Week 5: Distance methods, DNA and protein models

Genome 570

February, 2014
The Jukes-Cantor model

The simplest and most symmetrical of models of DNA evolution. In a small interval of time $dt$, the probability of change at a site is $u dt$, and it is equally likely to go to each of the other three bases.
A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).
A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).

- If you set the probability of change in that model to $\frac{4}{3} u dt$, it would be indistinguishable from the actual Jukes-Cantor model.
A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).
- If you set the probability of change in that model to \( \frac{4}{3} u dt \), it would be indistinguishable from the actual Jukes-Cantor model.
- If any nonzero number of these fictional events occur on a branch, the probability of ending up with (say) base C is \( \frac{1}{4} \).
A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).

- If you set the probability of change in that model to \( \frac{4}{3} u \, dt \), it would be indistinguishable from the actual Jukes-Cantor model.

- If any nonzero number of these fictional events occur on a branch, the probability of ending up with (say) base C is \( \frac{1}{4} \).

- The number of these events that occur in a branch has Poisson distribution with expected number \( \frac{4}{3} u \, t \), so the probability of no event is \( e^{-\frac{4}{3} u t} \).
The Jukes-Cantor model

Probability of no event: $e^{-\frac{4}{3}ut}$

Probability of some event: $1 - e^{-\frac{4}{3}ut}$

Probability of changing to C given start at A, have rate $u$, time $t$:

$$\text{Prob (C|A, u, t)} = \frac{1}{4} \left( 1 - e^{-\frac{4}{3}ut} \right)$$

Fraction of sites different:

$$f_D = \frac{3}{4} \left( 1 - e^{-\frac{4}{3}ut} \right).$$

Solving, the distance as function of the fraction of sites different is

$$D = \hat{u}t = -\frac{3}{4} \ln \left( 1 - \frac{4}{3}f_D \right)$$
Fraction of sites different versus branch length

\[ \frac{3}{4} \left( 1 - e^{-\frac{4}{3}rt} \right) \]

... as predicted by the Jukes-Cantor model.
Branch length versus fraction of sites different

\[-\frac{3}{4} \ln \left( 1 - \frac{4}{3} \text{diff} \right)\]

Week 5: Distance methods, DNA and protein models – p.6/63
If you don’t correct for multiple changes

This happening because the long path between A and C is shortened more, proportionally, than the shorter paths between A and B and between C and B. The only way to do this is to put in a spurious branch leading to B.

In effect, there is a “war” between the long paths and the short paths.

With properly corrected distances, each path approaches its true length when we look at a very large number of sites.
Least squares methods

These are just differently weighted least squares methods, as mentioned previously.

Fitch and Margoliash (1967) used a weight of $1/D_{ij}^2$, so the quantity to be minimized is

$$Q = \sum_{ij} \frac{(D_{ij} - d_{ij})^2}{D_{ij}^2}$$

Cavalli-Sforza and Edwards (1967) used the unweighted least squares method:

$$Q = \sum_{ij} (D_{ij} - d_{ij})^2$$

These amount to different assumptions about the how the size of a distance will affect its variance.
The Minimum Evolution method

Kidd and Sgaramella-Zonta (1971) and (independently) Rzhetsky and Nei (1992ff.) came up with this method:

- Search through tree space as usual
- For each tree estimate branch lengths by least squares, not allowing negative branch lengths
- Then actually evaluate the tree *not* by the sum of squares, but by the total sum of branch lengths.

This does fairly well, in spite of the mixture of two optimization criteria.

Note that it is *not* directly related to parsimony, in spite of its name.
The UPGMA algorithm

1. Assign each species weight $w_i = 1$
2. Pick the values of $i$ and $j$ that are for the smallest of the $D_{ij}$
3. Make a new group \((ij)\)
4. Assign time depth $D_{ij}/2$ to the connection between $i$ and $j$.
5. Its weight is $w_i + w_j$
6. The distance between \((ij)\) and another (say $k$) is computed as
   \[
   D_{(ij),k} = \frac{w_i D_{ik} + w_j D_{jk}}{w_i + w_j}
   \]
7. Delete $i$ and $j$ from the table, put in a row and column for \((ij)\).
8. If there is only one group left, stop. Otherwise go to step 2.

