

Introduction to Phylogenies: Likelihood methods

- Principle of maximum likelihood
- Computing likelihoods on trees
- Rate variation among sites

One minute responses on phylogenetics

- I enjoyed the phylogenies and explanation of distance methods.
- I was happy to finally find out why everyone in systematics seems to use neighbor joining instead of more accurate methods.
- Phylogeny was interesting and easy to understand.
- I am confused about the phylogeny portion still, but suspect I'll be OK after looking over more info. Could you recommend a tree building "primer"? I recall you mentioning a book but also that it was more advanced. *I recommended a paper: Felsenstein, J (1988) Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22: 521-565.*

One minute responses on phylogenetics

- I'm not entirely sure what the take-home message of the phylogeny lecture is. What will we need to be able to do? Judge the relative merits of the various methods? Implement each method? I'm still confused about what the point of this is.
- I see the goals as:
 - Be able to interpret a phylogeny (rooted or unrooted)
 - Understand the general concept of each method
 - Be able to carry out hand calculations for simple parsimony and UPGMA cases
 - Have a general idea of the strengths and weaknesses of each method
 - Recognize problem cases where phylogeny inference will probably fail

The idea of maximum likelihood

- I roll a die and it comes up 6 three times in a row
- What is the chance that it's a fair die?
- Impossible to tell unless we know something about the set of possible dice
- To calculate $P(\textit{hypothesis}|\textit{roll})$ we need to know about all possible hypotheses
- Sometimes we don't know that

The idea of maximum likelihood

- Instead, we could calculate $P(\text{roll}|\text{hypothesis})$
- If the die is fair, the chance of this outcome is

$$\left(\frac{1}{6}\right)^3 = 0.00463$$

- Under the theory that the die only has 6's, it would be 1.0
- We could then say that the data supports the second hypothesis much more strongly
- Without knowing whether there are any dice like that around, this is the best we can do

Application to trees

- We would like to know $P(\text{tree}|\text{data})$
- This would require considering all possible trees, which is unfeasible
- Instead, we will calculate $P(\text{data}|\text{tree})$ and prefer the tree for which it's highest
- This requires us to consider all possible data sets (of this size) but that's relatively easy
- Principle of Maximum Likelihood: choose the tree which makes the data most probable

