

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|QNX_Y_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.
- (5) Mention if the trees seem reasonable biologically or taxonomically reasonable by comparison with standard taxonomies (NCBI Taxonomy Page).
- (6) Do the two trees have the same topology? See important notes, below!
- (7) Do the trees have the same branch lengths?
- (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.
- (9) Are the differences taxonomically or biologically significant?
- (10) Do the trees show evidence of paralogous evolution?
- (11) Which nodes are orthologous and which are paralogous bifurcations?
- (12) Do the trees show evidence of either gene conversion or horizontal gene transfer?

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 phylogenetic 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.