Some important dates in history (billions of years ago)

- Origin of the universe: 15 ±4
- Formation of the solar system: 4.6
- First self-replicating system: 3.5 ±0.5
- Prokaryotic-eukaryotic divergence: 1.8 ±0.3
- Plant-animal divergence: 1.0
- Invertebrate-vertebrate divergence: 0.5
- Mammalian radiation beginning: 0.1

(86 CSH Doolittle et al.)
The three kingdoms

Two important early observations

- Different proteins evolve at different rates, and this seems more or less independent of the host organism, including its generation time.

- It is necessary to adjust the observed percent difference between two homologous proteins to get a distance more or less linearly related to the time since their common ancestor. (Later we offer a rational basis for doing this.)

- A striking early version of these observations is next.
Rates of macromolecular evolution

How does sequence variation arise?

- **Mutation:**
  - (a) Inherent: DNA replication errors are not always corrected.
  - (b) External: exposure to chemicals and radiation.

- **Selection:** Deleterious mutations are removed quickly. Neutral and rarely, advantageous mutations, are tolerated and stick around.

- **Fixation:** It takes time for a new variant to be established (having a stable frequency) in a population.
Modeling DNA base substitution

- Standard assumptions (sometimes weakened)
  - Site independence.
  - Site homogeneity.
  - Temporal homogeneity: stationary Markov chain.

- Strictly speaking, only applicable to regions undergoing little selection.

Some terminology

- In evolution, homology (here of proteins), means similarity due to common ancestry.

- A common mode of protein evolution is by duplication. Depending on the relations between duplication and speciation dates, we have two different types of homologous proteins. Loosely,

  - **Orthologues**: the “same” gene in different organisms; common ancestry goes back to a speciation event.
  - **Paralogues**: different genes in the same organism; common ancestry goes back to a gene duplication.

- Lateral gene transfer gives another form of homology.
Speciation vs. duplication

Beta-globins (orthologues)

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>M</td>
<td>V</td>
<td>H</td>
<td>T</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>A</td>
<td>A</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>A</td>
<td>Q</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>M</td>
<td>W</td>
<td>S</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>W</td>
<td>A</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>E</td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>D</td>
<td>L</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>P</td>
<td>D</td>
<td>A</td>
<td>V</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>P</td>
<td>K</td>
</tr>
<tr>
<td>A</td>
<td>H</td>
<td>G</td>
<td>K</td>
</tr>
<tr>
<td>K</td>
<td>L</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>F</td>
<td>L</td>
</tr>
</tbody>
</table>

- means same as reference sequence
- means deletion
Beta-globins: uncorrected pairwise distances

- DISTANCES between protein sequences (calculated over: 1 to 147)
  - Below diagonal: observed number of differences
  - Above diagonal: number of differences per 100 amino acids

<table>
<thead>
<tr>
<th></th>
<th>hum</th>
<th>mac</th>
<th>bov</th>
<th>pla</th>
<th>chi</th>
<th>sha</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum</td>
<td>----</td>
<td>5</td>
<td>16</td>
<td>23</td>
<td>31</td>
<td>65</td>
</tr>
<tr>
<td>mac</td>
<td>7</td>
<td>----</td>
<td>17</td>
<td>23</td>
<td>30</td>
<td>62</td>
</tr>
<tr>
<td>bov</td>
<td>23</td>
<td>24</td>
<td>----</td>
<td>27</td>
<td>37</td>
<td>65</td>
</tr>
<tr>
<td>pla</td>
<td>34</td>
<td>34</td>
<td>39</td>
<td>----</td>
<td>29</td>
<td>64</td>
</tr>
<tr>
<td>chi</td>
<td>45</td>
<td>44</td>
<td>52</td>
<td>42</td>
<td>----</td>
<td>61</td>
</tr>
<tr>
<td>sha</td>
<td>91</td>
<td>88</td>
<td>91</td>
<td>90</td>
<td>87</td>
<td>----</td>
</tr>
</tbody>
</table>

Beta-globins: corrected pairwise distances

- DISTANCES between protein sequences (calculated over: 1 to 147)
  - Below diagonal: observed number of differences
  - Above diagonal: number of differences per 100 amino acids
  - Correction method: Jukes-Cantor

