Likelihood ratio test of the molecular clock

Unconstrained topology:
2n-3 = 11 edge lengths
logL = -5216.392
(10.4 higher than constr.)

Constrained topology:
n-1 = 6 node depths
logL = -5226.835

The Genetic Code

First 12 nucleotides at the 5' end of the \textit{rbcL} gene in corn:

\begin{align*}
5' & - \text{ATG} | \text{TCA} | \text{CCA} | \text{CAA} - 3' \\
3' & - \text{TAC} | \text{AGT} | \text{GGT} | \text{GTT} - 5'
\end{align*}

DNA double helix

\begin{center}
\textbf{Genetic Code}
\end{center}

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
U & C & A & G \\
\hline
\hline
UUU & Phe & UCU & Ser & UAU & Tyr \\
UUC & Phe & UCC & Ser & UAC & Tyr \\
UUA & Leu & UCA & Ser & UAA & Stop (UAG) \\
UUG & Leu & UCG & Ser & UGA & Stop (UGG) \\
\hline
CUU & Leu & CUC & Leu & CAU & His \\
CUA & Leu & CCA & Pro & CAC & His \\
CUG & Leu & CCG & Pro & CAA & Gin \\
\hline
AUU & Ile & AUC & Ile & AAA & Lys \\
AUA & Ile & ACA & Thr & AAG & Lys \\
\hline
GUU & Val & GCU & Ala & GAA & Glu \\
GUC & Val & GCC & Ala & GAG & Glu \\
GUA & Val & GCA & Ala & GGU & Gly \\
GUG & Val & GCG & Ala & GGC & Gly \\
\hline
\end{tabular}
\end{center}

Codon models treat codons as the independent units, not individual nucleotide sites.

\begin{center}
\textbf{transcription}
\end{center}

\begin{center}
\textbf{translation}
\end{center}

\begin{center}
5' - AUG | UCA | CCA | CAA - 3' \quad \text{mRNA}
\end{center}

N-Met | Ser | Pro | Gln - C \quad \text{polypeptide}

Paul O. Lewis ~ Phylogenetics, Spring 2020
http://www.langara.bc.ca/biology/mario/Assets/Geneticode.jpg

**Table I. Part of Muse and Gaut’s 61 × 61 instantaneous rate matrix**

<table>
<thead>
<tr>
<th>Codon before substitution (the ‘from’ state)</th>
<th>Codon after substitution (the ‘to’ state)</th>
<th>TTT (Phe)</th>
<th>TTC (Phe)</th>
<th>TTA (Leu)</th>
<th>TTG (Leu)</th>
<th>CTT (Leu)</th>
<th>CTC (Leu)</th>
<th>...</th>
<th>GGG (Gly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT (Phe)</td>
<td></td>
<td>---</td>
<td>απ₇</td>
<td>βπ₂</td>
<td>βπ₃</td>
<td>βπ₄</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>TTC (Phe)</td>
<td></td>
<td>απ₁</td>
<td>---</td>
<td>βπ₂</td>
<td>βπ₃</td>
<td>0</td>
<td>βπ₄</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>TTA (Leu)</td>
<td></td>
<td>βπ₁</td>
<td>βπ₂</td>
<td>---</td>
<td>απ₃</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>TTG (Leu)</td>
<td></td>
<td>βπ₁</td>
<td>βπ₂</td>
<td>απ₃</td>
<td>---</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>CTT (Leu)</td>
<td></td>
<td>βπ₁</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>---</td>
<td>απ₄</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>CTC (Leu)</td>
<td></td>
<td>0</td>
<td>βπ₁</td>
<td>0</td>
<td>0</td>
<td>απ₅</td>
<td>---</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>GGG (Gly)</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>


Interpreting codon model results

\[ \omega = \frac{\beta}{\alpha} \] is the nonsynonymous/synonymous rate ratio

<table>
<thead>
<tr>
<th>( \omega )</th>
<th>Mode of selection</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega &lt; 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \omega = 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \omega &gt; 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Amino acid models

