Phylogenetic Tree Deduction Using Python

Dane Laufer

Introduction/Background

Phylogenetics is the process of inferring evolutionary relationships between organisms by analyzing their DNA sequence data. This process allows us to create a phylogenetic tree which tells us which organisms are closely related in the context of evolution.

Hypothesis/Problem

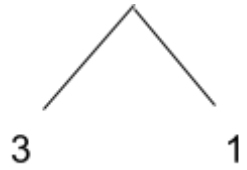
My goal in this project is to create an algorithm that can give us a phylogenetic tree by analyzing the genome data of different organisms.

Methods

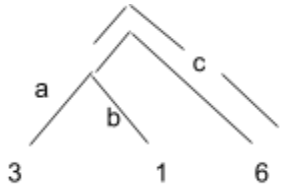
There are multiple methods that I used in this project

1. Gene Location
    a. Created a python script which takes a gene reference sequence, and then searches through the organism's genome to look for matches for the first 30 base pairs of the sequence. The program then saves the entire gene sequence for that organism.
2. Sequence comparison
    a. After I collected the gene sequence of each organism, I compared the sequences against each other and created a table that shows the differences between organisms.
3. Phylogenetic tree assumption

    I will walk through the process of making a phylogenetic tree. First choose the 2 most closely related samples:



    Then find the sample that is most closely related to the first 2 and add it.



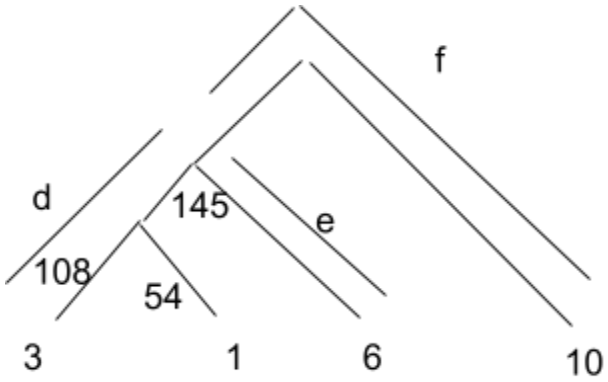
    Calculate the distances using the following formulas

    a = (d(3,1) +d(3,6) - d(1,6))/2 = (162+253-199)/2=108

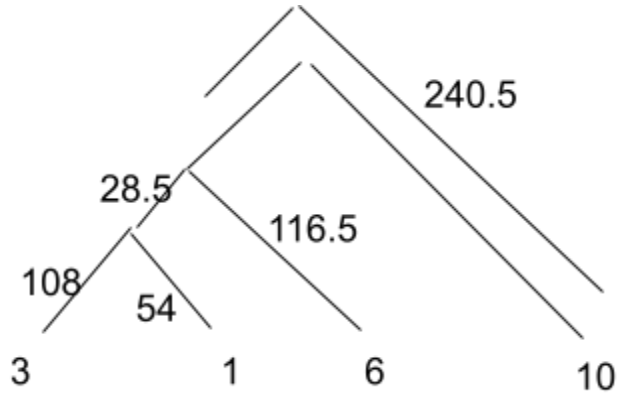
    b = (d(3,1) +d(1,6) - d(3,6))/2 = (162+199 -253)/2 = 54

    c = (d(3,6) + d(1,6) - d(3,1)/2 = (253 + 199 - 162)/2 = 145

Then add the sample that is most closely related to the 3 current samples and add it:



Calculate the lengths using the following formulas:
d = (d(3,6) + d(3,10) - d(6,10))/2 = (253 +377 -357)/2 = 136.5
e = d(3,6) - d = 253 - 136.5 = 116.5
f = d(6,10) - e = 357 - 116.5 = 240.5
Note that you can repeat this final step for as many samples that you have.
Here is the tree we ended up with:



Data

For this project, I used the whole genome sequence from 10 different E-coli organisms. And used referenced sequences for 10 different E-coli genes. I got the data from the National Library of medicine.
The genes I used were: nhaR, nhaA, thrA, thrC, thrB, dnaJ, ribF, mog, yaaA, and satP

Results

After running the script to locate the genes, I found that 4 E-coli samples had 8 of the genes and the other 6 E-coli samples only had the 2 genes that the other 4 were missing. So I split the data into 2 groups:
Group 1: Samples: 1, 3, 6, 10 Genes: nhaR, nhaA, thrA, thrC, thrB, dnaJ, ribF, mog
Group 2: Samples: 2, 4, 5, 7, 8, 9 Genes: yaaA, satP

Group 1:

