Botany 940 - Review 1

- Parsimony
- Likelihood Principle
- Ingredients of molecular evolution models:
  - Substitution rate matrix
  - Distribution of rates
- Bootstrap
- Divergence times

Next week: Bayesian methods; Testing monophyly
Parsimony

a: TAATAATTTAAAA
b: TATTTTTTTATAA
c: TAAATAATTTTTT
d: TAATAAATTTAAT

ab|cd L=11
ac|bd L=14
ad|bc L=12
12 volunteers:
choose a number between 1 and 12
Get a new alignment and a new MP tree
Repeat…
Likelihood principle

- Probability model: can be used to make predictions

- Data + probability model $\rightarrow$ likelihood
  probability of the data under the model, measures the fit of the model to the data.
Maximum Likelihood

- **ML tree**: find the tree (and parameters) that maximize the likelihood

<table>
<thead>
<tr>
<th>Tree</th>
<th>score: -log likelihood</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>cd</td>
<td>30.34</td>
</tr>
<tr>
<td>ac</td>
<td>bd</td>
<td>31.02</td>
</tr>
<tr>
<td>ad</td>
<td>bc</td>
<td>31.02</td>
</tr>
</tbody>
</table>
Bootstrap with Maximum Likelihood

- 12 volunteers:
  choose a number between 1 and 12
- Get a new alignment and a new ML tree
- Repeat…
Ingredients of molecular evolution models

- **Topology**
- **Branch lengths**: in number of substitutions per site. Usually unconstrained (no clock)
- **Substitution rate matrix**

<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>-1</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>C</td>
<td>.33</td>
<td>-1</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>G</td>
<td>.33</td>
<td>.33</td>
<td>-1</td>
<td>.33</td>
</tr>
<tr>
<td>T</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
<td>-1</td>
</tr>
</tbody>
</table>

<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>-1</td>
<td>.20</td>
<td>.60</td>
<td>.20</td>
</tr>
<tr>
<td>C</td>
<td>.20</td>
<td>-1</td>
<td>.20</td>
<td>.60</td>
</tr>
<tr>
<td>G</td>
<td>.60</td>
<td>.20</td>
<td>-1</td>
<td>.20</td>
</tr>
<tr>
<td>T</td>
<td>.20</td>
<td>.60</td>
<td>.20</td>
<td>-1</td>
</tr>
</tbody>
</table>

- **JC**
- **K80, k=3**
- **F81, unequal base freq.**
Ingredients of molecular evolution models

Common models: reversible. The root doesn’t matter.

<table>
<thead>
<tr>
<th>Model</th>
<th>Base freq.</th>
<th>ti/tv ratio</th>
<th># param</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>equal</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>K80</td>
<td>equal</td>
<td>free</td>
<td>1</td>
</tr>
<tr>
<td>F81</td>
<td>free</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>HKY85</td>
<td>free</td>
<td>free</td>
<td>4</td>
</tr>
<tr>
<td>F84</td>
<td>free</td>
<td>free</td>
<td>4</td>
</tr>
<tr>
<td>TN93</td>
<td>free</td>
<td>two ti/tv ratios are free</td>
<td>5</td>
</tr>
<tr>
<td>GTR</td>
<td>free</td>
<td>All rate ratios are free</td>
<td>8</td>
</tr>
</tbody>
</table>
Ingredients of molecular evolution models

- Distribution of rates across sites
  - Some sites are slow, some are fast
  - Model: each site chooses a rate $r$ at random, from some distribution

\[ \begin{align*}
\text{r=0.2} & \quad \text{r=1} & \quad \text{r=1.2} & \quad \text{r=0.8}
\end{align*} \]
Ingredients of molecular evolution models

- **Distribution of rates across sites**
  - Model: Gamma distribution of rates, continuous and discretized for calculations
Ingredients of molecular evolution models

- Distribution of rates across lineages
  - Covarion models: hidden on/off process for absence/presence of constraint. 1-3 parameters.
Ingredients of molecular evolution models

- Partitions in the data
  - 1st, 2nd and 3rd codon positions
  - several genes are available,
  - or several exons and introns,
  - or nuclear + mt DNA.

- Allow ‘blocks’ to have their own models and/or parameters
Estimation of Divergence Times

- Times and rates are confounded. Sites can evolve for 1 million years at $5 \times 10^{-7}$ subst./year/site or for 10,000 years at $5 \times 10^{-5}$ subst./year/site, the alignment will look the same: 0.5 subst./site on average.

- Difficult problem:
  - fossil calibrations usually few and uncertain
  - rates vary a lot