Genetics/MBT 541

Spring, 2001

Lecture 6

Joe Felsenstein
Department of Genetics
joe@genetics

Phylogeny methods, part 6
(Testing trees)
Likelihood ratio statistic used to place confidence limits on transition-transversion ratio for the 14-species primate data set
Likelihoods in tree space – a 3-species clock example
The constraints for a molecular clock

To test for a molecular clock:

- Obtain the likelihood with no constraint of a molecular clock (For primates data with $T_s/T_n = 30$ we get $\ln L_1 = -2616.86$
- Obtain the highest likelihood for a tree which is constrained to have a molecular clock: $\ln L_0 = -2679.0$
- Look up $2(\ln L_1 - \ln L_0) = 2 \times 62.14 = 124.28$ on a $\chi^2$ distribution with $n - 2 = 12$ degrees of freedom (in this case the result is significant)
A resampling approach to distributions of the likelihood ratio statistics

Goldman (1993) suggests that, in cases where we may wonder whether the Likelihood Ratio Test statistic really has its desired $\chi^2$ distribution we can:

- Take our best estimate of the tree
- Simulate on it the evolution of data sets of the same size
- For each replicate, calculate the LRT statistic
- Use this as the distribution and see where the actual LRT value lies in it (e.g.: in the upper 5%?)

This, of course, is a parametric bootstrap.
Two trees to be tested by paired sites tests
<table>
<thead>
<tr>
<th>Tree</th>
<th>site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>231</th>
<th>232</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td></td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>-0.431</td>
<td>⋯</td>
<td>+0.012</td>
<td>+0.010</td>
</tr>
</tbody>
</table>

**Differences in log likelihoods site by site**
Difference in log likelihood at site

Histogram of log likelihood differences
Paired sites tests

- Winning sites test (Prager and Wilson, 1988). Do a sign test on the signs of the differences.
- $z$ test (me, 1993 in PHYLIP documentation). Assume differences are normal, do $z$ test of whether mean (hence sum) difference is significant.
- $t$ test. Swofford et. al., 1996: do a $t$ test (paired)
- Wilcoxon ranked sums test (Templeton, 1983).
- RELL test (Kishino and Hasegawa, 1989 per my suggestion). Bootstrap resample sites, get distribution of difference of totals.
In this example …

• Winning sites test. 160 of 232 sites favor tree l. 
  \[ P < 3.279 \times 10^{-9} \]
• \( z \) test. Difference of log-likelihood totals is 0.948104 
  standard deviations from 0, \( P = 0.343077 \). Not 
  significant.
• \( t \) test. Same as \( z \) test for this large a number of sites.
• Wilcoxon ranked sums test. Rank sum is 4.82805 
  standard deviations below its expected value, 
  \( P = 0.00001378765 \)
• RELL test. 8,326 out of 10,000 samples have a positive 
  sum, \( P = 0.3348 \) (two-sided)
for each parameter value, find data values (unshaded) that account for 95% of the probability

then, given a data value, the parameters that are in the 95% confidence region are those for which that data value is in the unshaded region

The logic of confidence intervals
A toy case: 3 species with a clock

If we have 3 species and assume a molecular clock, in the Jukes-Cantor model of base change there are only 4 classes of data outcome that differ: \( xxx, xxy, xyx, yxx \), and \( xyz \).

Ignoring the first and the last (which loses us some information about tree depths) our choice among the three possible rooted tree topologies will depend on the probabilities of the three outcomes \( xxy, xyx, yxx \). Each tree predicts that one of these is highest and the other two equal:

<table>
<thead>
<tr>
<th>Tree</th>
<th>( xxy )</th>
<th>( xyx )</th>
<th>( yxx )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(((A,B),C))</td>
<td>( p )</td>
<td>( q )</td>
<td>( q )</td>
</tr>
<tr>
<td>(((A,C),B))</td>
<td>( q )</td>
<td>( p )</td>
<td>( q )</td>
</tr>
<tr>
<td>( (A,(B,C)))</td>
<td>( q )</td>
<td>( q )</td>
<td>( p )</td>
</tr>
</tbody>
</table>

(where \( p > q \) and \( p + 2q = 1 \))
Expected frequencies of xxy, xyx, yxx:

\[ (p_1, p_2, p_3) \]

\[ (p, q, q) \]

\[ (q, p, q) \]

\[ (q, q, p) \]

The three trees then correspond to three lines in an equilateral triangle
Test of 3–species Tree with a Clock

(Felsenstein, 1985)

(informative characters)

possible data

confidence region

statistic: number of steps different between best and next best tree

<table>
<thead>
<tr>
<th>Chars</th>
<th>S (0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9–13</td>
<td>5</td>
</tr>
<tr>
<td>14–20</td>
<td>6</td>
</tr>
<tr>
<td>21–29</td>
<td>7</td>
</tr>
</tbody>
</table>
Bayesian approach to interval estimation

In the Bayesian framework, one can avoid the separate calculation of confidence intervals. The posterior distribution of trees shows us how much credence to give different trees (for example, it assigns probabilities to different tree topologies).

The unresolved issue is how to summarize this posterior distribution in the best way. In this respect Bayesian methods leave you in a situation analogous to having the cloud of bootstrap-sampled trees without yet having summarized them.
References


