**Homework 5**

**Problem 1 (30pts)**

Derive weights for sequences

ACTA

ACTT

CGTT

AGAT

using Thompson, Higgins, and Gibson method

Use the outline below (a-d) to solve this problem

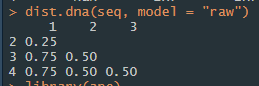
a) compute pairwise distances between sequences

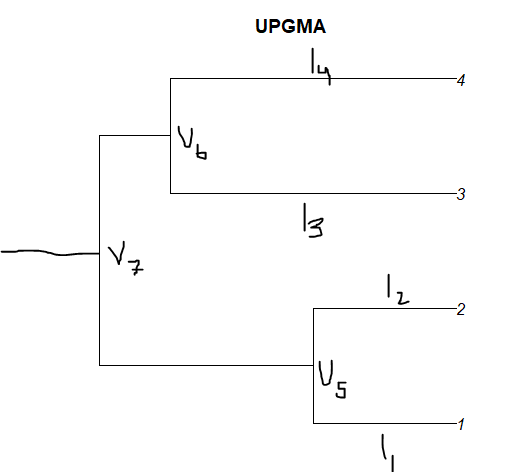
b) apply *UPGMA* method to join sequences and consequently the clusters)

c) build phylogenetic tree

d) derive sequence weights

See attached R script.





V5 = 2I1 = 2I2 = 0.25

* I1 = I2 = 0.25/2 = 0.125

Dist(1,2),3 = (dist1,3 + dist2,3) / 2 = 0.75 + 0.5 / 2 = 0.625

Dist(1,2),4 = (dist1,4 + dist2,4) / 2 = 0.75 + 0.5 / 2 = 0.625

|  |  |  |
| --- | --- | --- |
|  | 1,2 | 3 |
| 3 | 0.625 |  |
| 4 | 0.625 | 0.5 |

V6 = 2I3 = 2I4 = 0.5

* I3 = I4 = 0.5/2 = 0.25

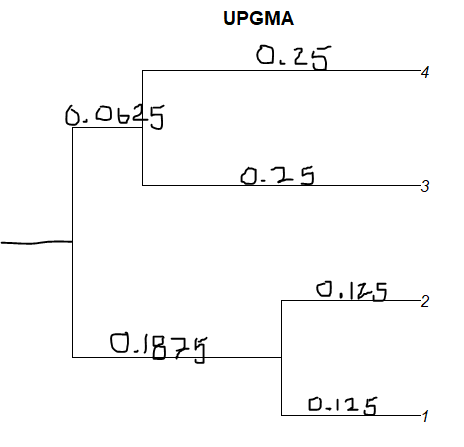
Dist(1,2),(3,4) = dist3,(1,2) + dist4,(1,2) / 2 = 0.625 + 0.625 / 2 = 0.625

|  |  |
| --- | --- |
|  | 1,2 |
| 3,4 | 0.625 |

V7 = (1,2),(3,4) / 2 = 0.625 / 2 = 0.3125

V7 – V6 = 0.3125 – 0.25 = 0.0625

V7 – V5 = 0.3125-0.125 = 0.1875



**Problem 2 (10pts)**

We assumed *additive* property when constructed UPGMA tree in problem 1.

What is limitation of this assumption (if any)?

In order to assume the additive property and still yield the correct tree, the sequences must have equal evolutionary rates. In addition, the data must be ultrametric, in which, for any three taxa, two must be equal and greater than the third. Should these requirements not be the case, this method could result in a tree that is incorrect.

**Problem 3 (20pts)**

The protein sequence of bacterial species “B3” was used to blast against swissprot protein database. The query returned significant hits to four other **bacterial** proteins (*B1,B2,B4, B5*), and one protein in **human** genome (*H*). No other mammalian species have shown presence of protein that is similar to B3. Phylogenetic tree construction by several methods resulted in a tree shown below. Explain the presence of this gene in humans.



Based on the tree above, an ancestor that eventually led to the protein in the human genome diverged from the ancestor to B2, B3, B4, and B5 with some evolutionary change. Because the B3 protein and human protein share a common ancestor, there are likely similar sequences in both genes that led to this tree. One possible explanation may be that a bacteria arising from the same ancestor as B1 eventually led to a bacterial species that infected early humans. During infection, this bacteria might have horizontally transferred a part of its genome to the human mitochondria, resulting in changes to the human genome that result in this protein. As a result, the human genome expresses a protein that evolves from a common ancestor with B3 and the two proteins share similarities that make one sequence a significant hit when the other is queried with blast.