<table>
<thead>
<tr>
<th></th>
<th>hum</th>
<th>mac</th>
<th>bov</th>
<th>pla</th>
<th>chi</th>
<th>sha</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum</td>
<td>----</td>
<td>5</td>
<td>17</td>
<td>27</td>
<td>37</td>
<td>108</td>
</tr>
<tr>
<td>mac</td>
<td>7</td>
<td>----</td>
<td>18</td>
<td>27</td>
<td>36</td>
<td>102</td>
</tr>
<tr>
<td>bov</td>
<td>23</td>
<td>24</td>
<td>----</td>
<td>32</td>
<td>46</td>
<td>110</td>
</tr>
<tr>
<td>pla</td>
<td>34</td>
<td>34</td>
<td>39</td>
<td>----</td>
<td>34</td>
<td>106</td>
</tr>
<tr>
<td>chi</td>
<td>45</td>
<td>44</td>
<td>52</td>
<td>42</td>
<td>----</td>
<td>98</td>
</tr>
<tr>
<td>sha</td>
<td>91</td>
<td>88</td>
<td>91</td>
<td>90</td>
<td>87</td>
<td>----</td>
</tr>
</tbody>
</table>
Human globins (paralogues)

<table>
<thead>
<tr>
<th></th>
<th>alpha-human</th>
<th>beta-human</th>
<th>delta-human</th>
<th>epsilon-human</th>
<th>myo-human</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>VLEPADETHVEAANGVGRAGETSGARALSMFLSLFPFTT</td>
<td>VHT.EE.SAT.L.</td>
<td>VHT.EE.A.M.L.</td>
<td>VHTAERE.AATSL.</td>
<td>VHTAERE.AATSL.</td>
</tr>
<tr>
<td>20</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td>beta-human</td>
</tr>
<tr>
<td>30</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td>delta-human</td>
</tr>
<tr>
<td>40</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td>epsilon-human</td>
</tr>
<tr>
<td>50</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td>myo-human</td>
</tr>
<tr>
<td>60</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>alpha-human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human globins: corrected pairwise distances

- **DISTANCES between protein sequences** (calculated over 1 to 141)
  - Below diagonal: observed number of differences
  - Above diagonal: estimated number of substitutions per 100 amino acids
  - Correction method: Jukes-Cantor

<table>
<thead>
<tr>
<th></th>
<th>alpha</th>
<th>beta</th>
<th>delta</th>
<th>epsil</th>
<th>gamma</th>
<th>myo</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>281</td>
</tr>
<tr>
<td>beta</td>
<td>82</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td>313</td>
</tr>
<tr>
<td>delta</td>
<td>82</td>
<td>10</td>
<td>----</td>
<td></td>
<td></td>
<td>281</td>
</tr>
<tr>
<td>epsil</td>
<td>89</td>
<td>35</td>
<td>39</td>
<td>----</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>gamma</td>
<td>85</td>
<td>39</td>
<td>42</td>
<td>29</td>
<td>----</td>
<td>313</td>
</tr>
<tr>
<td>myo</td>
<td>116</td>
<td>117</td>
<td>116</td>
<td>119</td>
<td>118</td>
<td>----</td>
</tr>
</tbody>
</table>
Correcting distances between DNA and protein sequences

- Why it is necessary to adjust observed percent differences to get a distance measure which scales linearly with time?

- This is because we can have multiple and back substitutions at a given position along a lineage.

- All of the correction methods (with names like Jukes-Cantor, 2-parameter Kimura, etc) are justified by simple probabilistic arguments involving Markov chains whose basis is worth mastering.

- The same molecular evolutionary models can be used in scoring sequence alignments.

