Problem 1

(10 points)

Please make an account and enroll in the Rosalind version of the class at: `https://rosalind.info/classes/enroll/22ad33c200/`.

Solve the *k-Mer Composition* problem. Use your favorite programming language to solve it.

Problem 2

(5 points)

Please be sure to show all corresponding work.

Assume the dictionary $\{A, C, G, T\}$ with the following distribution:

$$P(A) = 0.1$$
$$P(G) = 0.2$$
$$P(C) = 0.2$$
$$P(T) = 0.5$$

a) What is the expected frequency of the sequence CG in a sequence of length 3 (i.e., probability)? Hint: how many ways can you put CG in a sequence of length 3 and what is its probability? (2 points)

b) What is expected frequency of the sequence CG in a sequence of length 5? (3 points)

Problem 3

(10 points)

Assume the following DNA sequence:

$$ATGATCGAGATC$$

a) Draw the simple de Bruijn graph with $k = 3$, aka, $DeBruijn_3(ATGATCGAGATC)$, that does not collapse any nodes. (1 points)

b) Draw all intermediate collapsed graphs that collapse on common nodes. There are three such cases so be sure to provide three such graphs. (3 points)

c) Redraw your final graph from (b) and find the Eulerian path that corresponds to the original sequence but do not label the edges with their corresponding k-mer. Instead, label the edges on the Eulerian path edges with with an unique increasing integers starting with 1 (e.g. 1, 2, ...). (2 points)

d) Find an Eulerian path that starts with a different k-mer. You can simply write down the edge labels here and make sure to write down the corresponding sequence. (2 points)

e) Draw the final Buijn graph with $k = 5$, $DeBruijn_5(ATGATCGAGATC)$, that collapses any duplicated nodes. (1 points)