### WAG (Whelan And Goldman) rate matrix:

|     | A     | R     | N     | D     | C     | Q     | E     | G     | H     | I     | L     | K     | M     | F     | P     | S     | T     | W     | Y     | V     |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A   | 0.55157 | 0.50985 | 0.73900 | 1.02704 | 0.90860 | 1.58285 | 1.41672 | 0.31695 | 0.19334 | 0.39792 | 0.90627 | 0.89350 | 0.21049 | 0.55157 | 1.48385 | 3.37079 | 2.12111 | 0.11313 | 0.24074 | 2.00601 |
| R   | 1.16376 | 0.63535 | 0.14730 | 5.42942 | 0.26526 | 0.61678 | 0.09882 | 0.24941 | 0.11392 | 0.15426 | 3.89490 | 2.58443 | 0.37356 | 0.25756 | 0.90860 | 1.54364 | 3.95629 | 0.47986 | 0.07403 | 0.66531 | 0.19203 |
| N   | 0.55441 | 0.55424 | 0.93068 | 0.24897 | 4.29411 | 0.57003 | 0.11392 | 0.03944 | 0.28429 | 0.86949 | 0.15426 | 0.90860 | 0.21049 | 0.43916 | 0.19822 | 0.12740 | 0.13153 | 0.08480 | 0.49767 | 0.09882 |
| D   | 0.67949 | 0.03944 | 0.24897 | 4.29411 | 0.57003 | 0.11392 | 0.03944 | 0.28429 | 0.86949 | 0.15426 | 0.90860 | 0.21049 | 0.43916 | 0.19822 | 0.12740 | 0.13153 | 0.08480 | 0.49767 | 0.09882 |
| C   | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 | 0.10271 |
| Q   | 5.35142 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 |
| E   | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 | 0.49767 |
| G   | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 | 0.58467 |
| H   | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 | 0.37356 |
| I   | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 | 1.54364 |
| L   | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 | 0.25756 |

### Relative amino acid frequencies

Amino Acid Models

There has been a regular progression of empirical amino acid models proposed since the initial effort by Margaret Dayhoff and colleagues. Later matrices are based on much larger protein sequence databases (LG is based on 50,000 sequences).

- **PAM ➢ 1978** (used closely-related amino-acid sequences)
- **JTT ➢ 1992** (based on much larger database)
- **WAG ➢ 2001** (Q MLE allowed more divergent sequences)
- **LG ➢ 2008** (added rate heterogeneity to Q MLE)


Transition Probabilities Revisited

JC69 rate matrix:

\[ Q = \begin{pmatrix}
-3\beta & \beta & \beta & \beta \\
\beta & -3\beta & \beta & \beta \\
\beta & \beta & -3\beta & \beta \\
\beta & \beta & \beta & -3\beta \\
\end{pmatrix} \]

JC69 transition probability matrix:

\[ P = \begin{pmatrix}
\frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t} \\
\end{pmatrix} \]

When I (earlier) derived these transition probabilities, I did it **analytically** (i.e. using algebra). We need to obtain transition probabilities **numerically** if the rate matrix is obtained empirically. 

\[ P = e^{Qt} \]
Matrix multiplication

\[ \mathbf{A} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} \]
Matrix inverse

\[
A = \begin{pmatrix}
1 & -1 \\
1 & 0
\end{pmatrix} \quad A^{-1} = \begin{pmatrix}
0 & 1 \\
-1 & 1
\end{pmatrix}
\]
Start with the instantaneous rate matrix $Q$

\[ P = e^{Qt} \]

\[
Q = \begin{pmatrix}
-3\beta & \beta & \beta & \beta \\
\beta & -3\beta & \beta & \beta \\
\beta & \beta & -3\beta & \beta \\
\beta & \beta & \beta & -3\beta \\
\end{pmatrix}
\]

Multiply $Q$ by $t$

\[
Qt = \begin{pmatrix}
-3\beta t & \beta t & \beta t & \beta t \\
\beta t & -3\beta t & \beta t & \beta t \\
\beta t & \beta t & -3\beta t & \beta t \\
\beta t & \beta t & \beta t & -3\beta t \\
\end{pmatrix}
\]

