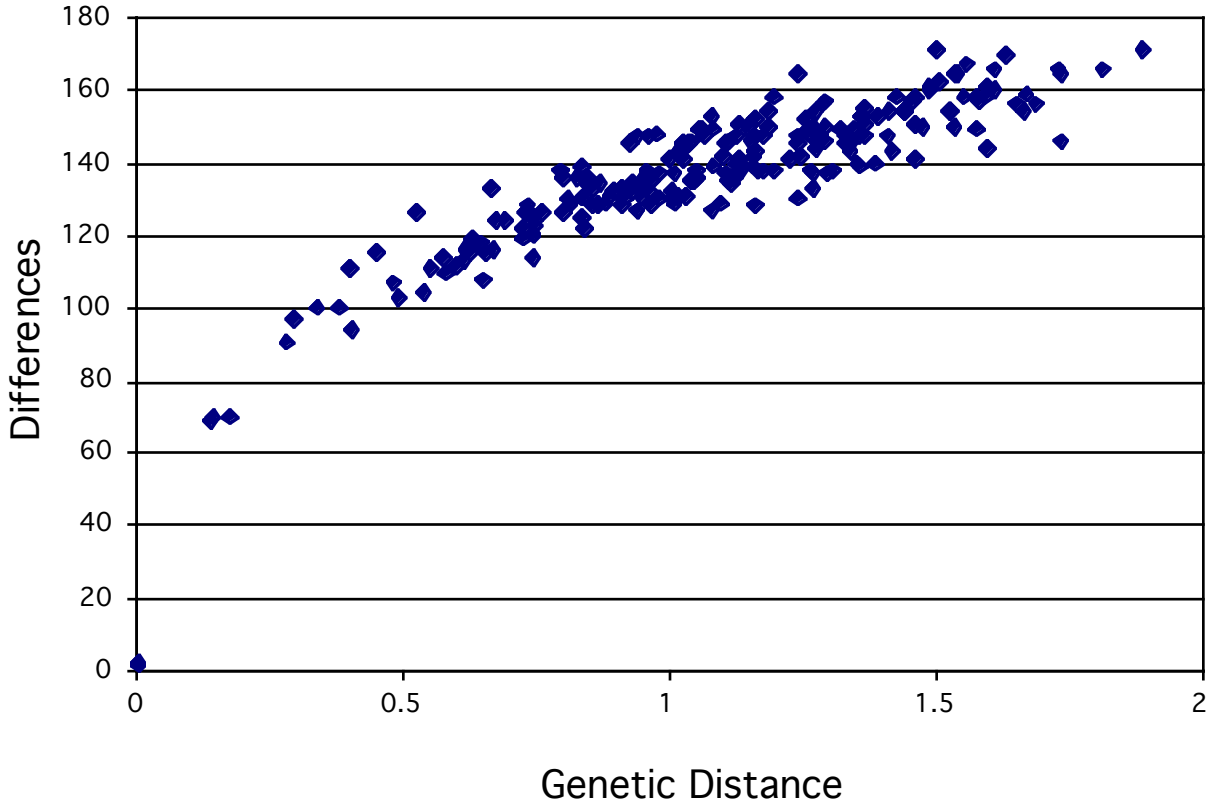
Lecture 9 – The Importance of Models

I. Introduction: As we've discussed, for long, there was a division among systematists regarding the relevance of models to phylogeny estimation. It's clear that model-based approaches are now dominant and it's worth examining why that is the case. The next several lecture topics focus on models, so let's first address why it is that they're critical.

Let's look at a saturation plot (a classical tool in molecular phylogeny and evolution).