Note that the weighting in effect weights each of the original tips equally, so that a group’s distance to an outside species or another group is the average of all the distances between species in one and species in the other. This is a natural “unweighted” choice.
### Sarich 1969, immunological distances

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>bear</th>
<th>raccoon</th>
<th>weasel</th>
<th>seal</th>
<th>sea lion</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>32</td>
<td>48</td>
<td>51</td>
<td>50</td>
<td>48</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>bear</td>
<td>32</td>
<td>0</td>
<td>26</td>
<td>34</td>
<td>29</td>
<td>33</td>
<td>84</td>
<td>136</td>
</tr>
<tr>
<td>raccoon</td>
<td>48</td>
<td>26</td>
<td>0</td>
<td>42</td>
<td>44</td>
<td>44</td>
<td>92</td>
<td>152</td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>34</td>
<td>42</td>
<td>0</td>
<td>44</td>
<td>38</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>seal</td>
<td>50</td>
<td>29</td>
<td>44</td>
<td>44</td>
<td>0</td>
<td>24</td>
<td>89</td>
<td>142</td>
</tr>
<tr>
<td>sea lion</td>
<td>48</td>
<td>33</td>
<td>44</td>
<td>38</td>
<td>24</td>
<td>0</td>
<td>90</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>84</td>
<td>92</td>
<td>86</td>
<td>89</td>
<td>90</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>136</td>
<td>152</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Find smallest element, its rows, columns

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>bear</th>
<th>raccoon</th>
<th>weasel</th>
<th>seal</th>
<th>sea lion</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>32</td>
<td>48</td>
<td>51</td>
<td>50</td>
<td>48</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>bear</td>
<td>32</td>
<td>0</td>
<td>26</td>
<td>34</td>
<td>29</td>
<td>33</td>
<td>84</td>
<td>136</td>
</tr>
<tr>
<td>raccoon</td>
<td>48</td>
<td>26</td>
<td>0</td>
<td>42</td>
<td>44</td>
<td>44</td>
<td>92</td>
<td>152</td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>34</td>
<td>42</td>
<td>0</td>
<td>44</td>
<td>38</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>seal</td>
<td>50</td>
<td>29</td>
<td>44</td>
<td>44</td>
<td>0</td>
<td>24</td>
<td>89</td>
<td>142</td>
</tr>
<tr>
<td>sea lion</td>
<td>48</td>
<td>33</td>
<td>44</td>
<td>38</td>
<td>24</td>
<td>0</td>
<td>90</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>84</td>
<td>92</td>
<td>86</td>
<td>89</td>
<td>90</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>136</td>
<td>152</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
We do the following averaging

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>bear</th>
<th>raccoon</th>
<th>weasel</th>
<th>seal</th>
<th>sea lion</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>32</td>
<td>48</td>
<td>51</td>
<td>(50 + 48)/2</td>
<td>98</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>bear</td>
<td>32</td>
<td>0</td>
<td>26</td>
<td>34</td>
<td>(29 + 33)/2</td>
<td>84</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>raccoon</td>
<td>48</td>
<td>26</td>
<td>0</td>
<td>42</td>
<td>(44 + 44)/2</td>
<td>92</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>34</td>
<td>42</td>
<td>0</td>
<td>(44 + 38)/2</td>
<td>86</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>* seal</td>
<td>49</td>
<td>31</td>
<td>44</td>
<td>41</td>
<td>0</td>
<td>24</td>
<td>89.5</td>
<td>142</td>
</tr>
<tr>
<td>* sea lion</td>
<td>24</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>84</td>
<td>92</td>
<td>86</td>
<td>(89 + 90)/2</td>
<td>0</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>136</td>
<td>152</td>
<td>142</td>
<td>(142 + 142)/2</td>
<td>148</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Week 5: Distance methods, DNA and protein models – p.13/63
Clustering seal and sea lion
## After clustering seal and sea lion

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th><em>bear</em></th>
<th><em>raccoon</em></th>
<th>weasel</th>
<th>SS</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dog</strong></td>
<td>0</td>
<td>32</td>
<td>48</td>
<td>51</td>
<td>49</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td><strong>bear</strong></td>
<td>32</td>
<td>0</td>
<td><strong>26</strong></td>
<td>34</td>
<td>31</td>
<td>84</td>
<td>136</td>
</tr>
<tr>
<td><strong>raccoon</strong></td>
<td>48</td>
<td><strong>26</strong></td>
<td>0</td>
<td>42</td>
<td>44</td>
<td>92</td>
<td>152</td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>34</td>
<td>42</td>
<td>0</td>
<td>41</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>SS</td>
<td>49</td>
<td>31</td>
<td>44</td>
<td>41</td>
<td>0</td>
<td>89.5</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>84</td>
<td>92</td>
<td>86</td>
<td>89.5</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>136</td>
<td><strong>152</strong></td>
<td>142</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering bear and raccoon

