Genome 559

Instructor: Dr. Mary Kuhner (through 2/26)

Office Hours: Tuesday 3-3:30 (in computer lab) or by appointment

Phone: (206) 543-8751

Web page:
Introduction to Phylogenies: Likelihood methods

- Principle of maximum likelihood
- Computing likelihoods on trees
- Rate variation among sites
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 $P(\text{hypothesis}|\text{roll})$ 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 $P(tree|data)$
- This would require considering all possible trees, which is unfeasible
- Instead, we will calculate $P(data|tree)$ and prefer the tree for which it’s highest
- 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, you’ve computed 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

Mutation probabilities for a branch of length 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Mutation probabilities for a branch of length 1

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"

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.07</td>
<td>0.01</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Example

\[
\begin{array}{cccc}
A & C & G & T \\
0.07 & 0.01 & 0.07 & 0.01 \\
\end{array}
\]

\[L = mA \times pA + mC \times pC + mG \times pG + mT \times pT\]

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.
Interpreting likelihoods

- Likelihood is $P(data|hypothesis)$
- 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

Phylogeny:

Sites:

1 2 3 4 5 6 7 8

C A C G A C G A
C G T A A C G A
C G A G A C G G ...
C A A A A A C G G
A A G T G C G C

Hidden Markov process:

Rates of evolution:

10.0 2.0 0.3
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 $\chi^2$

• Unfortunately, different trees are not nested hypotheses

• Also, this test is only asymptotically correct (infinite data)