Q4. Compare your alignments from #2 and #3. In what way did the free start and end gaps improve the alignment?

- Firstly, having free start and end gaps improved the score by 636 points. (Alignment2 Score - Alignment1 Score = 1711-1075 = 636). Moreover, Since the 2 sequences, the spike protein and the pfizer vaccine, have varying lengths, the second alignment was able to offer a more accurate alignment since it was free to have as many gaps at the end for the smaller sequence in order to find the best alignment amidst the longer sequence. Thus, compared to the first alignment, it wasn't having to compensate for its gaps at the beginning with finding and 'forcing' alignments for a better score due to the 'greediness' of the algorithm, when perhaps the better alignments of the 2 sequences were in fact better farther down the other longer sequence. Moreover, the second alignment was able to have the shorter sequence insert trailing gaps in order to compensate for the rest of the larger sequence. Overall this is a much more logical alignment since the query/shorter sequence is able to begin alignment where it finds the most matches or it finds the most optimal start position with respect to the longer sequence, without incurring penalties.

Q5. Based on this annotated image of the full sequence of Pfizer mRNA vaccine, why did it make sense to ignore start gaps and end gaps in your alignment in #3?

- As mentioned a bit in the previous response to Q4, it makes sense to ignore start gaps and end gaps since it allows the query sequence to begin matching/alignment in which is deemed more optimal for scoring, without having to be forced to include matches in order to compensate for the gapping. Focusing more on the question, we can see that the beginning of the vaccine strand represents 5 prime UTR bases, Start codon(ATG), and signal peptide bases. Moreover the ending of the vaccine strand represents the Stop codon(TAG) and the 3 prime UTR bases. All of these bases are important to the overall function of the vaccine, but the actual encoding and specificity of it comes within the orange portion that resembles the Spike Protein sequence. Therefore, being able to insert start and end gaps to the query/shorter sequence representing the Spike Protein makes sense since it is able to align optimally with the Pfizer vaccine Spike Protein encoding, without incurring any penalties to do so.

Q6. How many mismatches (counting gaps as a mismatch if any) are there between the real spike protein and the Pfizer version, in the coding portion of the RNA sequences?

- Based on my answer based on the output of Q3 (outputQ3.txt), there were a total of 1033 mismatch pairings for bases and then I counted an insignificant amount of about 11 gaps beside those mismatch pairings, so a total of 1044 mismatches totals.

Q9. How many mismatches are there in the two amino acid sequences (counting gaps as a mismatch if any)? What exactly is different between the amino acid sequence of the vaccine and the amino acid sequence of the real spike protein?

- There are only **2** mismatches found within the *middle* of the spike protein encoding for the vaccine and the actual spike encoding. The mismatch has to do with 2 Proline amino-acids. The two encodings for the spike protein don't differ anywhere else.

Q10. Describe why your findings in #9 make sense in the context of this article

- After reading the article, the findings of the sequence alignment for the protein makes sense because of the characteristics of the spike protein in the corona virus and also the changes to the said spike protein sequence for the vaccine. Vaccines essentially work by introducing the virus to our cells in order to build their immunity and antibodies to that said virus. Although, an issue that was encountered when developing the vaccine was how easily spike proteins transform from their prefusion shape to their postfusion shape when fusing to our cells. To combat this, the vaccine changed 2 amino acids to Prolines. Prolines are the most rigid of the 20 other amino acids and by making these changes to a key joint of the spike protein, it allowed the spike protein to stabilize on the cell but not actually fuse with the cell. That's why it can be seen in the vaccine amino acid sequence to be almost identical to the Coronavirus's spike protein, but yet just differs in 2 amino acids to make the spike protein non-lethal.
- …RLDKVEAE…
- …| | | x x | | |...
- …RLDPPEAE…

Q11. Why did the vaccine makers introduce so many synonymous mutations into the vaccine? Why didn't they just copy the spike protein sequence exactly (apart from the amino acids discussed in question 10)?

- The GC values of the 2 sequences are 37.3% for the *virus* spike protein encoding and 56.9% for the *vaccine* spike protein encoding.
- There could be a few reasons as to why the vaccine makers introduce so many synonymous mutations into the vaccine. Firstly as a premise, the vaccine makers were still able to replicate the amino acid structure of the spike protein for the vaccine, except for the 2-Proline mentioned above for fusion stability. While this may be true, the same codons were not used in order to get that respective amino acid sequence. Calculating and comparing the GC content for the 2 protein encoding sequences we can see that the virus's spike protein has a GC content of 37.3% and the vaccine spike protein has a GC content of 56.9%. With this being said, the vaccine makers were still able to translate differing codon sequences to have almost identical protein, which was done intentionally, but why?
- For one, GC pairs will be more strongly bonded and more stable due to 3 hydrogen bonds between the bases. But more importantly, Codon optimization plays a big part in mRNA vaccinations since it has to do with codon bias and mRNA stability (important forprotein gene expression and druggability). By increasing the GC content without altering the amino-acid sequence, the vaccine makers were trying to reach an optimal GC content in which the targeted organism would have a codon bias in which it would emphasize the gene expression for the edited spike protein so that the human body could create antibodies.