BRAC University
Department of Computer Science and Engineering

CSE 443: Bioinformatics-I (C)

Quiz 02: Summer 2025 Time: 25 Minutes Marks: 15

Name		ID		Section	

1. Do we need to use a reference genome in RNA-Seq data analysis? Why? or Why not? [3]

2. A dataset of vectors mostly falls within the range [0,10], but it contains a few extreme outliers with expression values near [100,100]. How will standard k-means be affected by these outliers? How can that be handled? [6]

3. k-means++ is an initialization algorithm for k-means clustering that aims to choose initial centroids more carefully to improve convergence.

   (a) Given the following 1D dataset points:

   $$\{2, 4, 10, 12, 20\}$$

   and $k = 2$, suppose the first centroid c_1 is chosen randomly as 2. Calculate the probabilities for each remaining point to be chosen as the second centroid based on their squared distances to c_1. [3]

   (b) If the second centroid is chosen as 12, explain how k-means++ initialization can lead to better clustering results compared to random initialization. Why 12 was selected instead of 20? [3]