---

Markov chain

- State space = \{A,C,G,T\}.
  
  \[ p(i,j) = \text{pr}(\text{next state } S_j | \text{ current state } S_i) \]

- Markov assumption:
  
  \[ p(\text{next state } S_j | \text{ current state } S_i \& \text{ any configuration of states before this}) = p(i,j) \]

  Only the present state, not previous states, affects the probs of moving to next states.
The multiplication rule

\[ pr(\text{state after next is } S_k \mid \text{current state is } S_i) = \sum_j pr(\text{state after next is } S_k, \text{next state is } S_j \mid \text{current state is } S_i) \]

\[ = \sum_j pr(\text{next state is } S_j \mid \text{current state is } S_i) \times pr(\text{state after next is } S_k \mid \text{current state is } S_i) \]

\[ = \sum_j p_{ij} \times p_{jk} \]

\[ = (i,k)\text{-element of } P^2, \text{ where } P=(p_{ij}). \]

More generally,

\[ pr(\text{state } t \text{ steps from now is } S_k \mid \text{current state is } S_i) = i,k \text{ element of } P^t \]

Continuous-time version

- For any \((s, t)\):
  - Let \(p_{ij}(t) = pr(S_j \text{ at time } t+s \mid S_i \text{ at time } s)\) denote the stationary (time-homogeneous) transition probabilities.
  - Let \(P(t) = (p_{ij}(t))\) denote the matrix of \(p_{ij}(t)\)'s.
    - Then for any \((t, u)\): \(P(t+u) = P(t) P(u)\).
  - It follows that \(P(t) = \exp(Qt)\), where \(Q = P'(0)\) (the derivative of \(P(t)\) at \(t = 0\)).
  - \(Q\) is called the infinitesimal matrix (transition rate matrix) of \(P(t)\), and satisfies
    \[ P'(t) = QP(t) = P(t)Q. \]
  - Important approximation: when \(t\) is small,
    \[ P(t) \approx I + Qt. \]
Interpretation of Q

- Roughly, $q_{ij}$ is the rate of transitions of $i$ to $j$, while $q_{ii} = -\sum_{j \neq i} q_{ij}$, so each row sum is 0 (Why?).
- Now we have the short-time approximation:
  \[ p_{i\rightarrow j}(t+h) = q_{ij}h + o(h) \]
  \[ p_{i\rightarrow j}(t+h) = 1 + q_{ii}h + o(h) \]
  where $p_{i\rightarrow j}(t+h)$ is the probability of transitioning from $i$ at time $t$ to $j$ at time $t+h$.
- Now consider the Chapman-Kolmogorov relation: (assuming we have a continuous-time Markov chain, and let $p_j(t) = \text{pr}(S_j \text{ at time } t)$)
  \[ p_j(t+h) = \sum_i \text{pr}(S_i \text{ at } t, S_j \text{ at } t+h) \]
  \[ = \sum_i \text{pr}(S_i \text{ at } t)\text{pr}(S_j \text{ at } t+h | S_j \text{ at } t) \]
  \[ = p_j(t) \times (1 + q_{jj}h) + \sum_{i\neq j} p_i(t) \times h q_{ij} \]
  i.e., $h^{-1}(p_j(t+h) - p_j(t)) = p_j(t)q_{jj} + \sum_{i\neq j} p_i(t)q_{ij}$, which becomes: $P' = QP$ as $h \downarrow 0$.

Probabilistic models for DNA changes

Orc: ACAGTGACGCCCCAAACGT
Elf: ACAGTGACGCTACAAACGT
Dwarf: CCTGTGACGTACAAACGA
Hobbit: CCTGTGACGTAGAAACGA
Human: CCTGTGACGTAGAAACGA
The Jukes-Cantor model (1969)

- Substitution rate:

  ![Diagram of DNA evolution](https://via.placeholder.com/150)

  The simplest symmetrical model for DNA evolution

Transition probabilities under the Jukes-Cantor model

- IID assumption:
  - All sites change independently
  - All sites have the same stochastic process working at them

- Equiprobability assumption:
  - Make up a fictional kind of event, such that when it happens the site changes to one of the 4 bases chosen at random equiprobably

- Equilibrium condition:
  - No matter how many of these fictional events occur, provided it is not zero, the chance of ending up at a particular base is \(1/4\).

- Solving differentially equation system \(P' = QP\)
Transition probabilities under the Jukes-Cantor model (cont.)

