Homework 4

CS 4364/5364
Spring 2021

Due: 12 April 2021

1. Assume that we compare the sequences below by counting the number of shared 3-mers (that is how many of the unique 3-mers in one sequence are also in the other). For the following sequences, which pair is more similar.

   - S_1 = `ACGTCGATC`
   - S_2 = `CCGGCGTCA`
   - S_3 = `ACGCTCGAT`

   Does this change if we look at 2-mers?

   Hint: It will help to create a list of all of the possible k-mers and their existence in each of the strings then use that list to calculate the count for each pair of sequences.

2. Given a set of sequences $S_1, S_2, S_3, ..., S_k$, we would like to find k substrings $T_1, T_2, T_3, ..., T_k$ of $S_1, S_2, S_3, ..., S_k$ respectively, such that the optimal SP score of the multiple sequence alignment of $T_1, T_2, T_3, ..., T_k$ is maximized.

   (a) Design a dynamic programming algorithm to solve the problem with match, mismatch, and indel penalties of α, β, γ respectively (i.e. not affine gap scoring).

   (b) What is the time complexity? Explain your answer.

   Note that when $k = 2$, this problem is the same as pairwise local alignment.