Bioinformatics Assignment 3

Elliott Capek

February 24, 2016

Problem One: Phylogenetic Trees

For the following set of nucleotides, we compute the number of substitutions between each sequence and plug it into the Jukes-Cantor distance formula to estimate the phylogenetic tree for the four species:

Human	*ACGTGCTGCTAGCTGACTGATCGATCGTACGTCTAG*
Chimp	*ACGTCCTGCAAGCTGACAGATCGATCCTACGTCTAG*
Mouse	*ACGACCAGCTAGGACAGACTTGGATCCTACCTTTAC*
Rat	*TCGACCAGCTAGGAAAGACTTGCATCCTACCATTAC*

$$\text{\# of Substitutions (Hamming distance)} \qquad \begin{pmatrix} * & 4 & 15 & 18 \\ & * & 13 & 16 \\ & & * & 4 \\ & & & * \end{pmatrix}$$

We divide each Hamming distance by the length of the sequence to get the probability per site of a mutation. We use this to compute the phylogenetic distances:

$$\text{Pairwise substitution probability} \qquad \begin{pmatrix} * & 0.111 & 0.416 & 0.5 \\ & * & 0.361 & 0.444 \\ & & * & 0.111 \\ & & & * \end{pmatrix}$$

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right) \rightarrow \qquad \text{Phylogenetic Distance} \qquad \begin{pmatrix} * & 0.12 & 0.60 & 0.82 \\ & * & 0.49 & 0.67 \\ & & * & 0.12 \\ & & & * \end{pmatrix}$$

From this we can see that:
·Humans and chimps are closely related
·Mice and rats are closely related
·Chimps and mice are slightly related
·Humans and rats are very unrelated

Unfortunately there is lots of redundant information here, so it is impossible to construct an accurate unrooted tree. We know that the chimp/human pair and mouse/rat pair were the last to diverge, and probably diverged at similar times, since both pairs have the same distance (0.12). However this conflicts with the much closer distance of humans to mice than humans to rats. The best we can do is average the Chimp-mouse, chimp-rat, human-mouse and human-rat distances to get 0.645. Subtracting off the two pair distances (0.24) we get an inner node length of 0.405:

```
     0.12          0.12
Chimp ---|  0.405  |---Mouse
         ----------
Human ---|         |---Rat
```

Problem Two: Histones

Here we do multiple-sequence alignments on various h3 protein sequences and analyze the trees made by Clustalw and Phyml.

a.)

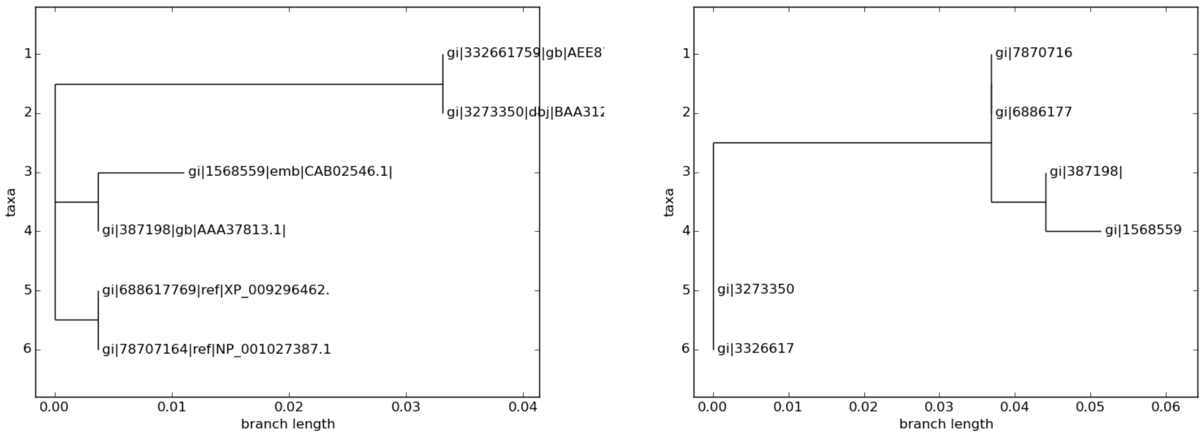

Figure 1: ClustalW tree on left, Phyml tree on right. 332-Arabidopsis, 156-human, 688-zebrafish, 387-mouse, 787-drosophila melanogaster, 327-tobacco.

