ABIN Assignment 1

Dr. Debarka Sengupta

September 4, 2023

Guidelines

1. Deadline: 10th September 2023, 11:59 pm (Midnight).
2. Late submission: 11th September 2023, 11:59 pm (Midnight) (50% of the obtained marks will be deducted).
3. Further late submissions after the 11th of September will be awarded zero.
4. All coding assignments must be submitted as .ipynb files with proper comments.
5. Standard IIIT-D plagiarism policy applies.
6. The assignments have to be submitted in the following manner:
   (a) Create a single Jupyter notebook with proper demarcation of questions and responses (text, code, command, explanation, output, graphs). The accepted language is Python.
   (b) A PDF file comprising the entire Jupyter Notebook and a separate Jupyter Notebook file needs to be uploaded (zip format) at the assignment link on Google Classroom before the deadline.
   (c) One group should submit only from the registered submission ID during group formation. Rest of the members must turn in the assignments with private comments mentioning the name, roll, and submission id. Any violation of above will disqualify your submission from evaluation.
   (d) The name of the PDF and jupyter notebook file should be a combination of the group number and assignment number. For example, `group1_1.pdf`, where 'group1' is the group number and '1' is the assignment number.
7. No shift of the deadline is allowed.
8. All technical problems should be addressed to the TAs for this course. If you want to chat with me, do it on Slack so that everyone benefits. Tag me in that case.

Question 1 (5 points)

Take a string of 10 nucleotides as input. Write a code to produce all possible strings with 0 or 1 or 2 mutations. Use these new strings for constructing a consensus string. Find the Hamming distance between the starting string and the consensus string. Also, write down the formula for the number of strings with 0, 1, 2 mutations.

Question 2 (5 points)

Write a script to download any FASTQC file from NCBI GEO and perform the following QC checks:

1. Read length.
2. Number of reads.
3. Plot nucleotide distribution across locations of the reads. State your observations about biases, if any.

Question 3 (5 points)

Generate 100 random 1KB sequences (as if they are upstream of some genes, you can use real data from UCSC for some co-regulated genes as well). Plant a motif of length 10 after introducing 0, 1, 2 random mutations (uniformly distributed). Use Gibbs sampler to identify the motif locations and the consensus motif.