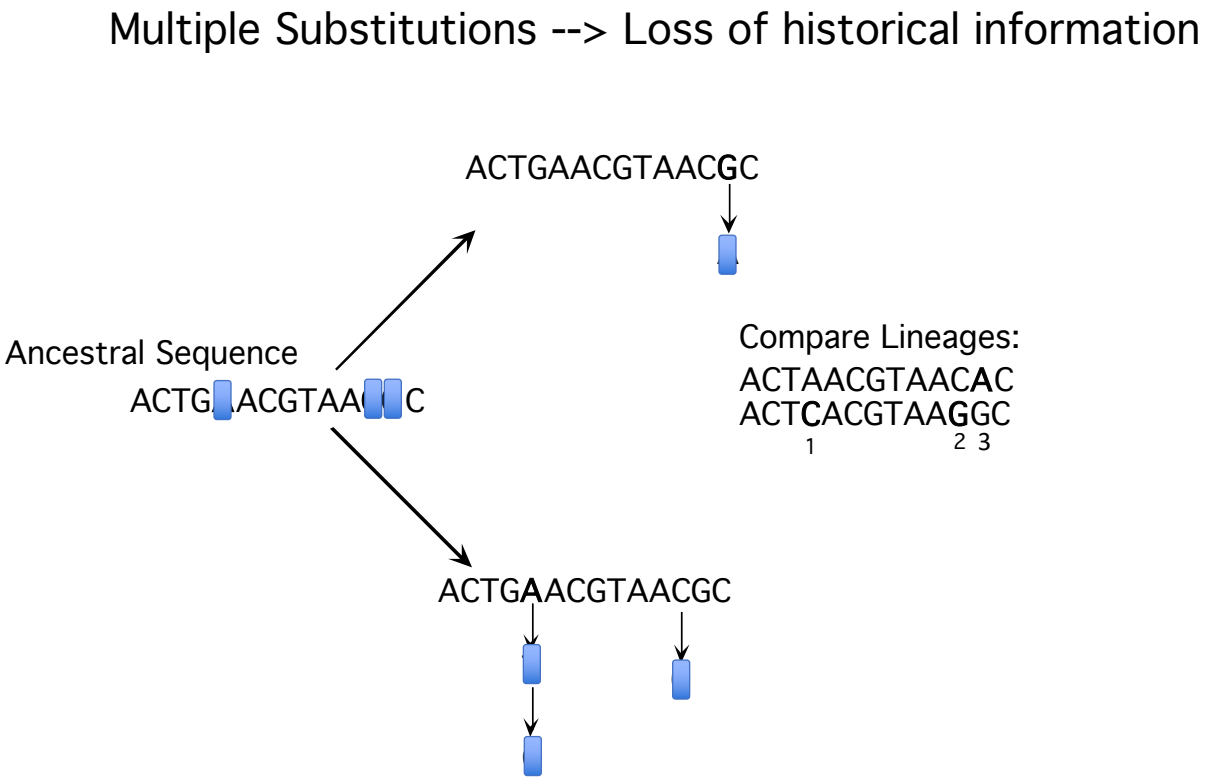
Here, I've plotted the absolute number of differences between each pair of taxa versus the genetic distance between those two taxa. The x-axis is a proxy for time since divergence between the two taxa.

So, observable differences accumulate linearly with time for only a very short time after two taxa diverge, such that two pairs taxa that have different divergence times, may have a similar number of differences between them.

This is an incredibly well-known phenomenon, called saturation.

It's also very well known that the cause of saturation is multiple substitutions at a site, or multiple hits.

II. Demonstration of multiple hits.



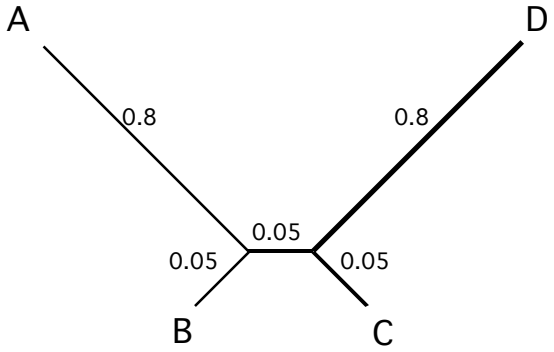
So even though there have been 4 substitutions, when we compare these two lineages, we only can detect 3 differences.

By adopting probabilistic models of sequence evolution, we can build an expectation for multiple hits into our phylogenetic estimates.

We can also see that incorporating branch-length information into our tree evaluation might be a good idea, because the **probability of multiple hits increases as a lineage persists longer**.