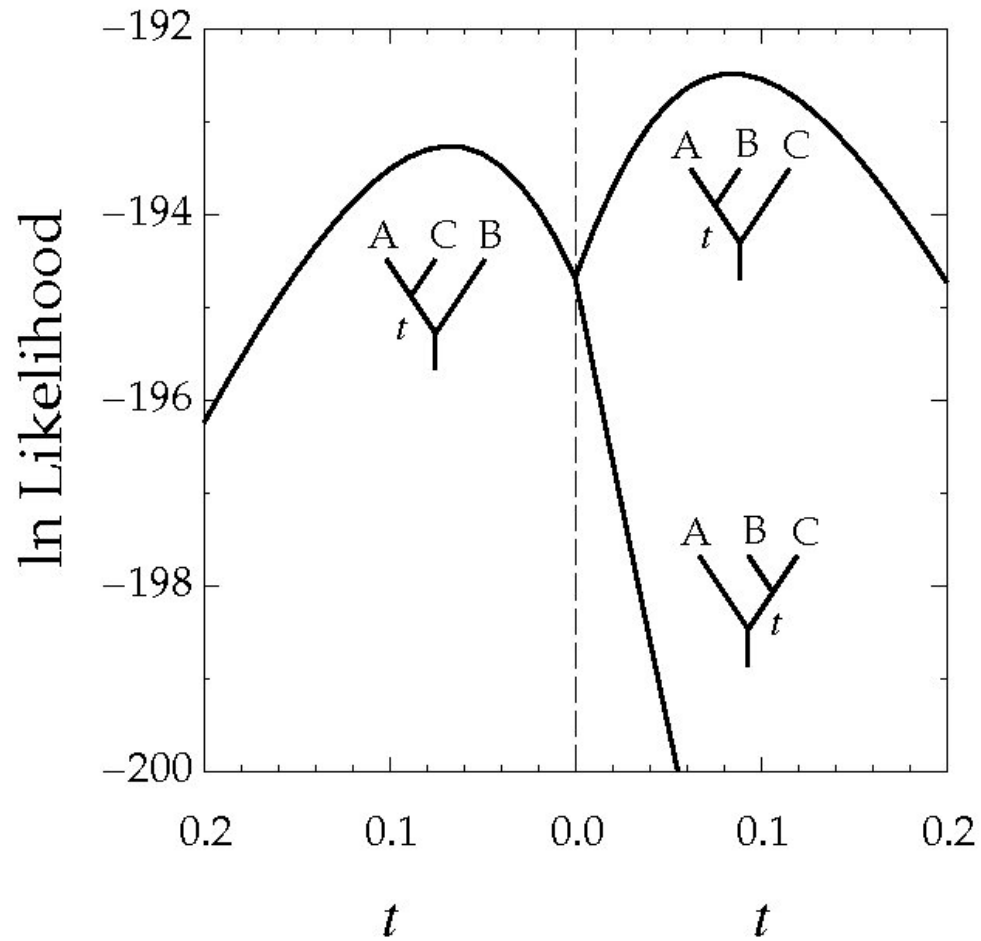
How to compute $P(data|tree)$

- Use a mutational model, just as with distances
- Start working down from tips of tree
- At each point, compute probability of data given the tree so far
- At the bottom you have $P(data|tree)$

How to compute $P(data|tree)$

- This algorithm is called “pruning” and is due to Felsenstein
- It is closely related to dynamic programming
- Note that it only gives us $P(data|tree)$ for a specified tree with specified branch lengths
- Tree search is still a problem

Shape of the likelihood function



Example

		A	C	G	T
Mutation probabilities for a branch of length 1	A	0.7	0.1	0.1	0.1
	C	0.1	0.7	0.1	0.1
	G	0.1	0.1	0.7	0.1
	T	0.1	0.1	0.1	0.7

Example

		A	C	G	T
Mutation probabilities for a branch of length 1	A	0.7	0.1	0.1	0.1
	C	0.1	0.7	0.1	0.1
	G	0.1	0.1	0.7	0.1
	T	0.1	0.1	0.1	0.7

Data: Human had A, chimp had G

Tree hypothesis: Each was 1 unit from the common ancestor

Matrix represents "probability of tree above this point, given that the ancestor had this particular base"

A	C	G	T
0.07	0.01	0.07	0.01

Example

A	C	G	T
0.07	0.01	0.07	0.01

$$L = mA * pA + mC * pC + mG * pG + mT * pT$$

where pA is the base frequency of A and mA is the number from the matrix above.

In this case if all bases equally frequent, $L=0.04$.

We could change the branch length to try to find a better likelihood.

Example

- This example did only one base pair
- In such cases, the branch length will maximize to either zero (if the bases are identical) or infinity (if they are not identical)
- With multiple base pairs, we multiply the likelihoods together
- This will give a more reasonable estimate of branch length!

Interpreting likelihoods

- Likelihood is $P(\text{data}|\text{hypothesis})$
- Can be compared among hypotheses
- Can NOT be compared among data sets
- If a data set has lots of information, its likelihood will be low for ANY hypothesis
- (What is the chance you were just dealt that exact card? Those exact 13 cards?)

