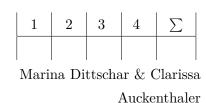
Sequence Bioinformatics

WS 2022/23

Tutor: Anupam Gautam



Blatt 7

(Abgabe am 14.12.2022)

Task 1: Mash sketches (4 points)

The code for the method SortedSet<Integer> computeSketch(int k, int s, String genome) starts at line 94 in file Mash_Auckenthaler_Dittschar.java. And the method String computeReverseComplement(String dna) starts at line 131 in file Mash_Auckenthaler_Dittschar

Task 2: Jaccard index (3 points)

The code for the method computeJaccardIndex(int s, SortedSet<Integer> sketchA, SortedSet<Integer> sketchB) starts at line 161 in file Mash Auckenthaler Dittschar.java.

Task 3: Mash distances (1 point)

The code for the method computeMashDistance (int k,double jaccardIJ) starts at line 184 in file Mash_Auckenthaler_Dittschar.java. By hand, we put in the resulting distances in the mega format (see Figure 1).

```
#mega
!Title:
!Format DataType=Distance DataFormat=UpperRight NTaxa=13;
[1] #Candidatus_Accumulibacter_sp._SK-12
[2] #Candidatus_Accumulibacter_phosphatis_isolate_UW-LDO-IC
[3] #Candidatus_Accumulibacter_sp._BA-93
[4] #Candidatus_Accumulibacter_phosphatis_isolate_UBA5574
[5] #Candidatus_Accumulibacter_sp._BA-92
[6] #Candidatus_Accumulibacter_phosphatis_isolate_UBA2327
[7] #Xanthomonadales_bacterium_UBA2790
[8] #Candidatus_Accumulibacter_sp._51_isolate_3
[9] #Candidatus_Accumulibacter_sp._66-26
[10] #Candidatus_Accumulibacter_sp._isolate_SCELSE-1
[11] #Candidatus_Accumulibacter_phosphatis_clade_IIA_str._UW-1
[12] #Candidatus_Accumulibacter_phosphatis_strain_Bin19
[13] #Candidatus_Accumulibacter_phosphatis_isolate_HKU-1
                 0.24704 0.22655 0.23012 0.31167 0.15643 0.35244 0.24233 0.22655 0.15336 0.25777 0.23392 0.29854 0.16929 0.23797 0.16929 0.24233 0.29854 0.23797 0.23797 0.26396 0.27089 0.24233 0.16929
[1]
[2]
[3]
[4]
[5]
                                    0.25216 0.15643 0.23392 0.31167 0.24233 0.22655 0.22319 0.22655 0.22001 0.16800
                                             0.24233 0.23012 0.28781 0.24704 0.23392 0.22001 0.22655 0.11945 0.23797
                                                      0.26396 0.32859 0.25777 0.25777 0.23392 0.22319 0.24704 0.13188
[6]
[7]
[8]
                                                                \hbox{0.32859 0.27089 0.25216 0.14501 0.23392 0.22319 0.27089} 
                                                                         0.32859 0.32859 0.31167 0.31167 0.31167 0.35244
                                                                                  0.27089 0.23392 0.27089 0.28781 0.25777
                                                                                           0.22655 0.27089 0.22655 0.23012
                                                                                                    0.21412 0.20877 0.23392
[11]
                                                                                                             0.23797 0.23392
[12]
                                                                                                                      0.23797
```

Figure 1: MEGA matrix of the resulting distances.

Task 4: Bacterial tree (2 points)

To run our script, navigate to the src folder and enter the following command: java assignment07/Mash_Auckenthaler_Dittschar.java 17 800 "assignment07/data-07"

To compute the neighbor-joining tree we used the program MEGA[1]. You can find the Matrix we used to compute the tree attached in file: bacterial_matrix.fasta. This is our resulting tree:

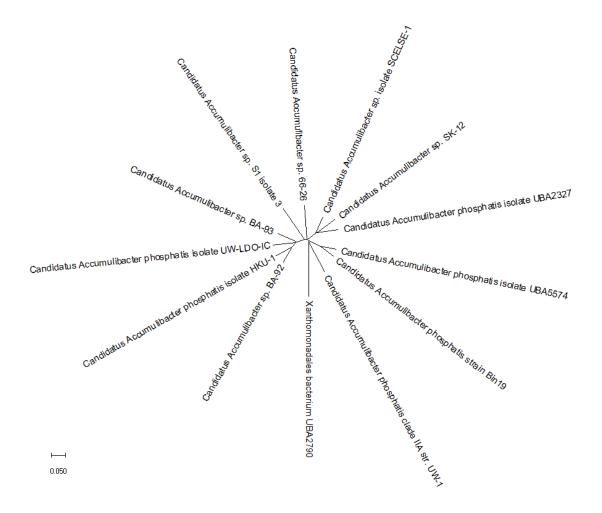


Figure 2: Our resulting bacterial tree

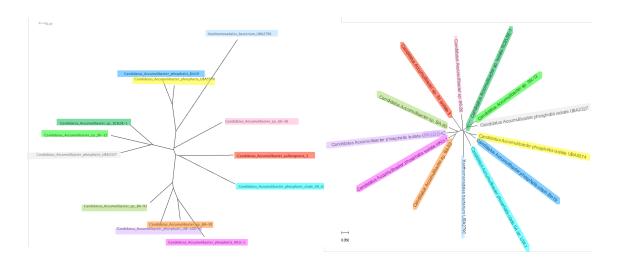


Figure 3: Comparison given tree and our resulting tree

We can see that the computed tree largely matches with the given tree. One thing that does not match is that there is a connection between the red and light neon blue taxa in the given tree that does not occur in our computed tree. Most importantly, the gray blue taxon is the most different in the given tree and as well in our computed tree.

References

[1] Koichiro Tamura, Glen Stecher, and Sudhir Kumar. Mega11: molecular evolutionary genetics analysis version 11. Molecular biology and evolution, 38(7):3022–3027, 2021.