Sequence Analysis -- Fall 2018

Homework 4. Due Oct 25

**MSA-to-Newick**

Write a program that reads a multiple sequence alignment in FASTA format and outputs a UPGMA tree in Newick format. Your internal data structures are a NxN matrix of distances (D), a N vector of strings (taxon_string) and a N vector of ancestral times (t). (Each split has a t).

Proceed to make a neighbor-joining tree, calculating branch lengths and ancestral times (t) as you go. First select the shortest distance. Calculate branch lengths. Then cross out one column and row and replace the other with the average of the remaining distances minus the branch lengths that you just assigned (if any).

To calculate UPGMA distances, take two nearest neighbors. Add any existing ancestral times to this distance and divide by two. Then subtract the ancestral times (if any) from each branch. Those are the two output branch lengths. Do this for each neighbor joining step. The last step defines the root of the tree.

====== START PSEUDOCODE ======

Read MSA

Calculate distance matrix D. Use Juke-Cantor distances based on %identities.
\[ x = -\log((pid-ranpid)/(1-ranpid)) \], where pid is the fraction identical and ranpid is 0.25

Read the taxon_strings from the FASTA "->" lines.

Initialize the ancestral time t(1..ntaxa) for each taxon to 0.0 because the current taxa are in the present.

Initialize number_of_columns to number of taxa

Do while number_of_columns > 1

    Find shortest distance A,B

    Calculate UPGMA branch length to A: \[ b(A) = 0.5 \times ( D(A,B) + t(A) + t(B) ) - t(A) \]
    Calculate UPGMA branch length to B: \[ b(B) = 0.5 \times ( D(A,B) + t(A) + t(B) ) - t(B) \]

    New taxon_string for col A. "(labelA:b(A),labelB:b(B))"
    New taxon_string for col B: "delete"
Update \( t(A) \) by adding \( b(A) \)

For all columns \( X \), if taxon string is not "delete", then
\[
\text{new } D(A,X) = \text{AVG( old } D(A,X), D(B,X) ) - b(A) .
\]
\[
\text{new } D(B,X) = 0.0 \quad \text{(ignored)}
\]

Reduce \text{number_of_columns} by one.

end while

Write out the one and only taxon_string that does not say "delete". If there is more than one, you have a bug.

======== END PSEUDOCODE ========

Check your program. Does it ever create negative distances. It is possible. So, set any negative distance to zero.

Try your program on the sequences posted under "test case". The answer should be as pasted below.

Check to see that your Newick string is accepted by UGENE if you open it "as Newick".

Submit your program and your answer as a plain text file with the Newick string.
ANSWER for test.msa: showing MSA, %ID matrix, Newick (note the use of ‘single quotes’. Spaces are preserved in UGENE if you use single quotes.), and the UGENE phylogram.

(\textbf{A}:0.097, (\textbf{E}:0.036, \textbf{D}:0.036):0.061):0.449, ((\textbf{C}:0.041, \textbf{B}:0.041):0.287, \textbf{A}:0.328):0.218);