Homework Assignment Number 5

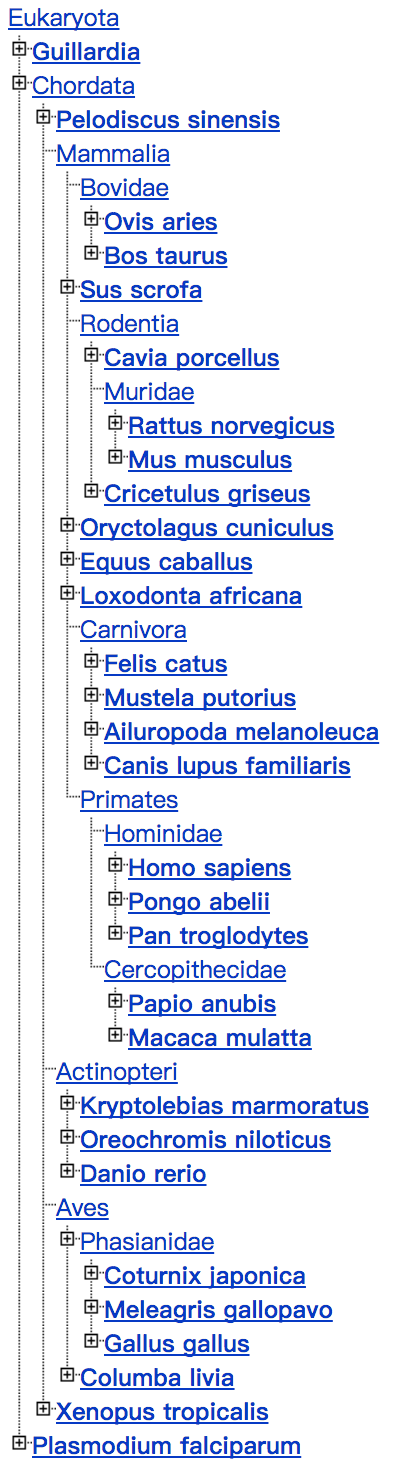
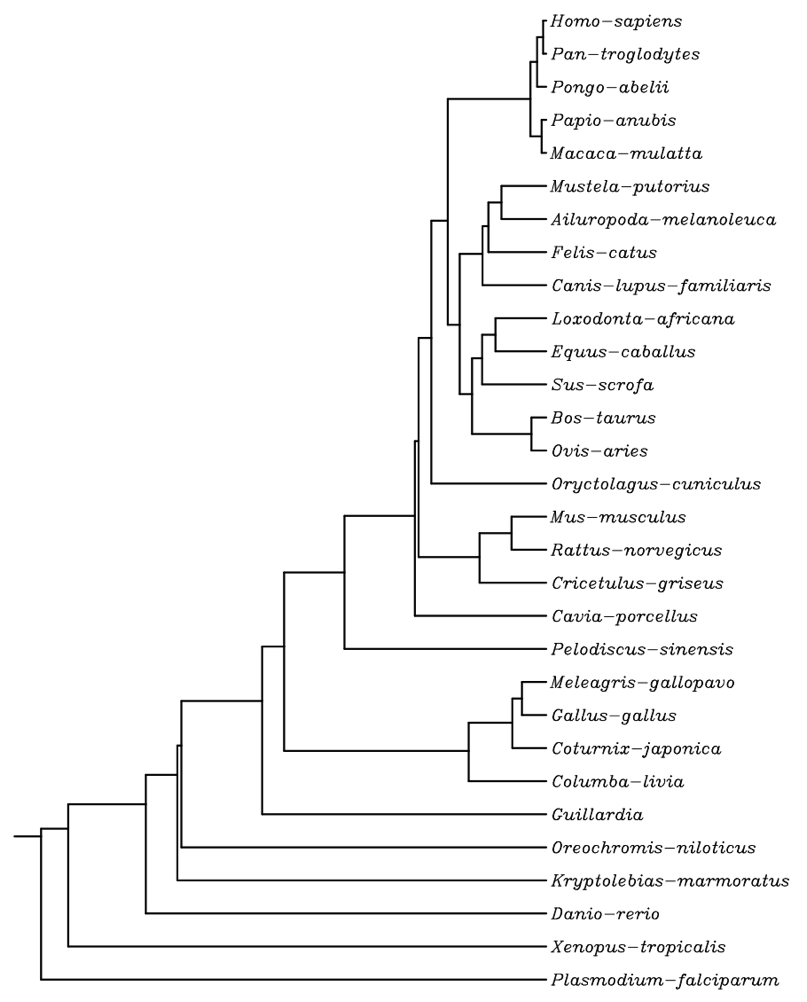
For this homework assignment take 20 to 30 protein sequences which are at least 30% identical or better and:

