Course: COMPSCI 260
Name: Divya Koyyalagunta
NetID: dk160
Problem: 4
Problem Set: 5
Due: Fri 7 Nov 2016, 5pm
Using free extension (yes/no): no

Statement of collaboration and resources used (put None if you worked entirely without collaboration or resources; otherwise cite carefully): Chapter 7, Phylogenetic Trees, Richard Durbin

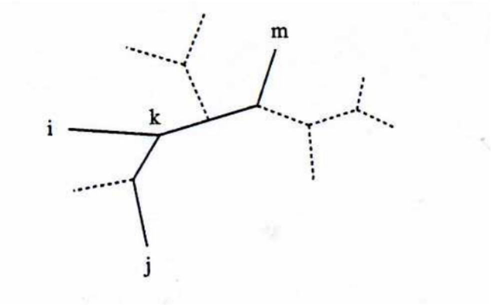
My solutions and comments for this problem are below.

PROBLEM 4

a)

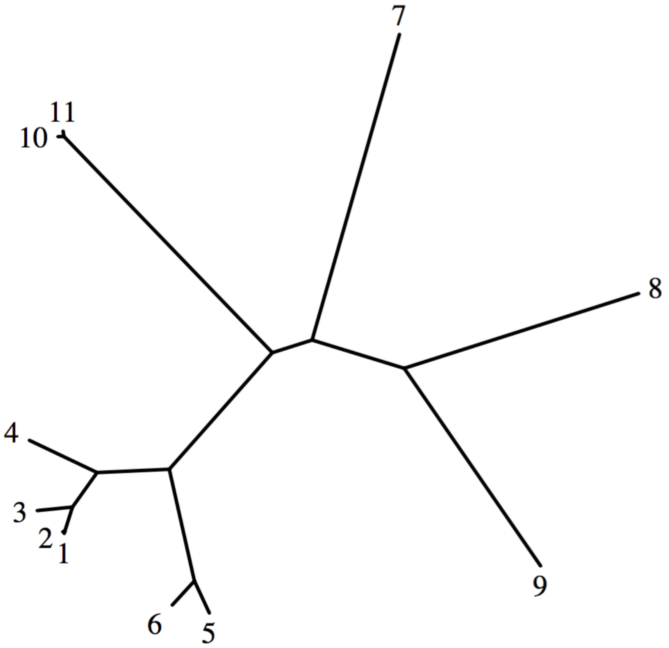
Index	Protein ID	Virus name	Animal host
1	AAL57308.1	Bovine coronavirus BCoV-LUN	cow
2	AAK83356.1	spike structural protein [Bovine coronavirus]	cow
3	AAR01015.1	S protein [Human coronavirus OC43]	human
4	AAL80031	Porcine hemagglutinating encephalomyelitis virus (HEV)	pig
5	YP_209233.1	Murine hepatitis virus strain JHM	mouse
6	NP_045300.1	E2 glycoprotein precursor [Murine hepatitis virus]	mouse
7	NP_040831.1	spike protein [Infectious bronchitis virus]	bird
8	NP_598310.1	spike protein [Porcine epidemic diarrhea virus]	pig
9	BAA06805	peplomer protein [Feline infectious peritonitis virus]	cat
10	AAP41037.1	spike glycoprotein [SARS coronavirus Tor2]	human
11	AAV49723	spike glycoprotein [SARS coronavirus PC4-241]	human, palm civet

c) The distances computed in part (b) are not ultrametric but they are additive. This means we cannot apply the UPGMA algorithm but we can apply the NJ algorithm. UPGMA assumes a constant rate of evolution across lineages (i.e. it only works under the assumption that the molecular clock hypothesis is valid). UPGMA constructs a rooted tree where the edge length can be viewed as times. In order for UPGMA to work, closest leaves must be neighboring leaves with a common parent node, but without the ultrametricity condition, an incorrect tree will be constructed. The additivity property is built in as the UPGMA tree is constructed, so it is possible for the ultrametricity property to fail but additivity to hold, in which case the NJ algorithm can be used.



from the phylogenetic tree above, we can see that for any three leave i, j, m, there is a node k such that the branches to them meet. and by additivity, we can see clearly that $d_{im} = d_{ik} + d_{km}, d_{jm} = d_{jk} + d_{km}, d_{ij} = d_{ik} + d_{jk}$. The formulas we use to build the NJ algorithm follow from here.

e) The following is the tree from the Phylip drawtree visualization:



First, we can see that certain clusters of nodes, for example 1/ 2, 10/11 and 5/6 are in very close proximity with each other, which is consistent with the animal hosts for each member of those clusters being the same. It also follows that branches 7, 8 and 9 (from the bird, pig and cat) are more distant because they are from distinct animal species.

Because NJ trees are unrooted, we have to make inferences about how the tree fits together. The two longest branches come from the node that connects {10,11} to the middle of the tree, and the node that connects 7 to the middle of the tree. These long branches indicate the possibility of these two clusters being outgroups, and for the root to be somewhere along these long branches. If we consider the possibility of the root being along the branch connecting 7, then there is a likelihood that the SARS virus arose in some type of birds, which then took different forms in multiple animal hosts. If we consider the possibility of the root being along the branch connecting {10,11}, then we can see that there is a possibility that the SARS virus arose in the palm civet (virus ID 11 has been found in humans and palm civet). This suggests that the SARS virus arose in the palm civet, and the same spike glycoprotein [SARS coronavirus PC4-241] "made the jump" into humans, as the same virus has been seen in humans. This also aligns with how virus ID 10, also seen in humans, is very close on the NJ tree, meaning there was some mutation of this virus form as it spread amongst humans, after initially being picked up from the palm civet. More mutations must have arisen as the SARS virus spread into multiple animal hosts, and the relative degree of mutation of the SARS virus seems to be greater between different animal species, and smaller between similar animal species.