- dog
- bear
- raccoon
- seal
- sea lion
- weasel
- cat
- monkey

13 13 12 12
After clustering bear and raccoon

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>BR</th>
<th>weasel</th>
<th>SS</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>40</td>
<td>51</td>
<td>49</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>* BR</td>
<td>40</td>
<td>0</td>
<td>38</td>
<td>37.5</td>
<td>88</td>
<td>144</td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>38</td>
<td>0</td>
<td>41</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>* SS</td>
<td>49</td>
<td>37.5</td>
<td>41</td>
<td>0</td>
<td>89.5</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>88</td>
<td>86</td>
<td>89.5</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>144</td>
<td>142</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering bear-raccoon with seal-sealion
After clustering those two clusters

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>BRSS</th>
<th>weasel</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>44.5</td>
<td>51</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>* BRSS</td>
<td>44.5</td>
<td>0</td>
<td>39.5</td>
<td>88.75</td>
<td>143</td>
</tr>
<tr>
<td>* weasel</td>
<td>51</td>
<td>39.5</td>
<td>0</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>88.75</td>
<td>86</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>143</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering weasel with BRSS
After adding weasel to that cluster

<table>
<thead>
<tr>
<th></th>
<th>* dog</th>
<th>* BRSSW</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>* dog</td>
<td>0</td>
<td>45.8</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>* BRSSW</td>
<td>45.8</td>
<td>0</td>
<td>88.2</td>
<td>142.8</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>88.2</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>142.8</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering the dog with BRSSW
After adding dog to it

<table>
<thead>
<tr>
<th></th>
<th>DBRWSS</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>0</td>
<td><strong>89.833</strong></td>
<td>143.66</td>
</tr>
<tr>
<td>*</td>
<td><strong>89.833</strong></td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>143.66</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
Clustering all the carnivores and pinnipeds

- dog
- raccoon
- seal
- sea lion
- weasel
- cat
- monkey

Distance values:
- 22.9
- 5.75
- 13
- 6.75
- 12
- 12
- 19.75
- 3.15
- 22.0166
- 44.9166
Finally, just monkey remaining

<table>
<thead>
<tr>
<th></th>
<th>DBRWSSC</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBRWSSC</td>
<td>0</td>
<td>144.2857</td>
</tr>
<tr>
<td>monkey</td>
<td>144.2857</td>
<td>0</td>
</tr>
</tbody>
</table>
UPGMA can mislead

**True tree**

A 13 14 4 10 D
2 2

**True distance matrix**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>17</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
<td>0</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>C</td>
<td>21</td>
<td>12</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>D</td>
<td>27</td>
<td>18</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>

**UPGMA tree**

B 6 6 8 10.833 A
2 2.833

Week 5: Distance methods, DNA and protein models – p.27/63
Neighbor-Joining

- For each tip, compute \( u_i = \sum_{j:j \neq i}^n D_{ij}/(n - 2) \). Note that the denominator is (deliberately) not the number of items summed.
- Choose the \( i \) and \( j \) for which \( D_{ij} - u_i - u_j \) is smallest.
- Join items \( i \) and \( j \). Compute the branch length from \( i \) to the new node (\( v_i \)) and from \( j \) to the new node (\( v_j \)) as

\[
\begin{align*}
v_i &= \frac{1}{2} D_{ij} + \frac{1}{2} (u_i - u_j) \\
v_j &= \frac{1}{2} D_{ij} + \frac{1}{2} (u_j - u_i)
\end{align*}
\]

(continued on next slide)
continued ...

- Compute the distance between the new node \((ij)\) and each of the remaining tips as

\[
D_{(ij), k} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}
\]

- Delete tips \(i\) and \(j\) from the tables and replace them by the new node, \((ij)\), which is now treated as a tip.

- If more than two nodes remain, go back to step 1. Otherwise, connect the two remaining nodes (say, \(\ell\) and \(m\)) by a branch of length \(D_{\ell m}\).
The NJ star decomposition

\[ (ij) \]
The NJ tree on the Sarich data

Same as UPGMA? No. Not clocklike.
Unweighted least squares on the Sarich data

Same as Neighbor-Joining tree? Close but not the same.
Fitch-Margoliash tree on the Sarich (1969) data

Same as the unweighted least squares tree? No.
Minimum evolution tree on the Sarich (1969) data

Same as either least squares tree? As NJ tree? No.
Kimura 2-parameter model

The simplest, most symmetrical model that has different rates for transitions than for transversions.
Parameters of K2P in terms of transition/transversion ratio