III. Example of the impact of multiple hits; long branches have a high probability of multiple hits.

Let's assume that the true tree for taxa A, B, C, & D is this:

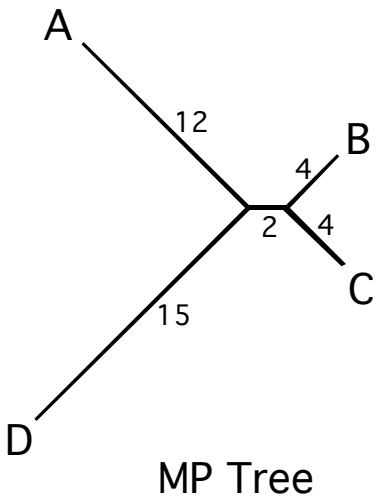


Furthermore, let's simulate nucleotide sequence data for these four taxa. We know the tree above is the true tree for these data because I used it to simulate them. Furthermore, we know that the data were generated using a model called Jukes-Cantor (which we'll learn later – it's a simple model that expects multiple hits).

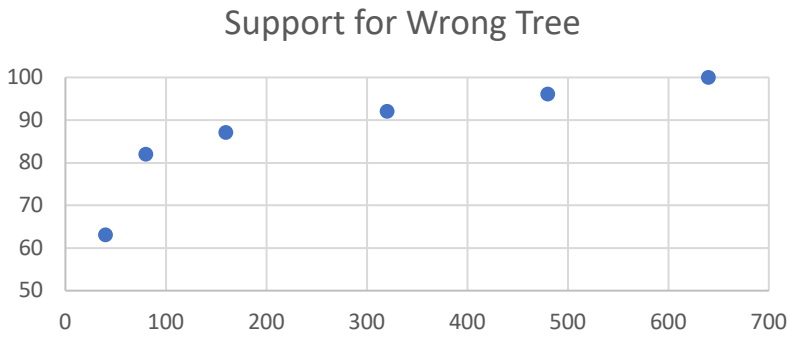
We have the following data, with 40 nucleotides and the four taxa in the tree.

```
A          ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
B          ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
C          ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
D          AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC
```

If we subject these sequences to Maximum Parsimony Analyses, we get the following 37-step tree (the true tree has 38 steps):



MP Tree



If we increase sequence length, this happens with certainty.

If we do multiple simulation replicates, this keeps happening.

Now let's subject the sequences to an ME search.

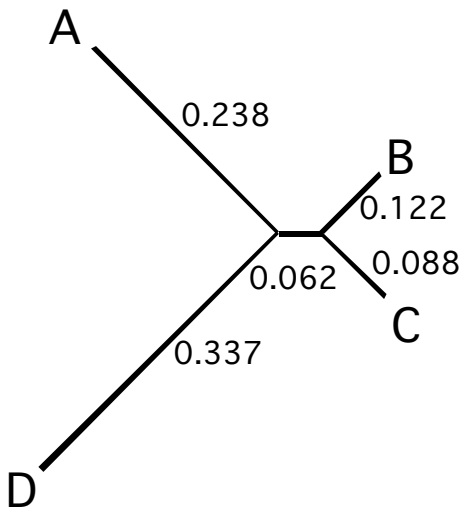
First, we need to convert the character by taxon matrix to a matrix of pairwise distances:

	A	B	C	D
A	-------	0.400	0.400	0.575
B	0.572	------	0.200	0.525
C	0.572	0.232	-------	0.475
D	1.091	0.903	0.752	-------

Above the diagonal are simply percent sequence divergences (*p*-distances). This is a distance that has not been corrected for multiple hits.

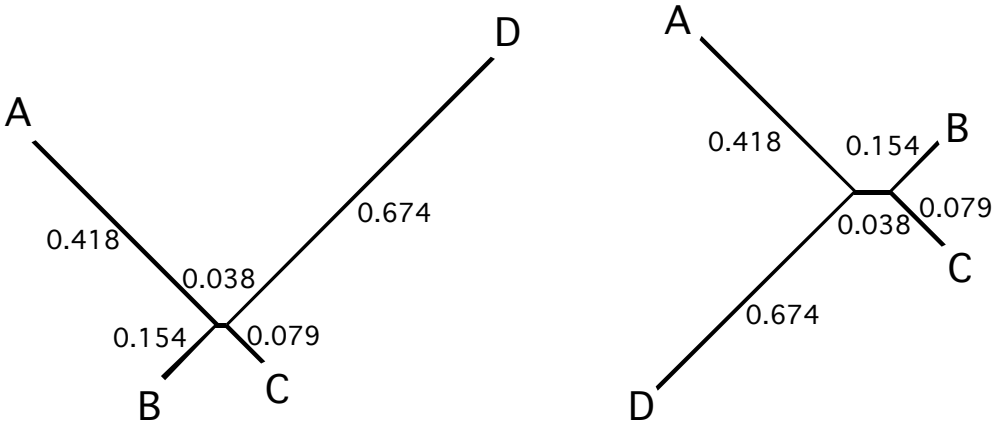
Below the diagonal are the corrected distances, corrected using the JC model.

