Assignment 1 Solutions

Usage Statement:

```
python3 nuc_count.py hs_ref_GRCh38.p2_chr22.fa.gz
python3 make_seq.py 1000000 0.2659 0.2332 0.2361 0.2649 > random_seq_1M.txt
```

Question 1:

Run `nuc_count.py` on `hs_ref_GRCh38.p2_chr22.fa`. How many times do each of the 4 nucleotides occur on chr22?

A: 11198038
C: 9819368
G: 9943117
T: 11155239
N: 53597

Question 2:

Run your modified `nuc_count.py` on `hs_ref_GRCh38.p2_chr22.fa.gz`. What are the frequencies of the 4 nucleotides on chr22?

Nucleotide Frequencies
A: 0.2659
C: 0.2332
G: 0.2361
T: 0.2649

Question 3:

Run the modified `nuc_count.py` for both human chr22 and your generated 'my_compare.txt' from part 4. Compare the two lists of frequencies. What are the differences? Can you provide a biological explanation for these differences?

Question 3.1:

What are the dinucleotide frequencies from chr22?

AA: 0.0795
AC: 0.0509
AG: 0.0748
AT: 0.0606
CA: 0.0763
CC: 0.0668
CG: 0.0163
CT: 0.0738
GA: 0.0628
GC: 0.0542
GG: 0.0681

GT: 0.0510
TA: 0.0473
TC: 0.0612
TG: 0.0769
TT: 0.0794

Question 3.2:
What are the dinucleotide frequencies for the generated `random_seq_1M.txt`. (Since the sequence was generated randomly, your frequencies may be slightly different.)
AA: 0.0712
AC: 0.0622
AG: 0.0625
AT: 0.0706
CA: 0.0622
CC: 0.0540
CG: 0.0550
CT: 0.0619
GA: 0.0627
GC: 0.0548
GG: 0.0561
GT: 0.0622
TA: 0.0705
TC: 0.0620
TG: 0.0622
TT: 0.0700

Question 3.3:
What are the differences? Can you provide a biological explanation for these differences?
There are a couple possible explanations, but the primary one we were looking for was the lack of CGs within chr22 as compared to the randomly created sequence (CG suppression). It was important to comment on the non-random nature of the genome. Repeats and selection could also play a role in this, but focus on comparison of just nucleotide frequencies or just the coding part of the genome was not given full credit.