Interpreting likelihoods

- As likelihoods are usually tiny, we generally report $\ln(L)$, the log likelihood
- This is a negative number, made best by making it closer to zero
- The appropriate comparison among trees is the difference in $\ln(L)$
- $\ln(L)$ differences become significant at roughly 2

Features of this approach

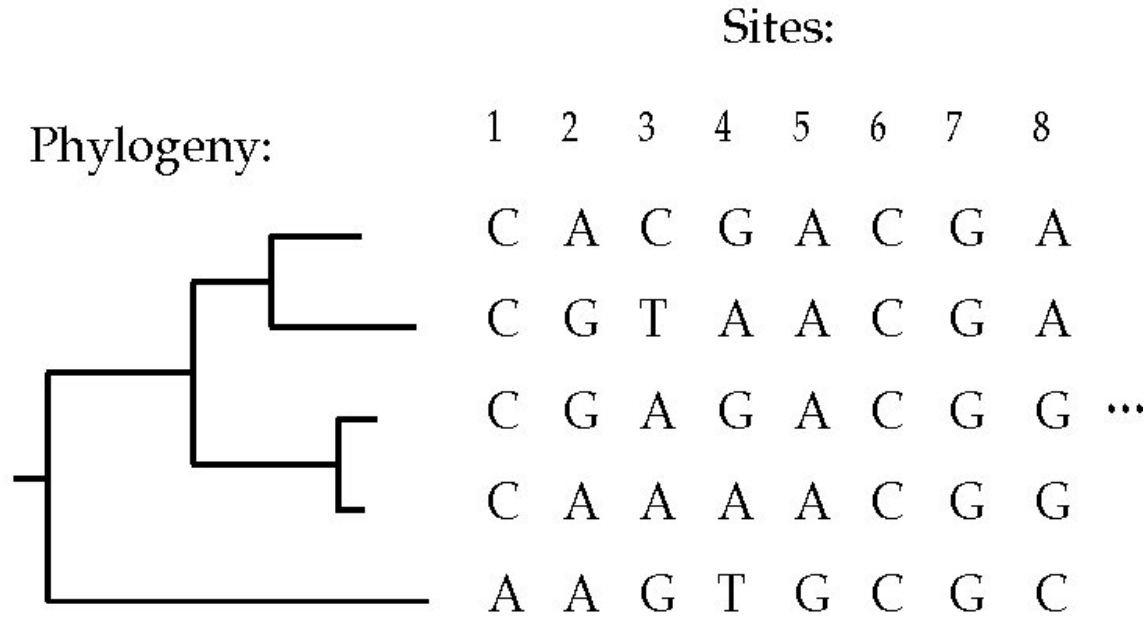
- Advantages:

- Maximum use of information in data
- Can use any available mutational model
- Powerful, robust, and consistent (if model is correct)
- Can tell us not only which tree to prefer, but by how much

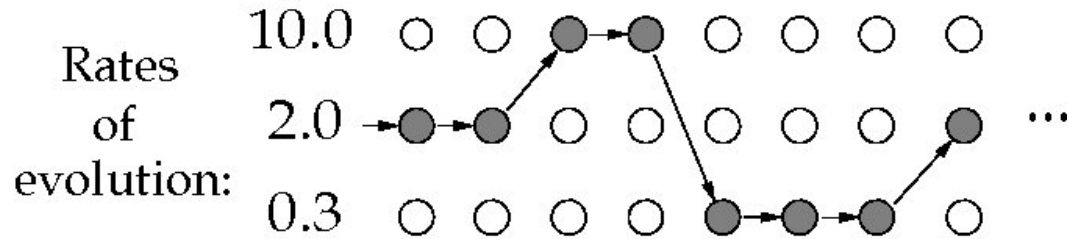
Disadvantages:

- Possible wrong answers if model is wrong
- Very, very slow
- User may be tempted to skimp on tree search to save time
- Not intuitive for many biologists

Rate variation among sites



Hidden Markov process:



Rate variation among sites

- Likelihood approach to rate variation sums over all possible combinations of rates
- Can allow correlation among rates at adjacent sites
- Optimizing the number of categories is difficult
- Slow algorithm becomes even slower!
- For HIV data this is essential

Likelihood ratio test

- Likelihood methods offer statistical tests of some questions:
 - Clock versus no clock
 - Rate variation versus rate constancy
- The two hypotheses must be nested (one is a special case of the other)
- LRT is distributed approximately as χ^2
- Unfortunately, different trees are not nested hypotheses
- Also, this test is only asymptotically correct (infinite data)