If we find the ME tree using *p*-distances (i.e., ignore the possibility of multiple hits):



ME Tree: p-distance

However, if we do the same ME search on distances corrected using the JC model (i.e., the true model), we get the following two trees:

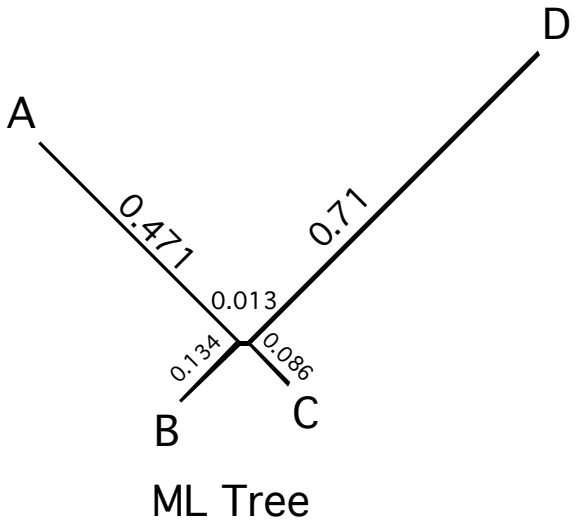


ME Trees: JC distance

These trees have the same sum of branch lengths (1.363 substitutions per site). One is the true tree and one is the same tree as before. At least we're getting better.

ME under the true model is consistent.

Now, if we find the ML tree, using the JC model (it's the simplest we can use), we get the true tree.



ML Tree

Notice that, although the ML tree is the true tree, we're still not getting accurate branch lengths. This is due to the fact that, with only 40 bases, we don't have enough data.

So, in the simulation, the methods that are agnostic with respect to multiple hits (MP and ME using *p*-distances) incorrectly unite the long-branch taxa (A & D).

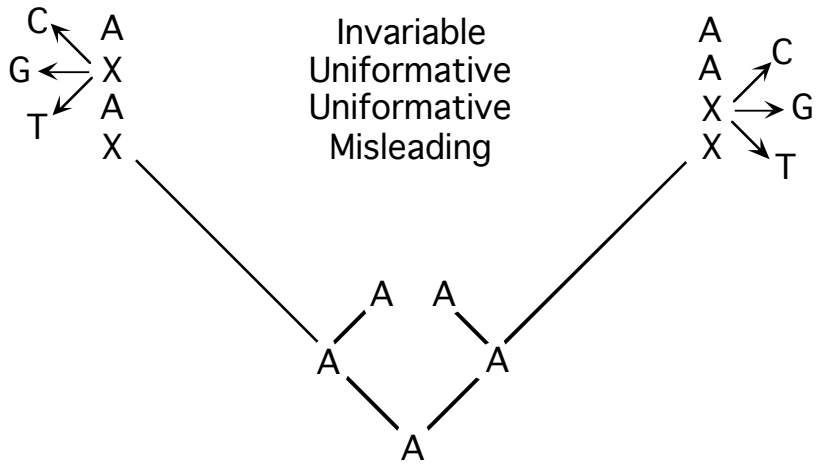
This is called **long-branch attraction** (by Felsenstein in 1978) and results from ignoring the possibility of multiple hits (or under estimating them).

ME on perfectly corrected distances finds two equally optimal trees, one of which is correct.

Under ML and assuming the true model, the optimal tree is the true tree.

IV. **Long-branch attraction.**

See Swofford et al. (1996: Pp. 385-514 in Hillis, Moritz & Mable[eds.] *Molecular Systematics*, 2nd edition, Sinauer).



Let's assume that there is an A in both the short-branch taxa. Given the shortness of the branches, it's almost certain that all the internal nodes have an A (i.e., the reconstruction shown will contribute the lion's share of the single-site likelihood). There are four possibilities.

1) The long-branch taxa could have A's, in which the site is invariant and can't help us discriminate among trees.

