

Introduction to Bioinformatics
Problem Set 8: Phylogenetic Analysis

1. Draw a phylogenetic tree of sequences evolving by your own hand. Define a parent sequence composed of random DNA 40 nucleotides long. Then define child1, by altering one of the nucleotides in the parent sequence. Then make child2 by inserting one or more nucleotides somewhere in the parent sequence and child3 by making the same modification to child 2.
 - 1a. Align the four sequences, using ALIGNMENT-OF, and make a tree of them, using TREE-OF
 - 1b. Do the same using the NO-GAPPED-COLUMNS option for ALIGNMENT-OF
 - 1c. Do the same, varying the size of the insertions (but keeping them the same for both child2 and child3)
2. Define a function called MUTATE that will mutate a sequence at a random position (SUBSTRING will be useful in this regard). Define a parent sequence composed of random DNA 40 nucleotides long and mutate it to make child1. Mutate child1 and its children and go back and mutate the parent again, and so forth, generating sequences whose relationships to each other are known (because you created them). Be sure to name them in a way that reminds you of the relationship! (The LABELED option in DEFINE will be useful). Align all the parent, children and grandchildren and from a visual inspection of the alignment, figure out a phylogenetic tree. Check your result using TREE-OF and compare it with the known relationship.
3. Are phylogenetic trees of individual proteins indicative of the relationships of whole organisms? Answer that question within CyanoBIKE. Your strategy will be to compare the trees derived from a comparison of two sets of orthologs (proteins of common evolutionary origin).
 - 2a. DEFINE core-proteins as the COMMON-ORTHOLOGS-OF *all-cyanobacteria*. You can get COMMON-ORTHOLOGS-OF from the Genome menu and *all-cyanobacteria* from the Data menu, Organism Subsets submenu.
 - 2b. DEFINE protein1. To do this, CHOOSE-FROM core-proteins a random protein from the list. DEFINE protein2 in a similar fashion. What are the descriptions of these proteins?
 - 2c. DEFINE a set, protein1-orthologs, comprising the orthologs of protein1, using the ORTHOLOG-OF function, found on the Genes-Proteins menu. DEFINE another set, protein2-orthologs, using protein2.
 - 2d. Use the TREE-OF function to create phylogenetic trees of both sets of orthologs. TREE-OF can be found on the Strings-Sequences/Phylogenetic-Tree menu. It asks for an alignment, which you can provide using the ALIGNMENT-OF function (which you are already familiar with). For *tree-project* provide a string that will serve as both a file name and a reminder to you of what the tree is about.
 - 2e. Compare the two trees. Are they similar in structure to each other?
 - 2f. You would probably find useful a list that tells you what organism is associated with what gene prefix. To do this, execute a loop or a map that lists for each organism its prefix (NAME-OF organism SHORT) and its full name. Sort the list alphabetically before displaying it (DISPLAY-LIST EACH ...).
4. Create a phylogenetic tree of some protein encoded by the phage you are annotating.