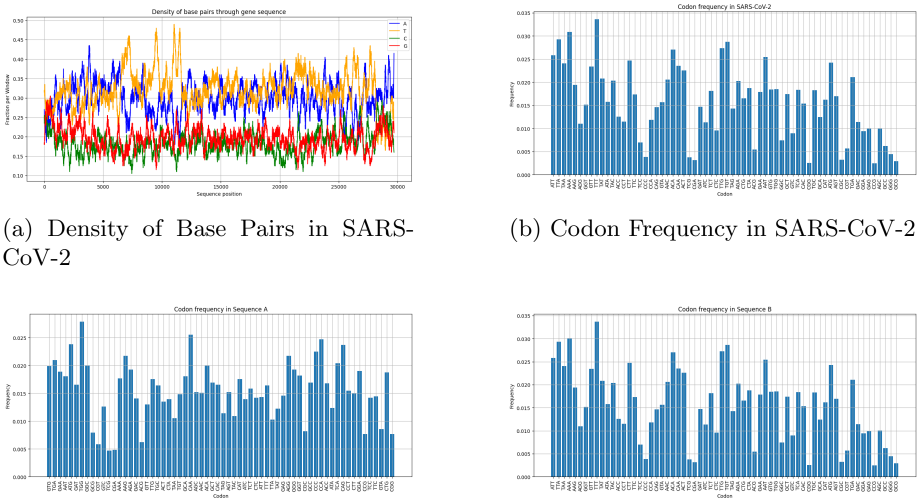
Implementation

The implementation went smoothly, largely due to my prior experience with Python. I used ChatGPT several times for help, specifically with plotting. It served as a useful reference, effectively providing documentation by explaining the options available and how they worked. This approach worked well because I wasn't always sure what I needed to do, and being able to see all possible options was helpful.

Questions

1.

all figures have a width of 14 and height of 6



(a) Density of Base Pairs in SARS-CoV-2

(b) Codon Frequency in SARS-CoV-2

(c) Codon Frequency in Gene Sequence A

(d) Codon Frequency in Gene Sequence B

2.

Kullback-Leibler values between:
Sequence A and SARS-CoV-2: $0.138\,011\,729\,443\,000\,6$
Sequence B and SARS-CoV-2: $9.028\,641\,025\,468\,603 \times 10^{-6}$

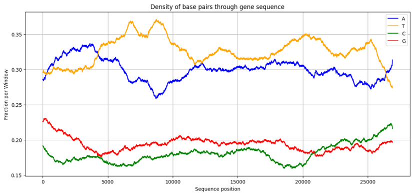
3.

it is most likely that Sequence B is a covid variant you can see this by observation in comparing figure 1b with figure 1d. It can also be seen quantitatively by see how much smaller the Kullback-Leibler value of Sequence B when compared to SARS-CoV-2

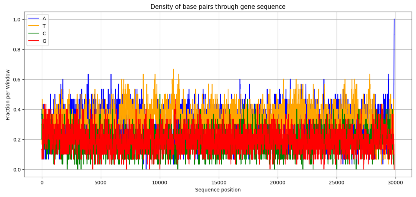
4.

if were were to use count to find the count of all the unique codons we would have to iterate through all the possible codons. the dictionary method is much better, just inching along the gene and every time you check the codon and dictionaries have very nice tools to tell if and how many of them have been found.

5.



(a) High nWind of 3000



(b) Low nWind of 30

from the figures you can see that a higher n window gives a higher correlation between A and T base pairs as well as between C and G base pairs which makes sense cause they are sister pairs.