**Problem 4 (10pts)**

Describe technical and theoretical challenges associated with building phylogenetic trees.

One common challenge when constructing a phylogenetic tree is picking the most appropriate method. Choosing a poor method for the sequences you have could mean constructing the wrong tree. As described in question 2, UPGMA demonstrates this challenge by its suitability for sequences that do not have equal evolutionary rates. In terms of the tree’s construction, if the routes to the leaves are different lengths on the tree, then the tree will not be constructed correctly, because the assumption is that these lengths are equal. In addition, phylogenetic trees have some limitations. For instance, length of time cannot be determined by the branches. Instead, the branches only show the evolutionary order. There is also a lot of information left out of phylogenetic trees because there is no telling how many branches led to a particular taxa after a divergence without analyzing all species with the tree.

**Problem 5 (10pts)**

Compare and contrast *parsimony, maximum likelihood, UPGMA,* and *neighbor-joining* methods

In the parsimony method, the phylogenetic tree is constructed such that the fewest necessary changes are used, starting at the root and working toward the tips. Thus the sequences are arranged on the tree such that as few evolutionary changes as possible are made to explain the arrangement. In the maximum likelihood method, all possible trees are considered, and the probability of the sequence is determined given each tree. The final result is provided based on the highest total probability of the sequences given that tree model. In the UPGMA method, you start with the individual sequences and then cluster them based on their closeness. At each step, more clusters are combined and their distance to the next cluster is used to determine which clusters to combine until the root of the tree is reached. The neighbor-joining method is similar to UPGMA, in which the tree is built starting with the two nearest taxa, then in considering the next two nearest taxa, the node of the first two is considered as one. However, the neighbor-joining method uses a slightly different algorithm that allows for the branch lengths to be different, a pitfall that occurs with UPGMA, as described in question 4.

**Problem 6 (20)**

Create multiple sequence alignment and phylogenetic tree **in R** using ape and clustalw by following steps below:

1. Install clutalw (depending you your OS) on your computer using <http://www.clustal.org/clustal2/> link
2. Open R. (all of the following steps will be implemented in R)
3. Set a working directory
4. Install package “ape” from your R session by typing:

intall.packages("ape ")

1. Load “ape” package by typing

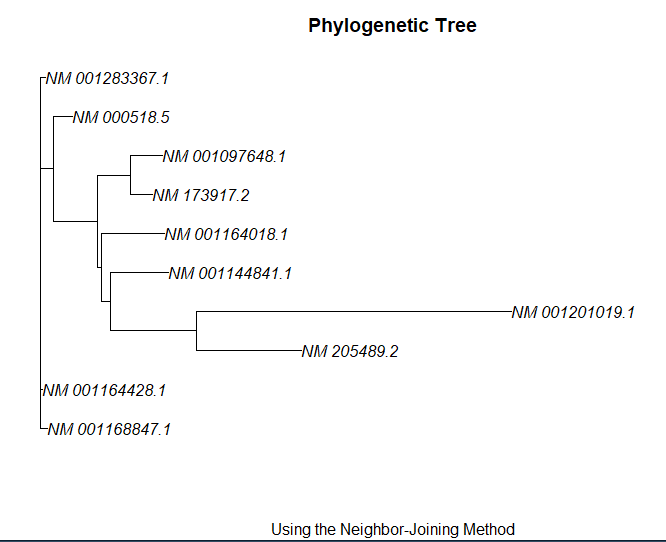
library("ape ")

1. Read accession numbers of sequences you downloaded for Homework 2 from GenBank; this step rather for exercising purposes since you have already downloaded these sequences.
2. Save the result from step 6 as <new.fas> file
3. Run clustalw by typing:

system(paste('"path\_to\_YOUR\_clustalw/clustalw2.exe" new.fas'))

1. Read alignment file (\*aln) it should be in your working directory
2. Create phylogenetic tree using neighbor-joining method
3. Plot the tree

Copied from R console:



**See attached R Script for problems 1 and 6**