Homework 5

CS 4364/5364
Spring 2021

Due: 29 April 2021

1. Given an additive tree $T = (E, V)$ for n species. (a) Describe an algorithm for reconstructing the distance matrix between all of the species. (b) What is the time complexity of the algorithm you described in (a) algorithm?

2. Given a set of sequences $R = (r_1, r_2, r_3, ..., r_\ell)$, which may contain single character changes (i.e. it may be one character from the alphabet, but should have been another). Design an algorithm that outputs a modified set of reads $R' = (r'_1, r'_2, r'_3, ..., r'_\ell)$ that replaces any changes such that the number of k-mers in R' that are erroneous is lowered (it may not be eliminated, we define this below).

   An *erroneous* k-mer is one that occurs at least once and less than 5 times. Note that one character change will impact up to $2k - 1$ overlapping k-mers.

   You can assume you have access to a k-mer conversion function $f(x) = y$ such that assigns an integer $y \in [1...\sigma^k]$ to each $x \in \Sigma^k$, and a k-mer count array $C[0...\sigma^k]$ where $C[y]$ contains the number of times $f^{-1}(y)$ occurs in R. (Here $f^{-1}(y)$ returns the k-mer x given an index y.)

   You can also assume that any window of $2k$ bases will only have 1 error, i.e. there will never be conflicts where two point mutations in the same k-mer. In the case that a character could be replaced with two different characters and satisfy this condition, prefer the one that has more total occurrences across the k overlapping windows.

   Note: There are multiple solutions, some examples include a greedy solution, a dynamic programming solution, and an ILP. You can choose any of these as long as you justify that your solution will not *increase* the number of errornious k-mers.

Example: assume $k = 3$ and the following read segment corrections would be made in this window of 6 characters given these k-mer frequencies:

$$\texttt{...ACTTG...} \longrightarrow \texttt{...ACCTG...}$$

x	$C[f(x)]$
ACA	100
ACC	$50 \to 51$
ACT	$1 \to 0$
ACG	9
ATG	2
CAT	4
CCT	$12 \to 13$
CTT	$4 \to 3$
CTG	$7 \to 8$
CGT	0
TTG	$2 \to 1$
GTG	3

In this example, changing the middle `T` to a `C` eliminates three occurrences of erroronious k-mers. The change in any values in the count array are illustrated, the original value is to the left, the counts after the change are shown to the right.