- Prob transition matrix:

\[
P(t) = \begin{pmatrix}
A & C & G & T \\
A & r(t) & s(t) & s(t) \\
C & s(t) & r(t) & s(t) \\
G & s(t) & s(t) & r(t) \\
T & s(t) & s(t) & r(t)
\end{pmatrix}
\]

Where we can derive:

\[
r(t) = \frac{1}{4} \left( 1 + 3e^{-\frac{3t}{4}} \right)
\]

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

Homework!

Jukes-Cantor (cont.)

- Fraction of sites differences

![Graph showing the fraction of sites differences over time]

Homework!
Kimura's K2P model (1980)

- Substitution rate:

\[ \begin{align*}
\alpha & \quad \beta \\
\beta & \quad \beta
\end{align*} \]

which allows for different rates of transition and transversions.

Transitions (rate $\alpha$) are much more likely than transversions (rate $\beta$).

Kimura (cont.)

- Prob transition matrix:

\[ P(t) = \begin{pmatrix}
  r(t) & s(t) & u(t) & s(t) \\
  s(t) & r(t) & s(t) & u(t) \\
  u(t) & s(t) & r(t) & s(t) \\
  s(t) & u(t) & s(t) & r(t)
\end{pmatrix} \]

Where

\[ s(t) = \frac{1}{4} (1 - e^{-4\beta t}) \]
\[ u(t) = \frac{1}{4} (1 + e^{-4\beta t} - e^{-2(\alpha + \beta) t}) \]
\[ r(t) = 1 - 2s(t) - u(t) \]

By proper choice of $\alpha$ and $\beta$ one can achieve the overall rate of change and $T_s/T_n$ ratio $R$ you want (warning: terminological tangle).
Kimura (cont.)

- Transitions, transversions expected under different $R$:

![Graph showing differences in transitions and transversions for different R values.]

Other commonly used models

- Two models that specify the equilibrium base frequencies (you provide the frequencies $A$; $C$; $G$; $T$ and they are set up to have an equilibrium which achieves them), and also let you control the transition/transversion ratio:

  - The Hasegawa-Kishino-Yano (1985) model:

<table>
<thead>
<tr>
<th>to :</th>
<th>$A$</th>
<th>$C$</th>
<th>$G$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>from :</td>
<td>$A$</td>
<td>$\alpha \pi_A + \beta \pi_A$</td>
<td>$\alpha \pi_C$</td>
<td>$\alpha \pi_T$</td>
</tr>
<tr>
<td></td>
<td>$G$</td>
<td>$-\pi_G + \beta \pi_G$</td>
<td>$\alpha \pi_C$</td>
<td>$\alpha \pi_T$</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>$\alpha \pi_A$</td>
<td>$-\pi_G$</td>
<td>$\alpha \pi_T + \beta \pi_T$</td>
</tr>
<tr>
<td></td>
<td>$T$</td>
<td>$\alpha \pi_A$</td>
<td>$\alpha \pi_G$</td>
<td>$-\pi_C + \beta \pi_C$</td>
</tr>
</tbody>
</table>
Other commonly used models

- The **F84 model** (Felsenstein)

\[
\begin{array}{c|cccc}
  \text{to :} & A & G & C & T \\
  \text{from :} & A & - & - & - \\
  & G & - & \alpha \pi_A + \beta \frac{\pi_A}{\pi_R} & - \\
  & C & - & - & - \\
  & T & - & - & - \\
\end{array}
\]

where \( \pi_R = \pi_A + \pi_G \) and \( \pi_Y = \pi_C + \pi_T \) (The equilibrium frequencies of purines and pyrimidines)

The general time-reversible model

- It maintains "detailed balance" so that the probability of starting at (say) A and ending at (say) T in evolution is the same as the probability of starting at T and ending at A:

\[
\begin{array}{c|cccc}
  & A & C & G & T \\
  A & - & \alpha \pi_C & \beta \pi_G & \gamma \pi_T \\
  C & \alpha \pi_A & - & \delta \pi_G & \epsilon \pi_T \\
  G & \beta \pi_A & \delta \pi_C & - & \nu \pi_T \\
  T & \gamma \pi_A & \epsilon \pi_C & \nu \pi_G & - \\
\end{array}
\]

- And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others).

