Week 5: Distance methods, DNA and protein models

Genome 570

February, 2010
Kimura 2-parameter model
Parameters of K2P in terms of transition/transversion ratio

\[ \alpha = \frac{R}{R+1} \]

\[ \beta = \left( \frac{1}{2} \right) \frac{1}{R+1} \]
Transition [sic] probabilities for K2P

\[
\begin{align*}
\text{Prob (transition}\,|\,t) &= \frac{1}{4} - \frac{1}{2} \exp \left(-\frac{2R+1}{R+1}t\right) + \frac{1}{4} \exp \left(-\frac{2}{R+1}t\right) \\
\text{Prob (transversion}\,|\,t) &= \frac{1}{2} - \frac{1}{2} \exp \left(-\frac{2}{R+1}t\right).
\end{align*}
\]
Transition and transversion when $R = 10$
Transition and transversion when $R = 2$
ML estimates for the K2P model

\[
\hat{t} = -\frac{1}{4} \ln \left[ (1 - 2Q)(1 - 2P - Q)^2 \right]
\]

\[
\hat{R} = \frac{-\ln(1-2P-Q)}{-\ln(1-2Q)} - \frac{1}{2}
\]
Likelihood for two species under the K2P model

\[ L = \text{Prob} (\text{data} \mid t, R) \]

\[ = \left( \frac{1}{4} \right)^n (1 - P - Q)^{n - n_1 - n_2} P^{n_1} \left( \frac{1}{2} Q \right)^{n_2} \]

where \( n_1 \) is the number of sites differing by transitions, and \( n_2 \) is the number of sites differing by transversions. \( P \) and \( Q \) are the expected fractions of transition and transversion differences, as given by the expressions three screens above.
The Tamura/Nei model, F84, and HKY

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>From A</td>
<td>$-\ \alpha R\pi_G/\pi_R + \beta\pi_G$</td>
<td>$\beta\pi_C$</td>
<td>$\beta\pi_T$</td>
<td></td>
</tr>
<tr>
<td>From G</td>
<td>$\alpha R\pi_A/\pi_R + \beta\pi_A$</td>
<td>$-\ \beta\pi_C$</td>
<td>$\beta\pi_T$</td>
<td></td>
</tr>
<tr>
<td>From C</td>
<td>$\beta\pi_A$</td>
<td>$\beta\pi_G$</td>
<td>$-\ \alpha Y\pi_T/\pi_Y + \beta\pi_T$</td>
<td></td>
</tr>
<tr>
<td>From T</td>
<td>$\beta\pi_A$</td>
<td>$\beta\pi_G$</td>
<td>$\alpha Y\pi_C/\pi_Y + \beta\pi_C$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

For the F84 model, $\alpha_R = \alpha_Y$

For the HKY model, $\alpha_R/\alpha_Y = \pi_R/\pi_Y$
Transition/transversion ratio for the Tamura-Nei model

\[ T_s = 2 \alpha_R \pi_A \pi_G / \pi_R + 2 \alpha_Y \pi_C \pi_T / \pi_Y \]

\[ + \beta \left( \pi_A \pi_G + 2 \pi_C \pi_T \right) \]

\[ T_v = 2 \beta \pi_R \pi_Y \]

To get \( T_s / T_v = R \) and \( T_s + T_v = 1 \),

\[ \beta = \frac{1}{2 \pi_R \pi_Y (1 + R)} \]

\[ \alpha_Y = \frac{\pi_R \pi_Y R - \pi_A \pi_G - \pi_C \pi_T}{(1 + R) \left( \pi_Y \pi_A \pi_G \rho + \pi_R \pi_C \pi_T \right)} \]

\[ \alpha_R = \rho \alpha_Y \]
Using fictional events to mimic the Tamura-Nei model

We imagine two types of events:

- **Type I:**
  - If the existing base is a purine, draw a replacement from a purine pool with bases in relative proportions $\pi_A : \pi_G$. This event has rate $\alpha_R$.
  - If the existing base is a pyrimidine, draw a replacement from a pyrimidine pool with bases in relative proportions $\pi_C : \pi_T$. This event has rate $\alpha_Y$.