1) Prepare your data. Please edit your sequence data so that each has a meaningful sequence name. For example remove the "sp|gi|P31415626|QNXY\_CJAPON", and replace it with the common name (e.g. RICE).

2) Make a multiple sequence alignment with them using EBI's ClustalW2. The sequences must be approximately the same length. Then:

3) Following the "Important notes" below, make two phylogenies, one using the UPGMA method and the other using the Neighbor Joining method. On the EBI ClustalW2 page, under Step 3, "Set your multiple sequence alignment options", look for the "Clustering" dropdown. There are two choices, NJ and UPGMA. The tree is presented at the bottom of the "Guide tree" tab on the results page.

4) Describe the resulting alignments and include graphic images of the alignments and phylogenies.



I found the protein of GAPDH in 30 species.

The right figure is the tree which base on the protein of GAPDH in 30 species.

The left figure forms the NCBI Taxonomy Page.

5) Mention if the trees seem reasonable biologically or taxonomically reasonable by comparison with standard taxonomies (NCBI Taxonomy Page).

By comparison with standard taxonomies in NCBI Taxonomy Page, the trees seem biologically and taxonomically reasonable.

(6) Do the two trees have the same topology? See important notes, below!

The two trees have almost the same topology, and the different is small.

(7) Do the trees have the same branch lengths?

The two trees have different branch lengths, but the different is small, just like their topology.

(8) If the two trees do not have the same topology or branch lengths, describe the differences and indicate why you think the two trees differ.

The differences are the location of some similar species. The evolutionary relationship of the gene GAPDH between these 30 species may not be the same with the evolutionary relationship of the whole genome between these 30 species totally. The evolutionary relationship of the whole genome shows the evolutionary relationship of many gene in species, so it may not be able to represent the evolutionary relationship of a given gene in species. Also, if some species have the same gene with highly similar sequences. In the process of independent evolution, the happen of some small gene mutations in this gene are possible, so the evolutionary relationships of the gene in these species may not represent the evolutionary relationships of the whole genome of these species.

(9) Are the differences taxonomically or biologically significant?

The differences only appear between similar species, so for the evolutionary relationship of these 30 species, these differences are not taxonomically or biologically significant.

(10) Do the trees show evidence of paralogous evolution?

The two trees both show the evidence of paralogous evolution, it is shown in these trees that the similar species have the same ancient ancestor, so they are paralogous.

(11) Which nodes are orthologous and which are paralogous bifurcations?

All the 30 species are different species and the number of the gene copy is unknown, so all of the nodes contact to the species above are orthologous, but I cannot to judge that whether the rest nodes are orthologous or paralogous bifurcations.

(12) Do the trees show evidence of either gene conversion or horizontal gene transfer?

I think that these two trees do not show obvious evidence of either gene conversion or horizontal gene transfer. If we want to research the gene conversion or horizontal gene transfer, more information about the DNA sequences is needed.

The best way to find 20 to 30 protein sequences which are 30% identical or better and an out group is to look at the results of your previous homework in which you performed a database search. Go down the list of similar proteins to either the top 20 to 30 sequences with > 30% identity and use those sequences to make your alignment. You may also include one more sequence further down the list to serve as the outgroup for your phylogenic trees. If you find less than 20 sequences with > 30% identity, then choose another sequence for this study.

Please send this assignment to he.jk@sustc.edu.cn or place it in your directory on the server by the due date. Please include sufficient output from your analyses including graphic files portraying your two trees to support each of your answers/conclusions.

Important notes:

A) Remember that UPGMA trees are rooted and Neighbor-Joining trees are UNrooted

B) Include your sequences in FASTA format (unaligned or aligned).

C) Rotation around internal branches does NOT change the topology of the tree. Don't focus on the tips when answering this question.