- (Neither of these has formulas for the transition probabilities, but those can be done numerically.)
Relation between models

Adjusting evolutionary distance using base-substitution model
The Jukes-Cantor model

Consider e.g. the 2nd position in α-globin2 Alu1. 

\[ r = \frac{1 + 3e^{-4\alpha t}}{4}, \quad s = \frac{1 - e^{-4\alpha t}}{4}. \]

**Definition of PAM**

- Let \( P(t) = \exp(Qt) \). Then the \( A,G \) element of \( P(t) \) is
  
  \[ pr(G \text{ now} | A \text{ then}) = \frac{1 - e^{-4\alpha t}}{4}. \]

  - Same for all pairs of different nucleotides.
  - Overall rate of change \( k = 3\alpha t \).

- **PAM = accepted point mutation**
  - When \( k = .01 \), described as 1 PAM
  - Put \( t = .01/3\alpha = 1/300\alpha \). Then the resulting \( P = P(1/300\alpha) \) is called the PAM(1) matrix.

- Why use PAMs?
Evolutionary time, PAM

- Since sequences evolve at different rates, it is convenient to rescale time so that 1 PAM of evolutionary time corresponds to 1% expected substitutions.

- For Jukes-Cantor, $k = 3\alpha t$ is the expected number of substitutions in $[0,t]$, so is a distance. (Show this.)
  - Set $3\alpha t = 1/100$, or $t = 1/300\alpha$, so 1 PAM = 1/300\alpha years.

Distance adjustment

- For a pair of sequences, $k = 3\alpha t$ is the desired metric, but not observable. Instead, $pr(different)$ is observed. So we use a model to convert $pr(different)$ to $k$.

- This is completely analogous to the conversion of $\theta = pr(recombination)$
  
  to genetic (map) distance (= expected number of crossovers) using the Haldane map function
  
  $\theta = 1/2 \times (1 - e^{-2d})$,

assuming the no-interference (Poisson) model.
Towards Jukes-Cantor adjustment

- E.g., 2nd position in a-globin Alu 1
- Assume that the common ancestor has A, G, C or T with probability 1/4.
- Then the chance of the nt differing
  \[ p_s = \frac{3}{4} \times (1 - e^{-\frac{8}{3}t}) \]
  \[ = \frac{3}{4} \times (1 - e^{-4k/3}), \text{ since } k = 2 \times 3\alpha \]

Jukes-Cantor adjustment

- If we suppose all nucleotide positions behave identically and independently, and \( n_s \) differ out of \( n \), we can invert this, obtaining
  \[ \hat{k} = -\frac{3}{4} \times \log \left( 1 - \frac{4}{3} \frac{n_s}{n} \right) \]
- This is the corrected or adjusted fraction of differences (under this simple model). \( \times 100 \) to get PAMs
- The analogous simple model for amino acid sequences has
  \[ \hat{k} = -\frac{19}{20} \times \log \left( 1 - \frac{20}{19} \frac{n_s}{n} \right) \]
  \( \times 100 \) for PAM.
Illustration

1. Human and bovine beta-globins are aligned with no deletions at 145 out of 147 sites. They differ at 23 of these sites. Thus $n_d/n = 23/145$, and the corrected distance using the Jukes-Cantor formula is (natural logs)

$$- \frac{19}{20} \times \log(1 - \frac{20}{19} \times \frac{23}{145}) = 17.3 \times 10^{-2}.$$  

2. The human and gorilla sequences are aligned without gaps across all 300 bp, and differ at 14 sites. Thus $n_d/n = 14/300$, and the corrected distance using the Jukes-Cantor formula is

$$- \frac{3}{4} \times \log(1 - \frac{4}{3} \times \frac{14}{300}) = 4.8 \times 10^{-2}.$$  

Correspondence between observed a.a. differences and the evolutionary distance (Dayhoff et al., 1978)