- **Type II:** No matter what the existing base is, replace it by a base drawn from a pool at the overall equilibrium frequencies:
  $\pi_A : \pi_C : \pi_G : \pi_T$. This event has rate $\beta$. 
Transition [sic] probabilities with the Tamura-Nei model

If the branch starts with a purine:
- No events: \( \exp(- (\alpha_R + \beta) t) \)
- Some type I, no type II: \( \exp(-\beta t) \left( 1 - \exp(-\alpha_R t) \right) \)
- Some type II: \( 1 - \exp(-\beta t) \)

If the branch starts with a pyrimidine:
- No events: \( \exp(- (\alpha_Y + \beta) t) \)
- Some type I, no type II: \( \exp(-\beta t) \left( 1 - \exp(-\alpha_Y t) \right) \)
- Some type II: \( 1 - \exp(-\beta t) \)
A transition probability

So if we want to compute the probability of getting a G given that a branch starts with an A, we add up

- The probability of no events, times 0 (as you can’t get a G from an A with no events)
- The probability of “some type I, no type II” times $\frac{\pi_G}{\pi_Y}$ (as the last type I event puts in a G with probability equal to the fraction of G’s out of all purines).
- The probability of “some type II” times $\pi_G$ (as if there is any type II event, we thereafter have a probability of G equal to its overall expected frequency, and further type I events don’t change that).
A transition probability

So that, for example

\[
\text{Prob } (G|A, t) = \\
\exp(-\beta t) \left( 1 - \exp(-\alpha_R t) \right) \frac{\pi_G}{\pi_R} \\
+ \left( 1 - \exp(-\beta t) \right) \pi_G
\]
A more compact expression

More generally, we can use the Kronecker delta notation $\delta_{ij}$ and the "Watson-Kronecker" notation $\varepsilon_{ij}$ to write

$$\text{Prob} (j \mid i, t) = \exp(-(\alpha_i + \beta)t) \delta_{ij}$$
$$+ \exp(-\beta t) (1 - \exp(-\alpha_i t)) \left( \frac{\pi_j \varepsilon_{ij}}{\sum_k \varepsilon_{jk} \pi_k} \right)$$
$$+ (1 - \exp(-\beta t)) \pi_j$$

where $\delta_{ij}$ is 1 if the two bases $i$ and $j$ are different (0 otherwise), and $\varepsilon_{ij}$ is 1 if, of the two bases $i$ and $j$, one is a purine and one is a pyrimidine (0 otherwise).
Reversibility and the GTR model

\[ \pi_i \text{ Prob } (j|i, t) = \pi_j \text{ Prob } (i|j, t) \]

The general time-reversible model:

<table>
<thead>
<tr>
<th>From:</th>
<th>To:</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>\pi_G \alpha</td>
<td>\pi_C \beta</td>
<td>\pi_T \gamma</td>
</tr>
<tr>
<td>G</td>
<td>\pi_A \alpha</td>
<td>\pi_G \delta</td>
<td>\pi_C \eta</td>
<td>\pi_T \epsilon</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>\pi_A \beta</td>
<td>\pi_G \delta</td>
<td>\pi_C \eta</td>
<td>\pi_T \epsilon</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>\pi_A \gamma</td>
<td>\pi_G \epsilon</td>
<td>\pi_C \eta</td>
<td>\pi_T \delta</td>
<td></td>
</tr>
</tbody>
</table>
The GTR model
Standardizing the rates

\[ 2\pi_A \pi_G \alpha + 2\pi_A \pi_C \beta + 2\pi_A \pi_T \gamma \]

\[ + 2 \pi_G \pi_C \delta + 2 \pi_G \pi_T \varepsilon + 2 \pi_C \pi_T \eta = 1 \]
General Time Reversible models – inference

A data example (simulated under a K2P model, true distance 0.2 transition/transversion ratio = 2)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>93</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>112</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>105</td>
<td>3</td>
<td>4</td>
<td>122</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>4</td>
<td>113</td>
<td>18</td>
<td>141</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>4</td>
<td>21</td>
<td>93</td>
<td>125</td>
</tr>
<tr>
<td>total</td>
<td>116</td>
<td>126</td>
<td>140</td>
<td>118</td>
<td>500</td>
</tr>
</tbody>
</table>
Averaging across the diagonal ...

