Simon Lee
AM 129

HW4 Report

Genetics Introduction

I got really excited about this assignment in particular because genomics/bioinformatics is what I want to do after the completion of my undergraduate degree. Studying genetics data is one of the world's leading practices due to all the biological data you can extract from genome sequences. From learning how a disease is caused or even spread, you can learn a lot from these genomes. These are also frequently being studied at the DNA, RNA and even protein levels and it can give valuable information about a species (taxa). In this assignment we manipulate some of the basic features of the *SARS-CoV-2* a.k.a the Coronavirus. Although we specifically only work with the DNA in this assignment, it is still important to be able to visualize some of the biological data since it is very hard to visualize it to begin with.

The purpose of a DNA sequence is to be transcribed into RNA and then eventually translated into a protein. And part of this assignment's task is to count all the different types of codons using a sliding method. The purpose for this sliding method is because a protein is constructed from amino acids and the start of every single protein is an 'M ' or the 'AUG' (RNA) or 'ATG' (DNA) codon. Until it reaches this start codon, the cells are actually not allowed to begin translating the mRNA. Once it does find the first occurrence of the start codon it will transcribe until reaching 1 of 3 different end codons. These three at the DNA level are 'TAA', 'TAG' and 'TGA'. Once it reaches an end codon it once again does not translate until it finds its next start codon. So though this sliding mechanism is overfitting of what our data looks like, it is justified. And based on other sources we know that there exists 29 different proteins on the COVID virus. These proteins have mutated often throughout this pandemic resulting in more harmful versions of this virus (delta variant, gamma variant).

Before we begin analyzing our graphs I just wanted to bring up the importance of the spike protein in particular. The recent *Pfizer* and *Moderna* vaccines are very new technologies based off of this sequencing. It is the first of its kind of what we call RNA vaccines. And to briefly mention what RNA is, it is DNA except its nucleotide base pairs exclude the 'T' and replace it with the 'U' to have 'A', 'C', 'G', and 'U'. But essentially what these vaccines contain are messenger RNA's (mRNA's), that contain the genetic information on how to construct the spike protein. And once this is injected into us our immune system builds the antibodies to combat this virus. So in a more simplified explanation, if the COVID-19 was to enter our body, our immune systems would be able to recognize the spike proteins which are the virus's distinct features and have our immune cells attack this virus. With these new RNA vaccines, we can develop effective prevention mechanisms if we were to simply just have the sequence of the virus. It is for these reasons that this vaccine development is under much controversy. With technology advancements, we are able to construct a working vaccine in less than a year when it used to take years of work to do such things. This brief introduction to genetics is some of the reasons why it is such a popular practice.

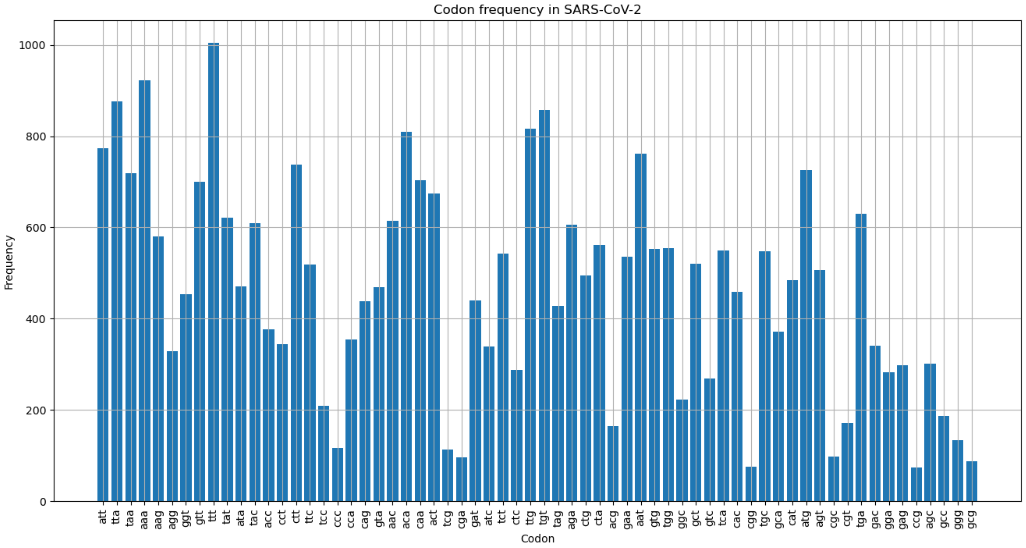
Codon Counter Figure

We briefly mentioned that one of the specific goals in this assignment was to count the frequencies of each codon for the *SARS-CoV-2.* And although this is very overfitting of our data, it is simply just counting the numbers in the sense that if there was a mutation of some sort, it could drastically change the way this virus behaves or looks. What I mean by this is that if there was a manipulation to a start or end codon (i.e 'ATG'->'ACG'), this could change where we begin to translate a protein and could give its structure drastically different features. Infact the delta variant is a mutation and its mutations actually improved its ability to bind with our cells and inject the genetic information to make us "more sick". Therefore these variants have become more dangerous due to how mutations change the way a virus can behave or look.