<table>
<thead>
<tr>
<th>Observed Percent Difference</th>
<th>Evolutionary Distance in PAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>35</td>
<td>47</td>
</tr>
<tr>
<td>40</td>
<td>56</td>
</tr>
<tr>
<td>45</td>
<td>67</td>
</tr>
<tr>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>55</td>
<td>94</td>
</tr>
<tr>
<td>60</td>
<td>112</td>
</tr>
<tr>
<td>65</td>
<td>133</td>
</tr>
<tr>
<td>70</td>
<td>159</td>
</tr>
<tr>
<td>75</td>
<td>195</td>
</tr>
<tr>
<td>80</td>
<td>246</td>
</tr>
<tr>
<td>85</td>
<td>328</td>
</tr>
</tbody>
</table>
### Scoring matrices for alignment

<table>
<thead>
<tr>
<th>Scoring Matrices</th>
<th>CSTPAGNDEQHRKMILVFYW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D:</strong> D = +6</td>
<td></td>
</tr>
<tr>
<td><strong>D:</strong> R = -2</td>
<td></td>
</tr>
</tbody>
</table>

### How scoring matrices work

From Henikoff 1996

BLOSUM62

```
134 LQQGELDLVMTSDILPRSELHYSPMFDFEVRVLVLAPDHPLASKTQIPEDLASETLLI
137 LDSNSVIDLVMGVPNRVEVEAEAFMDNPVLVVIAPFDHPLAGERAISLARLAETFVM
```

How scoring matrices work.
Statistical motivation for alignment scores

Alignment: AGCTGATCA... AACCCTTTA...
Hypotheses: H = homologous (indep. sites, Jukes-Cantor)
R = random (indep. sites, equal freq.)

\[ pr(data | H) = pr(AA | H)pr(GA | H)pr(CC | H)... \]
\[ = (1 - p)^t p^d, \text{ where } a = \# \text{ agreements, } d = \# \text{ disagreements, } p = \frac{3}{4}(1 - e^{-8\alpha}). \]
\[ pr(data | R) = pr(AA | R)pr(GA | R)pr(CC | R)... \]
\[ = (\frac{1}{4})^t (\frac{3}{4})^d \]
\[ \Rightarrow \log \left( \frac{pr(data | H)}{pr(data | R)} \right) = a \log 1 - p\frac{1}{4} + d \log p\frac{3}{4} = a\times\sigma + d \times (-\mu). \]

- Since \( p < 3/4 \), \( \sigma = \log((1-p)/(1/4)) > 0 \), while \( -\mu = \log(p/(3/4)) < 0 \).
- Thus the alignment score \( a\times\sigma + d \times (-\mu) \), where the match score \( \sigma > 0 \), and the mismatch penalty is \( -\mu < 0 \).

Large and small evolutionary distances

- Recall that
  - \( p = (3/4)(1-e^{-8\alpha}) \),
  - \( \sigma = \log((1-p)/(1/4)) \),
  - \( -\mu = \log(p/(3/4)) \).
- Now note that if \( \alpha t = 0 \),
  - then \( p = 6\alpha t \), and \( 1-p = 1 \), and so \( \sigma = \log 4 \), while \( -\mu = \log 8\alpha t \) is large and negative.
  - That is, we see a big difference in the two values of \( \sigma \) and \( \mu \) for small distances.
- Conversely, if \( \alpha t \) is large,
  - \( p = (3/4)(1-\epsilon) \), hence \( p/(3/4) = 1 - \epsilon \), giving \( \mu = \log(1-\epsilon) = \epsilon \), while \( 1-p = (1+3\epsilon)/4 \), \( (1-p)/(1/4) = 1+3\epsilon \), and so \( \sigma = \log(1+3\epsilon) = 3\epsilon \).
  - Thus the scores are about 3 (for a match) to 1 (for a mismatch) for large distances. This makes sense, as mismatches will on average be about 3 times more frequent than matches.
- the matrix which performs best will be the matrix that reflects the evolutionary separation of the sequences being aligned.
What about multiple alignment

- Phylogenetic methods: a tree, with branch lengths, and the data at a single site.
- See next lecture for how to compute likelihood under this hypothesis

Acknowledgments

- Terry Speed: for some of the slides modified from his lectures at UC Berkeley
- Phil Green and Joe Felsenstein: for some of the slides modified from his lectures at Univ. of Washington