AM 129 Homework 4 Report Deliverables

Owen Morehead

Nov 10 2020

Both figures are shown at the end of the report. The window width for the density plot I chose was **nWind=500**. This window size seemed to match the figure shown in the homework assignment and the window size seemed like a good middle ground value. I found that if I chose a window size too small, say nWind=20, all the data was overlapping because the fraction of base pairs in each window was too similar. On the other hand, when I choose a very large window size say, Nwind=20,000, we get to see a larger view as the window size is nearly the entire size of the data sequence and so there is a much larger amount of data pairs to count through. This produced an interesting plot as we can see how many of each base pair are in the sequence in total (most of the total sequence that is). In terms of the fraction of the window size (nWind = 20,000), I found that approximately: 0.30 were A, 0.33 were T, 0.18 were C, and 0.19 were G.

To generate the second plot, the *count* method works well to extract how many base pairs of a given type are in each window. This is because there are only four different base pairs (A, T, C, G). There are many more different codons (triplets of base pairs) as can be seen in the histogram. We do not know these codon values initially until we loop over the data and therefore if we were to do something like geneStr.count('ata') to generate the codon histogram, we would first have to devise a way to tell the program every codon name to count and then we would be able to put those keys and values in a dictionary. The pure dictionary method we chose to implement to count each codon uses less code and is more efficient. By running through the data and adding one to the dictionaries value every time the same codon (key) in the dictionary shows up, we can tally and plot all the different codons in the data very easily even if we don't know what the codons are initially.

Other than my brief data interpretation above, I won't add much else as I don't really have experience with genetics or sequence data analysis. However I did browse the internet on this Covid sequence data and find it very cool and interesting that we can and have isolated this virus and extracted such a detailed data sequence in which people can study and discover important aspects of the virus.

Figures:

Fig 1 Codon Frequency

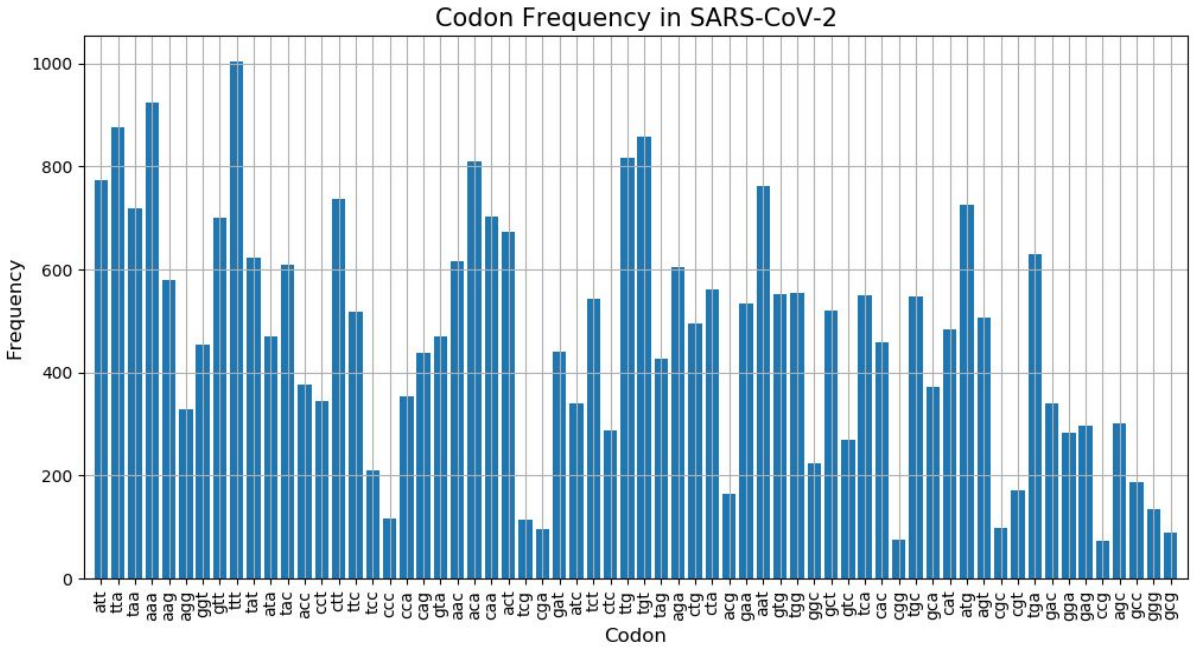

Fig 2 Base Pair Density