2 & 3) Similarly, if only one of the long branch taxa has a substitution to nucleotide X (= G, C, or T), the site won't help us choose a topology.

4) If both long-branch taxa experience a substitution, and X_1 does not equal X_2, parsimony will not favor the true tree, but won't favor a wrong tree either. The only scenario in which parsimony will favor a single tree is if $X_1 = X_2$. In this case it will favor an incorrect tree!

The reason that model-based methods avoid this bias is that they consider branch lengths when evaluating a tree.

Look back at the simulated data

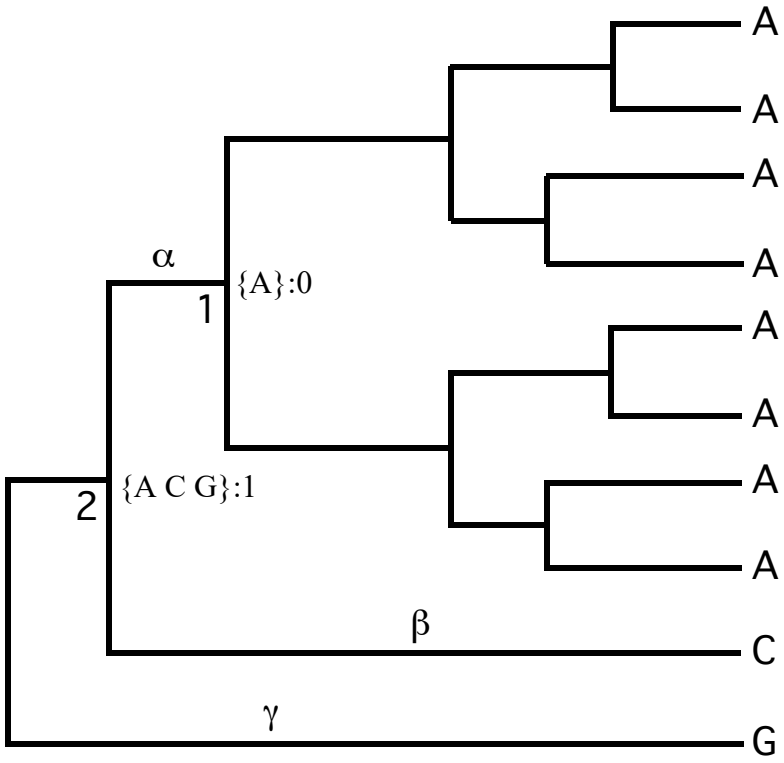
```
A    ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
B    ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
C    ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
D    AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC
```

The 2 sites that distinguish among topologies under parsimony are misleading and favor the LBA tree.

Remember that these data were simulated on the known tree.

IV. The importance of branch lengths.

Let's assume the following tree and distribution of character states for a particular site.



If you were to do a parsimony (Fitch) optimization to assign character states to internal nodes 1 and 2, you would have an A at the node labeled 1 (in the unrooted case).

Furthermore, a parsimony optimization at node 2 would infer either an A, C, or G; all are equally parsimonious.

Let's look at the implications of each of these three equally parsimonious reconstructions qualitatively under likelihood.

First, **under likelihood**, we would infer from the large number of A's that this is not a very rapidly evolving site.

If there's a C at node 2, there would have to be a substitution to A along the short branch α, no change along the long branch β and a change (to a G) along the long branch γ.

If there's a G at node 2, there would also have to be a substitution to A along the short branch α, a change to C along the long branch β and no change along the long branch γ.

Now, if there's an A at node 2, there would be no change along the short branch α, and one change along each of the long branches β & γ. Remember that branch lengths are expressed as expected number of substitutions per site, so the reconstruction that requires no change on a short branch & changes on long branches has a higher probability than those that require a change on the short branch and stasis on either of the long branches.

Now remember that each of these possibilities is considered when calculating the single-site likelihood for this character, but that SSL is going to be dominated by the last of these three reconstructions.

So, while parsimony voices no preference for any of these three reconstructions, likelihood prefers the reconstruction of an A at node 2, based on the inference that this is a low rate site and the length of branch α.

This illustrates a fundamental difference between model-based methods and parsimony. In incorporating process into our optimality criteria, we can avoid some of the biases that can mislead phylogenetic estimation.

Models that expect multiple hits to occur can be really important.

We'll spend the next couple lectures describing commonly used models.