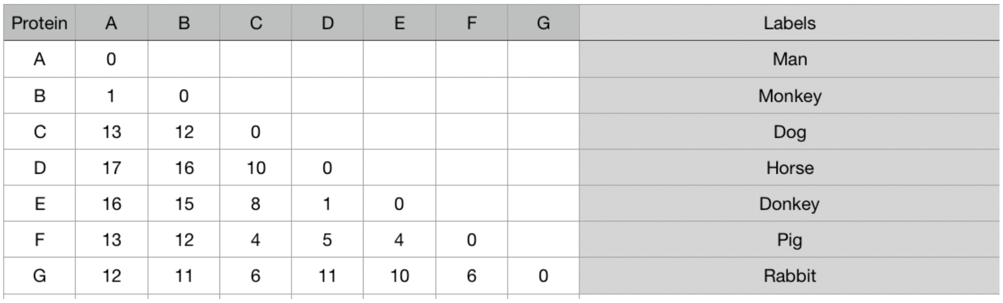
CMPE 549: Bioinformatics, Fall 2019
Assignment 2 - CALCULATION OF PHYLOGENIES: The UPGMA Method
Due: 24/12/2019, 23:55

In this assignment you are required to implement the Unweighted Pair Group Method using Arithmetic Averages (UPGMA) phylogenetic tree building algorithm.

TASKS:

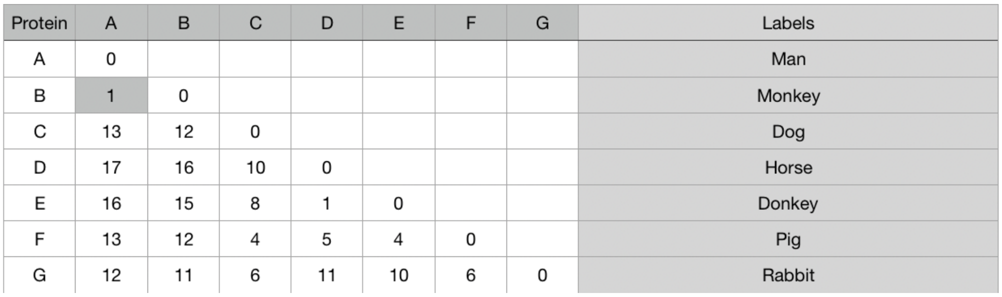
1. UPGMA builds a tree for a collection of sequences based on the *distances* between pairs of sequences. In this assignment, the distances are already provided. **Your first task is to write a function which reads in such pairwise distances and their labels from text files *distance_matrix.txt* and *labels.txt*.** The details of the *distance_matrix.txt* file are presented in Figure 1. It includes some typical *cytochrome c* comparisons (from Fitch and Margoliash, Science Vol. 155, 20 Jan. 1967). The numbers in the cells show mutation distances between the *cytochrome c* molecules of various species. For instance, the mutation distance between the amino acid sequences of donkey and horse is 1. The *labels.txt* file consists of the labels of corresponding species as a list. You can use the actual labels or a naming similar to Figure 1 (please provide the mapping).



Protein	A	B	C	D	E	F	G	Labels
A	0							Man
B	1	0						Monkey
C	13	12	0					Dog
D	17	16	10	0				Horse
E	16	15	8	1	0			Donkey
F	13	12	4	5	4	0		Pig
G	12	11	6	11	10	6	0	Rabbit

Figure 1: Selected *Cytochrome C* comparisons

2. Your second task is to locate the smallest cell in the table and get its coordinates. For instance, the smallest distance's coordinates are (1,0) which belongs to monkey (B) and man (A).

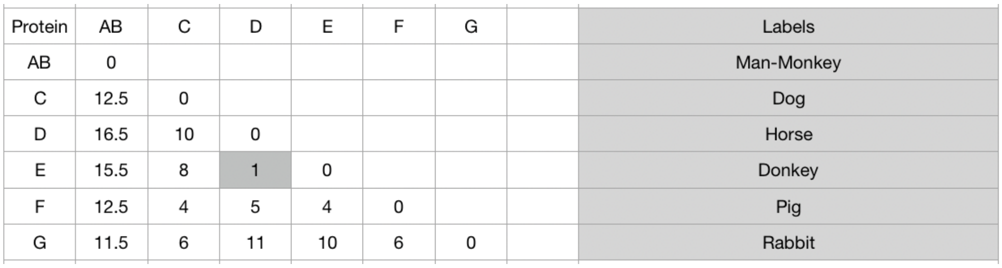


Protein	A	B	C	D	E	F	G	Labels
A	0							Man
B	1	0						Monkey
C	13	12	0					Dog
D	17	16	10	0				Horse
E	16	15	8	1	0			Donkey
F	13	12	4	5	4	0		Pig
G	12	11	6	11	10	6	0	Rabbit

Figure 2: Select the smallest cell: Step 1

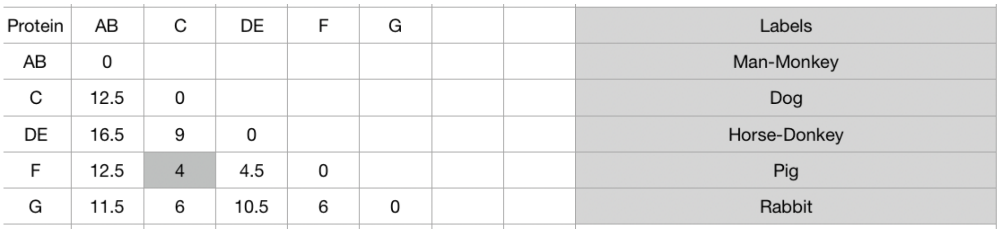
3. Your third task is to join the entries of the table on the smallest cell, and average the corresponding data entries. For instance, we first join man (A) and monkey (B) and update the cells dog (C) and man-monkey (AB) by averaging the values of C-A and C-B which are 13 and 12, respectively. Finally, update the cell C-AB as 12.5.