If we let $R$ be the expected ratio of transition changes to transversions, it turns out that

$$\alpha = \frac{R}{R+1}$$

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

Computing the net probabilities of change ("transition" in the stochastic process sense rather than as molecular biologists use it), it turns out that

\[
\text{Prob (transition} | t) = \frac{1}{4} - \frac{1}{2} \exp \left( -\frac{2R+1}{R+1} t \right) + \frac{1}{4} \exp \left( -\frac{2}{R+1} t \right)
\]

\[
\text{Prob (transversion} | t) = \frac{1}{2} - \frac{1}{2} \exp \left( -\frac{2}{R+1} t \right).
\]
Transition and transversion when $R = 10$
Transition and transversion when $R = 2$

![Graph showing the relationship between time (branch length) and differences for transitions, transversions, and total differences when $R = 2$.]
**ML estimates for the K2P model**

The sufficient statistics for comparison of DNA sequences under the K2P model are simply the observed fractions \( P \) and \( Q \) of transitions and transversions. Then the maximum likelihood estimates of the branch length \( \hat{t} \) and of \( \hat{R} \) are obtained by finding the values that predict exactly those fractions. That done by solving the equations given three slides ago.

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

\[
\hat{R} = \frac{-\ln (1 - 2P - Q)}{-\ln (1 - 2Q)} - \frac{1}{2}
\]
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 four screens above.

This is a function of the two parameters \( t \) and \( R \). The values that maximize it were given above, if we estimate both of them. But if \( R \) is given rather than estimated, there is no closed-form equation solving for \( t \), it has to be inferred numerically by finding the value of \( t \) that maximizes \( L \).
The Tamura/Nei model, F84, and HKY

<table>
<thead>
<tr>
<th>To:</th>
<th>A</th>
<th>G</th>
<th>C</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>$-\alpha_R \pi_G/\pi_R + \beta \pi_G$</td>
<td>$\beta \pi_C$</td>
<td>$\beta \pi_T$</td>
<td></td>
</tr>
<tr>
<td>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>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>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: \( \exp(-\alpha_R t) \)
- 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: \( \exp(-\alpha_Y t) \)
- 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 $\pi_G / \pi_Y$ (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

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

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

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

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

The condition of *reversibility* in a stochastic process is written in terms of the equilibrium frequencies of the states \( \pi_i \) and the transition probabilities as

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

The general time-reversible model:

<table>
<thead>
<tr>
<th>From:</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</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></td>
<td>( \pi_C \delta )</td>
<td>( \pi_T \varepsilon )</td>
</tr>
<tr>
<td>C</td>
<td>( \pi_A \beta )</td>
<td>( \pi_G \delta )</td>
<td></td>
<td>( \pi_T \eta )</td>
</tr>
<tr>
<td>T</td>
<td>( \pi_A \gamma )</td>
<td>( \pi_G \varepsilon )</td>
<td>( \pi_C \eta )</td>
<td></td>
</tr>
</tbody>
</table>
It is the most general model possible that still has the changes satisfy the condition of reversibility – so that one cannot tell which direction evolution has gone by examining the sequences before and after.
Standardizing the rates

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

\[ + 2\pi_G \pi_C \delta + 2\pi_G \pi_T \varepsilon + 2\pi_C \pi_T \eta = 1 \]

This is done so that the expected rate of change is 1 per unit time.
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>

The K2P model is a special case of the GTR model (as are all the other models that have been mentioned here). So if all goes well we should infer parameters that come close to specifying a K2P model.
**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

$$\hat{A}_t = \log \left( \hat{P} \right)$$

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

There are a great many other models, but these are the most commonly used. They allow for unequal expected frequencies of bases, and for inequalities of transitions and transversions.
Variation of rates of evolution across sites

The basic way all these models deal with rates of evolution is using the fact that a rate $r$ times as fast is exactly equivalent to evolving along a branch that is $r$ times as long, so that it is easy to calculate the transition probabilities with rate $r$:

$$P_{ij}(r, t) = P_{ij}(rt)$$

The likelihood for a pair of sequences averages over all rates at each site, using the density function $f(r)$ for the distribution of rates:

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

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

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

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

To get a mean of 1, set \( \beta = \frac{1}{\alpha} \) 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

Here are density functions of Gamma distributions for three different values of $\alpha$:

- $\alpha = 1/4$, $CV = 2$
- $\alpha = 1$, $CV = 1$
- $\alpha = 4$, $CV = 1/2$
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

\[ E_r \left[ e^{-b r t} \right] = \left( 1 + \frac{1}{\alpha} b t \right)^{-\alpha} \]

and just use that to replace \( e^{-bt} \) in the formulas for the transition probabilities.