But going back to the assignment. The codon table is constructed of 64 unique codon tables. Using simple combinatorics math we are able to derive this number 64 from there being no order and replacement being okay for each nucleotide base.

$$4 \text{ base} * 4 \text{ base} * 4 \text{ base} = 64 \text{ different combinations}$$

Doing so we obtain the following graph from the sliding mechanism:

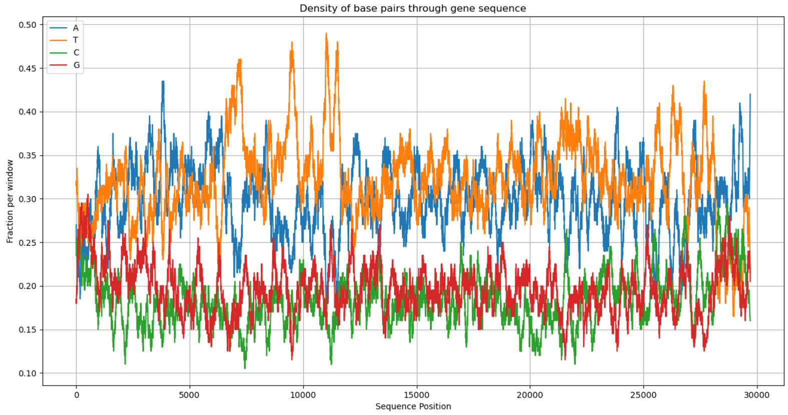


There is not much to analyze here other than the fact that these are the frequencies. Just to point out a few of the notable codons, there are about a little over 700 different start codons in this sliding mechanism. What we can infer from this data is that if there were a mutation that deleted a base pair, it could totally change how the structure and the function are entirely changed. There are also a little over 400 ‘TAG’, a little over 600 ‘TGA’ and a little over 700 ‘TAA’ end codons. So once again a shift or mutation can really alter how these proteins become translated. It really shows how interesting a genome is and shows how many “what if” scenarios that could possibly occur.

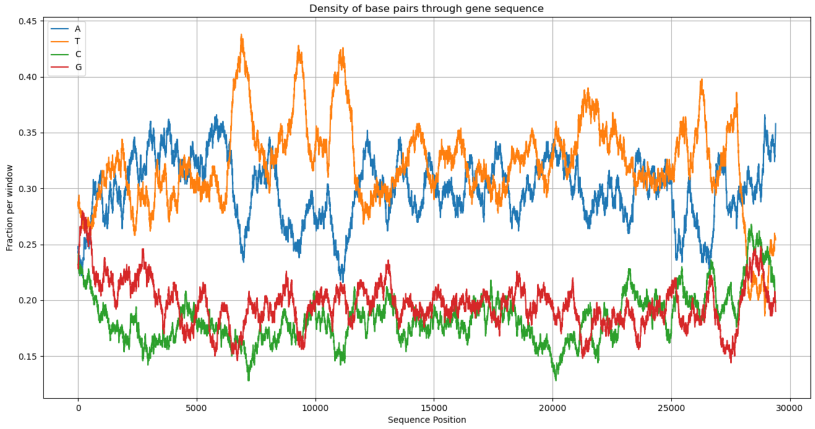
Density Graphs

The final part of the project is to visualize how much a single nucleotide base appears in a specific window. We are once again going to use the sliding mechanism and the objective of this function was to graph the fraction that equates to density. Just to run an example lets say we have a `nWind = 200` , and the letter a occurred 58 total times in this window then we would know that the density of the nucleotide base a is .29. We continue to do this until the very end by sliding the sequence by 1. So what we are left with is around 27,000 different fractions that resemble the density. But this is only true for when the window size is 200. It would be a little less if the window was larger and it would be a lot more plots if the window size was smaller. So below we plot some of the graphs from this program.

`nWind = 200`



`nWind = 500`



The relationship we want to see within this program is that the larger the window size is, the more spread out the data will be. The reason this is true is because if we were to divide frequencies by a bigger window size we would have a smaller density number. And if we were to do the opposite all the lines would be much more clustered. So what we want to see from this mapping is that the window sizes directly affect how our density graph will look.

Conclusion:

Overall this program was very fun. Visualization as mentioned is a big component of scientific computing especially in biological studies. By being able to visualize how frequently a codon appears and how much a nucleotide base appears in a given window can truly tell a lot about a sequence. It is crazy to think about how a few codons or nucleotide bases can totally alter this virus that has torn through our world. It is also pretty crazy to think that these genomes are what governs not only this virus but also us. After all, the human genome is just a longer version of these A’s, C’s, G’s and T’s. But I really enjoyed this assignment overall. I think it will play a big role in my academic career and I was glad that it just happened to be in the field that I am most interested in.