In order to raise $e$ to the power $Qt$, we must first **diagonalize** $Qt$. Diagonalization involves finding eigenvalues and eigenvectors such that $Qt$ can be represented as the following matrix multiplication...
\[ P = e^{Qt} \]

\[ Q_t = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -4\beta t & 0 & 0 \\ 0 & 0 & -4\beta t & 0 \\ 0 & 0 & 0 & -4\beta t \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix} \]
\[ P = eQ^t \]

\[
e^Q^t = \begin{pmatrix}
1 & -1 & -1 & -1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 \\
\end{pmatrix}
\begin{pmatrix}
e^0 & 0 & 0 & 0 \\
0 & e^{-4\beta t} & 0 & 0 \\
0 & 0 & e^{-4\beta t} & 0 \\
0 & 0 & 0 & e^{-4\beta t} \\
\end{pmatrix}
\begin{pmatrix}
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
-\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\
-\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\
-\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\
\end{pmatrix}
\]
$P = e^{Qt} = \begin{pmatrix}
\frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} - \frac{1}{4} e^{-4\beta t} & \frac{1}{4} + \frac{3}{4} e^{-4\beta t}
\end{pmatrix}$
Stems and loops


Paul O. Lewis ~ Phylogenetics, Spring 2020

Compensatory substitutions
Muse (1995) stem model

• Apply normal substitution model (e.g. JC69, K80, HKY85, etc.) to loop regions
• Apply a special "stem model" to the stem regions
• The stem model has 16 states: AA, AC, AG, AT, CA, CC, ..., TT
  where each state consists of one possible pairing of two nucleotides across a stem

Muse (1995) stem model

- Idea is to compare null (independence) model to the alternative (dependence) model using a likelihood ratio test
- **Independence** model assumes each site evolves independently
- **Dependence** model allows rate of evolution to be *higher* for changes that *improve stem stability* and lower for changes that destroy stability
Independence model
(applied to one stem site)

\[ L_{\text{stem ``site''}} = \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} + \frac{3}{4} e^{-\beta t} \right) \right] \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} - \frac{1}{4} e^{-\beta t} \right) \right] \]

G → G (no change)  U → C (change)
The instantaneous rate equals the slope at time 0 of the transition probability curve.

\[
Pr(\text{no change}) = \frac{1}{4} + \frac{3}{4} e^{-4\beta t}
\]

\[
Pr(\text{change}) = \frac{1}{4} - \frac{1}{4} e^{-4\beta t}
\]
Instantaneous rate for a change at one of the two sites is $\beta$

$$\Pr(1\ of\ 2\ sites\ change) = \begin{bmatrix} \frac{1}{4} + \frac{3}{4}e^{-4\beta t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \end{bmatrix}$$

$$\Pr(both\ sites\ change) = \begin{bmatrix} \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \end{bmatrix}$$

Instantaneous rate for a change at both sites is zero!

slope = $\beta$

slope = 0
Independence model rate matrix*

*I have used a JC model here; Muse used HKY85
# Dependence model rate matrix

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AC</th>
<th>AG</th>
<th>AU</th>
<th>CA</th>
<th>CC</th>
<th>CG</th>
<th>CU</th>
<th>GA</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>-</td>
<td>β</td>
<td>β</td>
<td>λβ</td>
<td>β</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>β</td>
<td>...</td>
</tr>
<tr>
<td>AC</td>
<td>β</td>
<td>-</td>
<td>β</td>
<td>λβ</td>
<td>0</td>
<td>β</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>AG</td>
<td>β</td>
<td>β</td>
<td>-</td>
<td>λβ</td>
<td>0</td>
<td>0</td>
<td>λβ</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>AU</td>
<td>β/λ</td>
<td>β/λ</td>
<td>β/λ</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>β/λ</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>CA</td>
<td>β</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>β</td>
<td>λβ</td>
<td>β</td>
<td>β</td>
<td>...</td>
</tr>
<tr>
<td>CC</td>
<td>0</td>
<td>β</td>
<td>0</td>
<td>0</td>
<td>β</td>
<td>-</td>
<td>λβ</td>
<td>β</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>CG</td>
<td>0</td>
<td>0</td>
<td>β/λ</td>
<td>0</td>
<td>0</td>
<td>β/λ</td>
<td>-</td>
<td>β/λ</td>
<td>β/λ</td>
<td>...</td>
</tr>
<tr>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>λβ</td>
<td>0</td>
<td>β</td>
<td>λβ</td>
<td>-</td>
<td>β</td>
<td>...</td>
</tr>
<tr>
<td>GA</td>
<td>β</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>