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: Distance Methods

- Distance matrixes
- Mutational models
- Distance phylogeny methods
Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimp</th>
<th>Orang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>aactc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chimp</td>
<td>aagtc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orang</td>
<td>tagtt</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

becomes

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>C</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>-</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>O</td>
<td>3</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>
Distance Methods

- Tree is built using distances rather than original data

- Only possible method if data were originally distances:
  - immunological cross-reactivity
  - DNA annealing temperature

- Can also be used on DNA, protein sequences, etc.
Large distances are underestimated by raw counts
A mutational model allows corrected distances

Jukes-Cantor model:

\[ D = -\frac{3}{4} \ln(1 - \frac{4}{3} D_s) \]

- \( D \) is the corrected distance (what we want)
- \( D_s \) is the raw count (what we have)
- \( \ln \) is the natural log
Mutational models for DNA

- Jukes-Cantor (JC): all mutations equally likely
- Kimura 2-parameter (K2P): transitions more likely than transversions
- Felsenstein 84 (F84): K2P plus unequal base frequencies
- Generalized Time Reversible (GTR): most general usable model

Models more complex than GTR would be useful but are very hard to work with.
Mutational models for protein sequence

- We have already seen these in alignment (BLOSUM etc.)

- Protein models are usually built from empirical data
Distances into trees

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.23</td>
<td>0.16</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>B</td>
<td>0.23</td>
<td>0</td>
<td>0.23</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>C</td>
<td>0.16</td>
<td>0.23</td>
<td>0</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>D</td>
<td>0.20</td>
<td>0.17</td>
<td>0.15</td>
<td>0</td>
<td>0.21</td>
</tr>
<tr>
<td>E</td>
<td>0.17</td>
<td>0.24</td>
<td>0.11</td>
<td>0.21</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Distances into trees

- Not all sets of distances fit a tree perfectly
- For those that do, finding the tree is simple
- If no tree fits perfectly, which one is best?
Least squares

• Least squares rule: prefer the tree for which the sum of

\[(\text{observed} - \text{expected})^2\]

is minimized.

• This means that getting a long branch wrong is penalized much more heavily than getting a short branch wrong

• Some least-squares methods add weights to this calculation to allow for long distances being less accurately measured than short ones
Minimum evolution and neighbor-joining

- Minimum evolution rule: for each topology, find the best branch lengths by least-squares
- Then, choose the topology with the lowest total branch lengths
- The popular neighbor-joining algorithm is a very fast approximation to ME
- Neighbor-joining gains its speed by considering very few trees
- It uses a clustering approach rather than a tree search
- Surprisingly, it works quite well
The molecular clock

• The molecular clock is the hypothesis that the rate of evolution is constant over time and across species.

• This is almost never true.

• It is most nearly true:
  – among closely related species
  – among species with similar generation time and life history
  – for genetic regions with the same function in all species, or no function
The molecular clock

• Even when the clock is doubtful, it is often assumed in order to:
  – put a root on the tree
  – infer the times at which species arose
  – estimate the rate of mutation

• When the data are not really clocklike, assuming a clock will often result in inferring the wrong tree
  – Branch lengths will certainly be wrong
  – Topology will often be wrong

• Statistical tests for clock violation are available and should be used
Distance methods summary

• All distance methods lose some information in making the distances

• Which algorithm you use is much less important than a good distance correction

• The more you know about the evolutionary process, the better you can correct the distances

• Distance methods are popular because they are fast and can be used with a variety of models
Judging tree-inference methods

Points to consider:

• Consistency: would it get the right answer with infinite data and a correct model?
  – Parsimony is not consistent
  – Distance methods with properly corrected distances are

• Robustness: how much is it hurt by a wrong model?
  – Distance methods can be highly vulnerable
  – Parsimony is more robust

• Power: how well can it do with limited data?

• Speed: can I stand to run it?
  – Methods that are consistent, robust and powerful tend to be slow
Judging tree-inference methods

Points to consider:

- **Availability**: can I find a program to do this?
  - The PHYLIP package is a good free source of phylogeny programs
  - Links to huge list of other available programs

- **Intended use**: what do I need from my phylogenies?