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>93</td>
<td>11.5</td>
<td>4.5</td>
<td>5</td>
<td>114</td>
</tr>
<tr>
<td>G</td>
<td>11.5</td>
<td>105</td>
<td>3.5</td>
<td>4</td>
<td>124</td>
</tr>
<tr>
<td>C</td>
<td>4.5</td>
<td>3.5</td>
<td>113</td>
<td>19.5</td>
<td>140.5</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>4</td>
<td>19.5</td>
<td>93</td>
<td>121.5</td>
</tr>
<tr>
<td>total</td>
<td>114</td>
<td>124</td>
<td>140.5</td>
<td>121.5</td>
<td>500</td>
</tr>
</tbody>
</table>
Dividing each column by its sum

(column, because \( P_{ij} \) is to be the probability of change from \( j \) to \( i \) )

\[
\hat{P} = \begin{bmatrix}
0.815789 & 0.0927419 & 0.0320285 & 0.0411523 \\
0.100877 & 0.846774 & 0.024911 & 0.0329218 \\
0.0394737 & 0.0282258 & 0.80427 & 0.160494 \\
0.0438596 & 0.0322581 & 0.13879 & 0.765432
\end{bmatrix}
\]
Rate matrix from the matrix logarithm

If the rate matrix is $A$, 

$$P = e^{At}$$

so that

$$\hat{A}t = \log(\hat{P})$$

$$= \begin{bmatrix}
-0.212413 & 0.110794 & 0.034160 & 0.046726 \\
0.120512 & -0.174005 & 0.025043 & 0.035554 \\
0.0421002 & 0.028375 & -0.236980 & 0.205579 \\
0.0498001 & 0.034837 & 0.177778 & -0.287859
\end{bmatrix}.$$
Standardizing the rates

If we denote by $\hat{D}$ the diagonal matrix of observed base frequencies, and we require that the rate of (potentially-observable) substitution is 1:

$$-\text{trace}(\hat{A}\hat{D}) = 1$$

We get:

$$\hat{t} = -\text{trace}(\hat{A}t\hat{D}) = -\text{trace} \left( \log(\hat{P})\hat{D} \right)$$

and that also gives us an estimate of the rate matrix:

$$\hat{A} = \log(\hat{P}) / - \text{trace} \left( \log(\hat{P})\hat{D} \right)$$
The rate estimates

\[ \hat{\mathbf{A}} = \begin{bmatrix}
-0.931124 & 0.485671 & 0.149741 & 0.204826 \\
0.528274 & -0.762764 & 0.109776 & 0.155852 \\
0.184549 & 0.124383 & -1.038820 & 0.901168 \\
0.218302 & 0.152710 & 0.779302 & -1.261850
\end{bmatrix}. \]

moderately close to the actual K2P rate matrix used in the simulation which was:

\[ \mathbf{A} = \begin{bmatrix}
-1 & 2/3 & 1/6 & 1/6 \\
2/3 & -1 & 1/6 & 1/6 \\
1/6 & 1/6 & -1 & 2/3 \\
1/6 & 1/6 & 2/3 & -1
\end{bmatrix} \]

(but if any of the eigenvalues of \( \log(\hat{\mathbf{P}}) \) are negative, this doesn’t work and the divergence time is estimated to be infinite).
The lattice of these models

- General 12-parameter model (12)
- General time-reversible model (9)
  - Tamura–Nei (6)
  - HKY (5)
  - F84 (5)
    - Kimura K2P (2)
    - Jukes–Cantor (1)
Variation of rates of evolution across sites

\[ L(t) = \prod_{i=1}^{\text{sites}} \left( \int_{0}^{\infty} f(r) \pi_{ni} P_{mni}(r \, t) \, dr \right) \]
The Gamma distribution

\[ f(r) = \frac{1}{\Gamma(\alpha) \beta^\alpha} r^{\alpha-1} e^{-\frac{r}{\beta}} \]

\[ E[x] = \alpha \beta \]

\[ \text{Var}[x] = \alpha \beta^2 \]

To get a mean of 1, set \( \beta = 1/\alpha \) so that

\[ f(r) = \frac{\alpha^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\alpha r}. \]

so that the squared coefficient of variation is \( 1/\alpha \).
Gamma distributions

- $\alpha = 0.25$, $cv = 2$
- $\alpha = 1$, $cv = 1$
- $\alpha = 11.1111$, $cv = 0.3$
Gamma rate variation in the Jukes-Cantor model