Phyml Tree:

```
(((gi|7870716:0.000000,gi|6886177:0.000000):0.000000,(gi|387198|:0.000000,gi|1568559:0.007289)
:0.007228):0.036831,gi|3273350:0.000000,gi|3326617:0.000000);
```

ClustalW Tree:

```
((gi|332661759|gb|AEE87159.1|:0.00000,gi|3273350|dbj|BAA31218.1|:0.00000):0.03309,
(gi|1568559|emb|CAB02546.1|:0.00735,gi|387198|gb|AAA37813.1|:0.00000):0.00368,
(gi|688617769|ref|XP_009296462.:0.00000,gi|78707164|ref|NP_001027387.1:0.00000):0.00368);
```

The PhyML tree accurately groups the mammals together, then the other animals, then the plants. However, ClustalW does a better job of grouping each specific group together (mammals, non-mammals, plants). PhyML doesn't show flies, zebrafish being more related to mammals than each other, which seems innaccurate. So in that regard PhyML does a better job. Although PhyML doesn't seem to be generating a binary tree, whereas ClustalW does. This could be because PhyML has a minimum threshold for grouping things together, whereas ClustalW just takes the most related things and groups them without a threshold.

b.)

These two found motifs look almost identical. This suggests that there is only one main motif in the sequences, so that one-occurence-per-sequence and any-number-of-occurrences both finds the same number: one.

Problem Three: More Trees

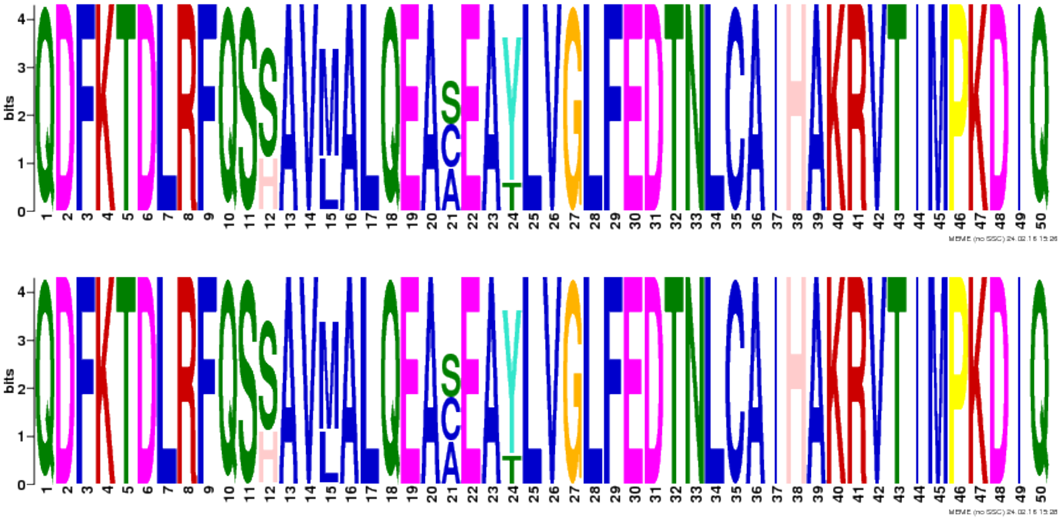

Figure 2: OOPS on left, ANR on right.

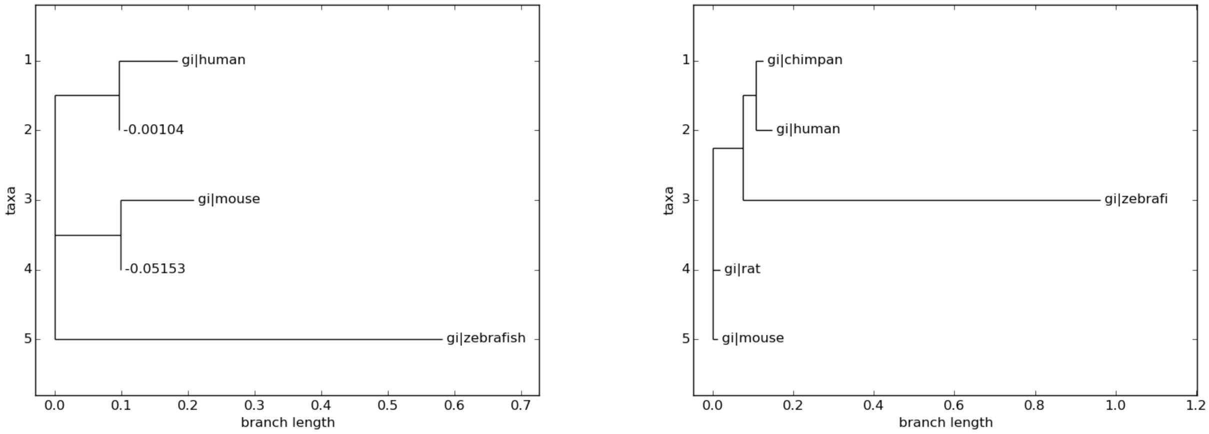

Figure 3: ClustalW on left, PhyML on right.