When you perform Task 2 and Task 3 for the given distances, the intermediary tables will be similar to the figures listed below:



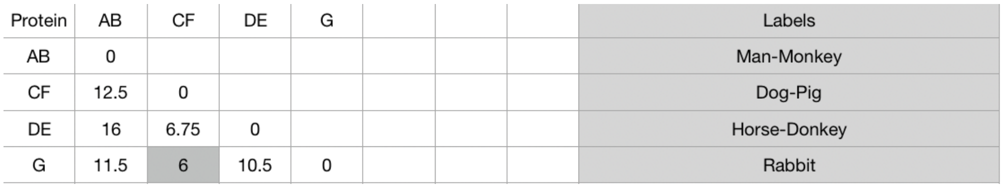
Protein	AB	C	D	E	F	G		Labels
AB	0							Man-Monkey
C	12.5	0						Dog
D	16.5	10	0					Horse
E	15.5	8	1	0				Donkey
F	12.5	4	5	4	0			Pig
G	11.5	6	11	10	6	0		Rabbit

Figure 3: Join cells: Step 2



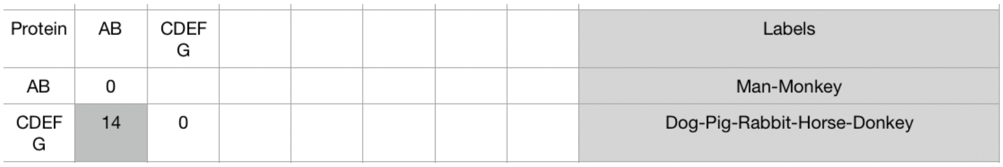
Protein	AB	C	DE	F	G			Labels
AB	0							Man-Monkey
C	12.5	0						Dog
DE	16.5	9	0					Horse-Donkey
F	12.5	4	4.5	0				Pig
G	11.5	6	10.5	6	0			Rabbit

Figure 4: Join cells: Step 3



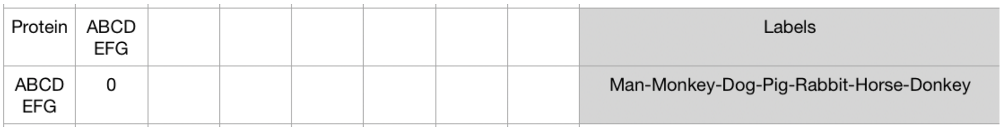
Protein	AB	CF	DE	G				Labels
AB	0							Man-Monkey
CF	12.5	0						Dog-Pig
DE	16	6.75	0					Horse-Donkey
G	11.5	6	10.5	0				Rabbit

Figure 5: Join cells: Step 4



Protein	AB	CDEFG						Labels
AB	0							Man-Monkey
CDEFG	14	0						Dog-Pig-Rabbit-Horse-Donkey

Figure 6: Join cells: Step 5



Protein	ABCD EFG							Labels
ABCD EFG	0							Man-Monkey-Dog-Pig-Rabbit-Horse-Donkey

Figure 7: Join cells: Step 6

4. Your fourth task is to print the intermediary and final clusters. You can simply print them as shown below (please pay attention to the parenthesis) or as a tree-like visualization.

1. (A,B)
2. (D,E)
3. (C,F)
4. ((C,F),G)
5. (((C,F),G),(D,E))
6. ((A,B),(((C,F),G),(D,E)))

NOTES:

1. You are **not** allowed to use any 3rd party libraries except NumPy.
2. You can use any programming language of your choice.
3. You're encouraged to use Jupyter Notebook either with Python or R.
4. **If you use Jupyter Notebook**, you can prepare your code and report, which explains your code and evaluations of your outputs, at the same place and only submit your Notebook.
5. **If you do not use Jupyter Notebook**, please submit a fully commented code as well as a report that includes; detailed explanation of the functions and parameters, *screenshots* from running your code, the outputs of your UPGMA algorithm for the given file, and the intermediary and resulting clusters.

Submission: Please submit your assignment using Moodle. Upload a single zip file named as **YourNameSurnameAssignment2.zip**. Your zip file should include the **report**, **source code**, **requirements.txt** that has the requirements to run your program, and the corresponding **READ.ME** file that explains how to run your code. Your READ.ME file should clearly explain how to run your program. Feel free to reach me from my e-mail address, selen.parlar@boun.edu.tr.

Late Submission: You are allowed a total of 3 late days on homeworks with no late penalties applied. You can use these 3 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 1 day late. After using these 3 extra days, 10 points will be deducted for each late day.