Assignment 4, BME501/CPS501

Introduction

This assignment is based on the Chapter Project for Chapter 5 of the course text, by St. Clair and Visick.

The basic purpose of the assignment is to look for molecular evidence that let's one decide which of several land mammals is most closely related to the marine mammals, the whale and the porpoise. Phylogenetic trees are based on alignment of sequences of a shared gene which is orthologous across the species in question. We use web-based software called phyogeny.fr. The multiple sequence alignment on which the analysis is based is called MUSCLE, which is similar to Clustal W which we described in class. The distance metric that the software uses by default is called HKY85, the Hasegawa-Kishino-Yano model, which is more complex than the three metrics we covered in class, since their model allows for variable base frequencies.

Background

Around 1750 Linnaeus set up a classification system which we still use today: phylum, class, order, family, genus, species, whereby plants and animals are divided into various groups according to their relatedness. About 100 years later, Darwin in 1859 proposed that new species arise from common ancestors as natural selection acts on individual variation. The nearest in time common ancestor of two species is called the "most recent common ancestor", or MRCA. Proposing MRCAs is a way of estimating what happened in the past. We assume that all mammals have a common ancestor, but we might like to know more recent common ancestors of certain species so that we can say which species are more closely related. For example, is a whale more closely related to a land mammal that is a carnivore or one that is a herbivore? When scientists tried to answer this question based on morphology or fossil evidence, the answer was unclear. In this project we use the differences in the sequences of a gene to answer the question. Such a gene is called a molecular clock, since elapsed time from divergence in the species to now is assumed to be proportional to the number of substitutions. The more changes in the gene between two species, the further back in time is their MRCA. We choose the gene CSN2, beta-casein, which is the major protein in milk. All mammals nurse their young, so all mammals have that gene.

Steps

1. Go to NCBI Gene, and search for CSN2. You will see on the lower half of the screen about 20 of 416 search results, but ignore that for now, and click on "**CSN2 orthologs from mammals**" on the top half of the screen. Consider the first animal at the top of the list "Bos taurus" **cattle**. Click on the down-arrow at the right. The drop-down box gives you some choices for RefSeq transcripts and RefSeq proteins. We want to compare nucleotide sequences to make the tree, so choose the top RefSeq transcript (**XM_010806178.3**). Click on that accession number, which takes you to the record for it, with length 1094 bp. Click on "**Send to**", "Coding Sequences", FASTA Nucleotide, Create File. Rename the downloaded sequence file to "**cattle.txt**". Change the comment line to something short and clear, like "> cattle XM_010806178". Similarly, get the coding sequence (CDS) for the following CSN2 genes: **house mouse**, **dog**, **whale**, and **porpoise**. (There is more than one kind of whale, so choose whichever whale you want, and similarly for the other animals). Do similar downloads and name changes for the other CSN2 genes.

For **Step 1** show the names and accession numbers of the mammals you chose.

2. Put all of these FASTA sequences, comments and all, into one file, "**all.txt**". Go to a site called **www.phylogeny.fr**. Click on the tab "Phylogeny Analysis" and choose "**One-click**". Upload the file,

all.txt, and start the analysis with all of the default parameters. You will see the six steps occurring: overview, data and settings, alignment, curation, phylogeny, tree rendering. When these steps complete, you can see the phylogenetic tree. Under Display:, choose "**Branch Lengths**". Look back at the tree and you will see now the branch lengths on the tree.

(a) Download the resulting tree as PNG or PDF or other format, so that you can **include the tree image in your report**. Notice that the porpoise and whale are more closely to each other than to any of the land mammals.

(b) What are the these marine mammals most closely related to: dog (carnivore), mouse (rodent) or cattle (herbivore)?

(c) What is the total distance from whale to dog, to an accuracy of two decimal places? (Add up the lengths of edges connecting the two species).

For Step 2, show answers for (a), (b), (c).

3. The default tree shown in step 2 is a phylogram, which means that the branch lengths are proportional to evolutionary distance. Click on "**radial**" tree style.
For Step 3 show the radial tree.

4. Make a new phylogram for **whale**, **cattle**, **hippopotamus**, **human** and **X**, where X is a mammal unique to your group.
For Step 4, (a) show the phylogram image with distances, and say (b) which of these land mammals is closest to the whale?

5. Click on the tab **alignments**, to see the result of multiple sequence alignment of the five sequences. The gaps need to be removed, and maybe some ambiguities at the edges of gaps. Click on the tab **curation**, where you can see the result of a program called Gblocks that has underlined in blue the parts of the alignment that can be used for distance measurement. Click on "**Cured alignment in FASTA format**", under Outputs. You will then see the gapless multiple sequence alignment, where all of the sequences have the same length and there are no gaps. **Download** or copy paste that file to a local file on your computer.
For Step 5, show this gapless cured alignment.

6. Write a **Python** program which compares each pair of sequences, counting the differences in the straightforward way, where the sequences are compared letter by letter from start to end. Call that count d, and from that value calculate D = d/n, where n is the length of the sequence (they are all the same length). Write a function **jukes**(D) to calculate K from D, and then calculate k = K * n. Let k be our distance measure, and make a 5 x 5 **matrix** of these distances k.
For Step 6, show your matrix, with the mammal names (abbreviated) across the top and down the left side.

7. By hand use **single linkage** to make a tree from the distance matrix.
For Step 7, show the tree.

8. Compare the tree you made by hand with the one phylogeny.fr made.
For Step 8, say what are the **differences** and what are the **similarities**?

9. For Step 9, submit your Python file as a separate **a4.py** file on D2L. (It has to be a .py file, and not buried in a PDF. There should be comments to show what you are doing. Step 9 is worth 50% of your mark).

10. Submit your report on D2L as a file **assignment4.pdf** on D2L.
Marking: one point each for Steps 1 to 8, and 8 points for Step 9. Total = 16.