For example, for the Jukes-Cantor distance, to get the fraction of sites different we do

$$D_s = \int_0^\infty f(r) \frac{3}{4} \left(1 - e^{-\frac{4}{3}ru} \right) \, dr$$

leading to the formula for $D$ as a function of $D_s$

$$D = -\frac{3}{4} \alpha \left[ 1 - \left(1 - \frac{4}{3}D_s \right)^{-1/\alpha} \right]$$
Gamma rate variation in other models

For many other distances such as the Tamura-Nei family, the transition probabilities are of the form

\[ P_{ij}(t) = A_{ij} + B_{ij} e^{-bt} + C_{ij} e^{-ct} \]

and integrating termwise we can make use of the fact that

\[ E_r \left[ e^{-brt} \right] = \left( 1 + \frac{1}{\alpha b t} \right)^{-\alpha} \]
# Dayhoff’s PAM001 matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
<th>K</th>
<th>M</th>
<th>F</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>98672</td>
<td>9</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>17</td>
<td>21</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>22</td>
<td>35</td>
<td>32</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>99131</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>19</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>4</td>
<td>1</td>
<td>982236</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>9</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>0</td>
<td>42</td>
<td>98590</td>
<td>6</td>
<td>53</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>99730</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>987627</td>
<td>1</td>
<td>23</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>56</td>
<td>0</td>
<td>35</td>
<td>98654</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>21</td>
<td>1</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>99351</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>8</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>99120</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>98729</td>
<td>2</td>
<td>12</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>22</td>
<td>99472</td>
<td>45</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>37</td>
<td>25</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>992620</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>98741</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>99460</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>992612</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>28</td>
<td>11</td>
<td>34</td>
<td>7</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>17</td>
<td>984038</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>22</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>32</td>
<td>98710</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>57</td>
<td>11</td>
<td>1</td>
<td>17</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Week 5: Distance methods, DNA and protein models – p.31/34
The codon model

Goldman & Yang, MBE 1994; Muse and Weir, MBE 1994

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>phe</td>
<td>UUU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>phe</td>
<td>UUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>leu</td>
<td>UUA</td>
<td>ser</td>
<td>UCA</td>
</tr>
<tr>
<td>G</td>
<td>leu</td>
<td>UUG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>leu</td>
<td>CUU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>leu</td>
<td>CUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>leu</td>
<td>CUA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>leu</td>
<td>CUG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>ile</td>
<td>AUU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ile</td>
<td>AUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ile</td>
<td>AUA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>met</td>
<td>AUG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>val</td>
<td>GUU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>val</td>
<td>GUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>val</td>
<td>GUA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>val</td>
<td>GUG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probabilities of change vary depending on whether amino acid is changing, and to what extent.
A codon-based model of protein evolution

In each cell:

\[ P_{ij}(v) \cdot a_{ij} \]

where \( P_{ij}(v) \) is the probability of codon change

and \( a_{ij} \) is the probability that the change is accepted

<table>
<thead>
<tr>
<th></th>
<th>AAA</th>
<th>AAC</th>
<th>AAG</th>
<th>AAT</th>
<th>ACA</th>
<th>ACC</th>
<th>ACG</th>
<th>ACT</th>
<th>AGA</th>
<th>AGC</th>
<th>AGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>lys</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asn</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lys</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asn</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thr</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thr</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thr</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thr</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arg</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ser</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arg</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

observation:

lys lys lys lys
asn asn
lys lys lys
asn asn
thr thr thr thr
thr thr thr thr
thr thr thr thr
arg arg arg arg
ser ser ser ser
arg arg arg arg

Week 5: Distance methods, DNA and protein models – p.33/34
Considerations for a protein model

Making a model for protein evolution (a not-very-practical approach)

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept it with probability that declines as the amino acids become more different.
- Fit this to empirical information on protein evolution.
- Take into account variation of rate from site to site.
- Take into account correlation of rates in adjacent sites.
- How about protein structure? Secondary structure? 3D structure?

(the first four steps are the “codon model” of Goldman and Yang, 1994 and Muse and Gaut, 1994, both in Molecular Biology and Evolution. The next two are the rate variation machinery of Yang, 1995, 1996 and Felsenstein and Churchill, 1996).