1 Jukes-Cantor

Jukes-Cantor model,

\[ d = \frac{3}{4} \ln(1 - \frac{4}{3}D) \]  

looks like this:

We do not have to plot this to see the behavior though, for we could simply draw the main features of such a function. First we observe that

\[ D \to \frac{4}{3}, \; d \to \infty. \]

This singularity can be seen in the graph at the value \( D = 0.75 \).

When the fraction of sites that differ between the sequences is zero \( (D = 0) \), we notice that \( d = 0 \) as well. Moreover, when difference between the sequences are small, the distance \( d \) is increasing linearly with \( D \), because at these initial times there is a low probability of several mutations per site. We can see this by the slope of the curve at small values of \( D \) (slope 1).

If \( D = 0.20 \), then \( d = 0.23 \).

2 UPGMA

\[
\begin{array}{cccc}
A & B & C & D \\
A & - & 0.2 & 0.3 & 0.4 \\
B & 0.2 & - & 0.4 & 0.4 \\
C & 0.3 & 0.4 & - & 0.5 \\
D & 0.4 & 0.4 & 0.5 & - \\
\end{array}
\]

\[
\begin{array}{ccc}
\{A,B\} & C & D \\
\{A,B\} & - & 0.35 & 0.4 \\
C & 0.35 & - & 0.5 \\
D & 0.4 & 0.5 & - \\
\end{array}
\]

The last matrix is not necessary to calculate in order to form the clades, but only to calculate the distances. But, remember that it is necessary to use the values of the original matrix, namely \((A,D)+(B,D)+(C,D)/3\).
3 A simple tree

- First the dissimilarity matrix is calculated:

<table>
<thead>
<tr>
<th></th>
<th>seq.1</th>
<th>seq.2</th>
<th>seq.3</th>
<th>seq.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq.1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>seq.2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>seq.3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>seq.4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

- The smallest distance (=1) cluster is formed \{seq.1,seq.2\}

- A new dissimilarity matrix is calculated, using the mean distance to seq.1 and seq.2:

<table>
<thead>
<tr>
<th></th>
<th>{seq.1,seq.2}</th>
<th>seq.3</th>
<th>seq.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq.3</td>
<td>2.5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>seq.4</td>
<td>3.5</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

- The next cluster is formed corresponding to the smallest distance(=2): \{seq.3,seq.4\}

- Repeat last dissimilarity matrix calculation (not needed, but interesting if you want to know the values of these distances, which are drawn into the tree below)

![Tree Diagram]

4 Tree reading

rooted: iii and iv are equal

Unrooted: all are equal. This can be seen if you draw all the trees as unrooted, i.e. in star shape.
5 Parsimony

a) As a total, we count 13 substitutions. From these, we notice that most (9) substitutions occur at the third position of the codons. See page 38 of your reader for definition of synonymous mutations. A mutation at the third position usually does not lead to a change of the amino acid, so the tendency seems to be that there exists more synonymous substitutions. However, if we carefully analyse the substitutions, we arrive at the following list of mutations:

- (1x) AAT ↔ AAC, corresponds to: N ↔ N
- (3x) GAG ↔ GAA, corresponds to: E ↔ E
- (1x) GTG ↔ GTT, corresponds to: V ↔ V
- (1x) TTT ↔ CTT, corresponds to: F ↔ L
- (2x) TTT ↔ TTA, corresponds to: F ↔ L
- (1x) TCC ↔ TTC, corresponds to: S ↔ F
- (2x) ATG ↔ ATA, corresponds to: M ↔ I
- (2x) ATG ↔ CTG, corresponds to: M ↔ L

(observe that the last four mutation occur in the last codon, and that we have considered here mutations on the consensus sequence ATG.)

This means that only 5 mutations are synonymous.

b) From the 13 substitutions, 8 are transitions and 5 are transversions. (see page 31 of your reader for definition of transitions/transversions).

c) The most informative sites are those in which at least two substitutions to the same nucleotide occur; sites 15, 27, 28, 30. (see page 82 of your reader).

d)

6 Parsimony, II

Here the informative sites are 4, 6, 7, 9 and 12.
7 IL-11

Because almost all fish species have IL11a and an IL11b, their common ancestor of fish probably had a gene duplication. Trout, carp (both missing IL-11a) and halibut (missing IL-11b) probably lost a gene, because they are the only ones of the fish that only have one of the IL-11’s. The mamalians have only one IL-11 as well, they probably did not experience the gene duplication.

8 Cyclin-dependent kinase

a) The common ancestor of eukaryotes had probably 1 CDK gene, because yeast, pombe and arabidophsis have still one CDK gene.

b) Gene duplication: Rice, common ancestor of animals, Xenopus, Human (which gave raise to CDK3). Gene loss: C. elegans, Chicken.

c) Never, because bootstrap value of Xenopus CDK1 and CDK2 clade is 100.

9 Bootstrapping, I

In all trees, D and E are found to form a clade. Also, in all cases, A, B, and C come together in clade (though in different ways). From the three possibilities of clustering A, B and C, we choose to draw the most probable one, as indicated below in the figure:
Note, that by drawing this tree with the probabilities observed, we lose some information, namely, that 19 times B was left alone while A and C formed a clade, and that 9 times A was on its own, while B and C formed a clade. The only information we can see now, is that in 72 cases A and B remained together.

10 Bootstrapping, II

a) 14. We know that 86 times W and X DID form a clade, so 14 times this did not occur.

b) No. We know from the tree that 94 times W,X,Y and Z form a branch, meaning that there is a maximal number of 6 times that W could form a clade with V.

c) YES. Because the only information that is guaranteed is that W and X come together 86 times (and that all 4 - W,X,Y,Z - form a clade in 94 cases). So, it is possible that in the other 14 cases, W, Y and Z form a clade.