==========
+ Write - Up
==========

1. Give a clear statement of the problem.

To attempt to predict whether or not a provided sequence can be labeled as CpG or non-CpG based off of log-odds ratios which were calculated using transition matrices from a provided set of CpG's in the hg38 chr12 set.

2. Give a brief description of the method of your procedure. This should be a few sentences describing the problem, solution, and results. It should be similar to an abstract in a scientific paper.

Given three sets of data, a list of 1211 CpG sequences, the chr12 genome, and a list of 10 test sequences, I parsed the data into list data structures in order to properly sort through the nucleotides. In order to predict whether or not the list of ten test sequences were CpG's or not, I first had to find a positive and negative set of 200 random sequences in order to calculate a transition matrix for each set. Using the transition matrices from known CpG regions for the positive and unknown for the negative, I was then able to calculate log-odds ratios for each of the remaining sequences and the test sequences. My resulting ratios for the remaining 2022 sequences are shown below in the histogram. Based off of the graph, the two data sequences converge around 0. Given this, we can conclude that R>0 are CpG's and R<0 are non-CpG's. This leads to the following results for the test sequences:

0 :	0.22748748394414517	CpG
1 :	-0.10037605507479315	non-CpG
2 :	0.16611083725873013	CpG
3 :	0.10373173209397915	CpG
4 :	0.13147377393278592	CpG
5 :	-0.01258033312608612	not-sure (possibly non-CpG)
6 :	-0.05531276090249152	not-sure (possibly non-CpG)
7 :	-0.3010451279868806	non-CpG
8 :	0.26486436562308097	CpG
9 :	0.11639229598239048	CpG

Address the question of whether the selection of positive and negative sequence sets is adequate for a method of identifying CpG regions. How can this method be improved?

For the positive set, selection was solely from CpG regions, which made this a good control set. But, for the negative set, we were picking randomly from the entire chr12 sequence meaning that some of the sequences selected could have be partially CpG regions. If we were to have only selected from non-CpG regions, this may have made our negative transition matrix more accurate.

3. Present the procedure in detail, including any relevant mathematics. If the solution involves programming, give an overview in words of how the program operates. The presentation should contain sufficient detail so that a reader can follow each step without difficulty. *

First, Parse each of the sequence files (CpG, chr12, test_sequences) into lists
Define the limits for selecting random seq from chr12
initialize a structure to track all previously checked sequences (prev_windows)
find 200 random cpg sequences
for each cpg sequence:
    find a random chr12 sequence that has not been used before and has len(cpg_seq)
    add newly added sequence window to prev_windows to check against
remove 200 cpgs from original set to get remaining 1011 cpgs
perform same step as before to get chr12 200 for the next 1011 from chr12
next find transition matrices for the positive and negative control sets:
    aggregate count of each pair then normalize sums by length of the row
using the two transition matrices, calculate log-odds ratios for remaining 2022 seq and testseq:
    for each seq:
        for each nucleotide pair:
            sum the probabilities in the transition matrix that apply
        normalize by the length of the seq
        append to list
plot the calculated ratios for 2022 seq
output the 10 test seq ratios

4. If the solution involves programming code, output must be clearly labeled with each item having a descriptive title. Programming code must be documented with comments describing each major step in the program and all relevant variables. Source code should be submitted in a separate file. *

==========
+ OUTPUT
==========

Positive Transition Matrix
        A   C   T   G
A [[0.1845766   0.26304733 0.11072729 0.44164877]]
C [[0.15626314 0.36087743 0.18244721 0.30041222]]
T [[0.08961764 0.3655013   0.17923529 0.36564577]]
G [[0.16468622 0.34165794 0.12078063 0.37287521]]

Negative Transition Matrix

	A	C	T	G
A	[[0.33535789	0.16740067	0.2545371	0.24270435]]
C	[[0.35344095	0.25438969	0.34455537	0.04761399]]
T	[[0.21650843	0.20225514	0.3315354	0.24970102]]
G	[[0.29877605	0.2031483	0.24732152	0.25075413]]

Test Sequence Log Odds Ratios:

0 : 0.22748748394414517
1 : -0.10037605507479315
2 : 0.16611083725873013
3 : 0.10373173209397915
4 : 0.13147377393278592
5 : -0.01258033312608612
6 : -0.05531276090249152
7 : -0.3010451279868806
8 : 0.26486436562308097
9 : 0.11639229598239048

5. Graphs must be titled and have each axis labeled clearly