Bioinformática / Algoritmos para Bioinformática 21/22

Phylogenetics

Consider the following multiple sequence alignment from a set of four sequences. Using a MSA with match = 1, mismatch = -1 and gap = -1 calculate:

s1:ATAGC
s2:ATGAC
s3:AACG
s4:AATCG

Task 1

Do manually on paper:

1) Implement the code from MSA and the above parameters define the multiple alignment for the above sequences. Note: if you haven't finished the implementation of all the functions request the resulting alignment.

Task 2

2) Calculate the distance matrix. Assume the metric distance as the number of distinct characters in pairwise alignment (assume the pairwise alignment given by the MSA).

3) Build the tree for the sequences using the UPGMA algorithm.

Task 3

4) Write a function called get_cluster that given a tree returns all the elements (leaves) in the tree as list.
Hint: Traverse tree and collect the elements in the leaves. If it is in an internal node (value == -1) call the function recursively. Note: The method extends a list allows to add the elements of list2 are added to the end of list1: list1.extend(list2).

5) Implement the method, exists_leaf. For a given input value, returns a boolean indicating if the value exists or not. The value should be found as a leaf in the tree.

6) Fill in the code for method *execute_clustering()* in class HierarchicalClustering.

7) Fill in the code for methods *create_mat_dist()* and *run()* in class UPGMA.