

Introduction to Bioinformatics
Problem Set 7: 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, INSERTing a mutation at a random location within the DNA. Define child2 in the same way. Then define child1a and child1b by mutating child1 and define child2a and child2b by mutating child2. To do this, define a function called MUTATE that will take any sequence and put a mutation at a random location. 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.
2. 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 ...).
3. Create a phylogenetic tree of some protein encoded by your metagenome contig. If you haven't extended your read into a contig, now is the time to do that. If you can't find a gene in your contig, maybe it's time to get a new read. If the gene does not encode a protein that is similar to any known protein, there's another reason to get a new read.