Models of Sequence Evolution I

Stochastic error and models of sequence evolution.
- Random error:
  - Data too sparse to recover true tree
  - Sequences too short to describe evolutionary history.
- Systematic error:
  - Erroneous model assumptions
  - Data biased toward the wrong tree
  - Unrealistic model fails to correct bias.

Systematic Error
- Sources of bias - possible departures from simple model assumptions.
- Multiples substitutions obscure some changes
- Unequal base (nucleotide) frequencies.
- Transition to transversion bias.
- Across site rate heterogeneity.

12 possible substitution types
Each site in the sequence can change in one of 12 ways.

Count of all 12 observed changes summed across all 7844 sites.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>from</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>652</td>
<td>752</td>
<td>639</td>
</tr>
<tr>
<td>C</td>
<td>481</td>
<td>-</td>
<td>276</td>
<td>1286</td>
</tr>
<tr>
<td>G</td>
<td>781</td>
<td>274</td>
<td>-</td>
<td>206</td>
</tr>
<tr>
<td>T</td>
<td>344</td>
<td>846</td>
<td>168</td>
<td>-</td>
</tr>
</tbody>
</table>

Springer 1999 sequence alignment of 11 mammals

We simplify to 6 ways for nucleotides to differ between taxa.

The case of 12 changes is nearly mathematically intractable. We simplify to 6 reversible changes. Big model simplification but usually works fine.
So six changes averaged across the diagonal.

Jukes Cantor 69: 1 substitution type

Jukes Cantor one parameter

Jukes Cantor also assumes equal probabilities of change.

Jukes Cantor also assumes equal probabilities of change.

In this case a serious departure from Jukes Cantor model

Corrects for multiple substitutions erased by time

describes the probability of change.

Realistic assumption?
Kimura80 (K2P) probability matrix

from

\[
\begin{pmatrix}
1 + (\alpha - 2 \beta) & \beta & \alpha & \beta a \\
\beta & 1 + (\alpha - 2 \beta) & \beta & \alpha \\
\alpha & \beta & 1 + (\alpha - 2 \beta) & \beta \\
\beta & \alpha & \beta & 1 + (\alpha - 2 \beta)
\end{pmatrix}
\]

to

Kimura80: 2 substitution types

transitions

HKY85 substitution matrix

\[
\begin{array}{cccc}
A & C & G & T \\
\alpha & \beta & \alpha & \beta \\
\beta & \alpha & \beta & \alpha \\
\alpha & \beta & \alpha & \beta \\
\beta & \alpha & \beta & \alpha \\
\end{array}
\]

 HKY85 model adds unequal base (nucleotide frequencies).

GTR substitution matrix adds 4 more substitution types.

\[
\begin{array}{cccc}
A & C & G & T \\
\alpha & \beta & \alpha & \beta \\
\beta & \alpha & \beta & \alpha \\
\alpha & \beta & \alpha & \beta \\
\beta & \alpha & \beta & \alpha \\
\end{array}
\]

Nucleotide frequencies: \( \pi_i \neq \pi_j \neq \pi_k \neq \pi_l \)
GTR: 6 substitution types and unequal base frequency

Models nested within the general time reversible (GTR) model

- Equal base frequencies
  - JC69 1 substitution type (ST)
  - K80 2 STs (transitions and transversions)
- Add unequal base frequencies
  - F81 1 ST
  - HKY/F84 2 STs
  - GTR 6 STs (A<->G, A<->T, ...)

Increasingly complex models

So different trees for different models

- How do we know if more complex models are better?
  - What is our criteria for considering one tree or one method better than another.
  - Parsimony at least gives us objective criteria for distinguishing among trees.
  - We settle on the tree(s) that most parsimoniously describe evolutionary history.
  - But we can study a whole range of alternatives.
  - Distance methods provide no alternative trees.

Maximum likelihood methods

- Like parsimony allows evaluation of alternative close trees.
- Like distance allows modeling different processes describing evolutionary history.
- But has many other advantages over the parsimony and distance methods.
- Includes Bayesian methods which are based on likelihood models.

Recall the Felsenstein zone: long branch attraction
Felsenstein zone

Four taxon tree with branches of lengths p (long) and q (short)

\[ \text{Branch length } a, b, c, d, e \]

The space of four tree shapes

Long branch attraction

Compare methods of tree inference.

Efficient methods: shorter sequence needed to converge to the true tree.

Consistent methods: the tree converges to the right tree as the sequence length increases.

No correction for multiple substitutions.

NJ and UPGMA: longer sequence needed to converge.

for many tree types the probability of converging to the wrong tree increases as the sequence length increases.

Correction for multiple substitutions.

Jukes-Cantor model

NJ: no more efficient than UPGMA.

Neighbor joining: More consistent than UPGMA. The problem area of the tree space is smaller so more likely to get the right tree.
Somewhat more efficient than NJ or UPGMA
Similar in consistency to NJ.

Maximum likelihood,
Kimura model
Compared to others ML methods are
Highly consistent
Most efficient
And importantly most robust to deviations from model assumptions.
This is true for all statistical methods not just phylogenetics.

We will see that maximum likelihood performs the best in almost every case.

Next week
- Maximum likelihood methods.
  - More on models
  - Criteria for deciding among models.
  - Criteria for deciding among trees.