I created tables that show the differences between the samples for each gene. Here they are:

dnaJ - 1131 nucleotides

	1	3	6	10
1	0	12	22	24
3	12	0	25	12
6	22	25	0	27
10	24	12	27	0

mog - 588 nucleotides

	1	3	6	10
1	0	4	8	6
3	4	0	8	9
6	8	8	0	7
10	6	9	7	0

nhaA - 1167 nucleotides

	1	3	6	10
1	0	15	58	79
3	15	0	59	83
6	58	59	0	89
10	79	83	89	0

nhaR - 906 nucleotides

	1	3	6	10
1	0	29	28	138
3	29	0	35	136
6	28	35	0	137
10	138	136	137	0

ribF - 942 nucleotides

	1	3	6	10
1	0	16	10	28
3	16	0	14	31
6	10	14	0	24
10	28	31	24	0

thrA - 2463 nucleotides

	1	3	6	10
1	0	49	41	44
3	49	0	64	59
6	41	64	0	53
10	44	59	53	0

thrB - 933 nucleotides

	1	3	6	10
1	0	18	4	15
3	18	0	18	20
6	4	18	0	15
10	15	20	15	0

thrC - 1287 nucleotides

	1	3	6	10
1	0	19	28	25
3	19	0	30	27
6	28	30	0	5
10	25	27	5	0

I then added all the tables together and made the overall table:

Group 1 - 9417 nucleotides

	1	3	6	10
1	0	162	199	359
3	162	0	253	377
6	199	253	0	357
10	359	377	357	0

Group 2:

I created tables that show the differences between the samples for each gene. Here they are:

yaaA - 777 nucleotides

	2	4	5	7	8	9
2	0	2	1	0	16	4
4	2	0	3	2	18	6
5	1	3	0	1	17	5
7	0	2	1	0	16	4
8	16	18	17	16	0	20
9	4	6	5	4	20	0

satP - 567 nucleotides

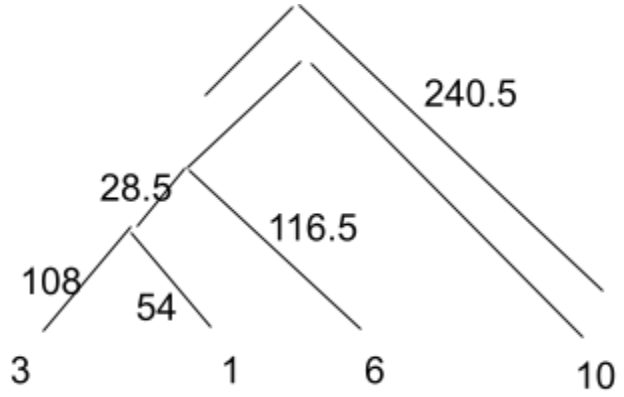
	2	4	5	7	8	9
2	0	6	6	0	12	4
4	6	0	0	6	12	2
5	6	0	0	6	12	2
7	0	6	6	0	12	4
8	12	12	12	12	0	10
9	4	2	2	4	10	0

Then I added the tables together and created the overall table:

Group 2 - 1344 nucleotides

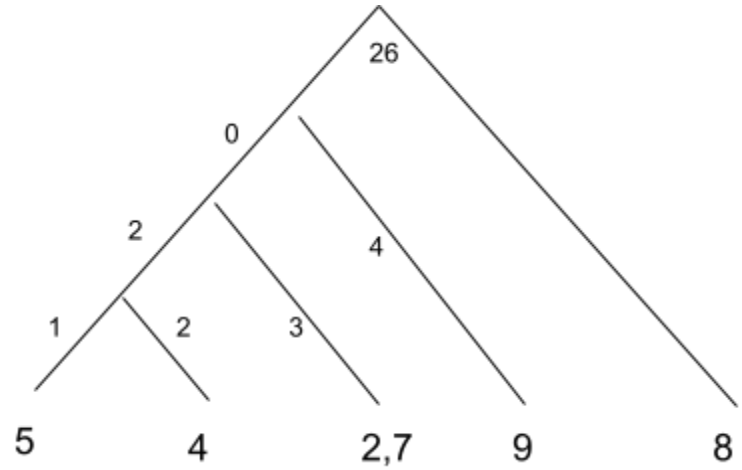
	2	4	5	7	8	9
2	0	8	7	0	28	8
4	8	0	3	8	30	8
5	7	3	0	7	29	7
7	0	8	7	0	28	8
8	28	30	29	28	0	30
9	8	8	7	8	30	0

Group 1:



Here is the tree that was created from the group 1 data. We found out that 3 and 1 were the most closely related and 6 was pretty closely related to 3 and 1 as well. Sample 10 was clearly the one that was furthest away from the others.

Group 2:



Here is the tree that was created from group 2 data. Samples 2 and 7 were identical in the 2 genes we looked at so they were combined into one branch. Samples 5 and 4 were also very closely related. Sample 8 was very distinct from the rest of all the samples and this is evident by how large the distance is compared to the other ones. In general, samples 5,4,2,7 and 9 were relatively similar.

Summary:

My goal of this project was to create a system that allows us to analyze genomic data from organisms and figure out how closely related they are. To do this, I collected 10 sample E-coli genomes and found gene reference sequences for 10 genes. I found that in group 1, samples 3,6, and 1 were relatively closely related and sample 10 was not very closely related to the others. In sample 2, 5 and 4 were very close and 2 and 7 were identical. Sample 8 was not closely related to the other samples but in general, samples 5, 4, 2, 7 and 9 were pretty closely related to each other.

The cool thing about this method is that it can now be expanded to look at more samples and genes. We could use this algorithm on a very large set of samples, analyzed on as many genes as we choose.