Note: I felt that 4 and 5 could be considered pretty similar in what they're asking for, but I tried my best to answer them separately.

[4.] Free start and end gaps improved the alignment by allowing the algorithm to "ignore" the Pfizer MRNA's 5' and 3' UTR without incurring any penalties for inserting gaps and/or mismatching against the reference SARS-Cov2 spike protein genome. Instead, matching began at their respective start codons ATG to the stop codons. This increased the alignment score from 1075 to 1711.

[5.] The Pfizer_MRNA had some 5' UTR stuff and whatnot prior to the appearance of its start codon ATG and some post stop codon stuff with the 3' UTR and whatnot whereas the SARS-Cov-2 spike protein reference didn't have them. By ignoring start and end gap penalties, we're able to match the 5' and 3' stuff with gaps/ignore them and properly optimize matching both genomes from their start codons to end codons. Had it been penalized, the algorithm would've matched the 5'/3' UTR and result in a significantly lower alignment score. It is also worth noting that the lengths of both genomes aren't [relatively] the same length, with the former being 4176 bases long vs the latter's 3822. Because of this, global alignment isn't the way to go, but instead ***semi-global*** alignment. The lecture slides said that it's for comparing a short query against a long reference, but in this case, it's a long query against a short reference. However, that shouldn't matter as much in this specific case/assignment.

[6.] There are 1369 mismatches between the real spike protein and the Pfizer version (Q2) and 1404 mismatches between the former and latter using free start and end gap penalties (Q3).

[9.] There are two (2) mismatches between the two amino acid sequences for both the cases where start/end gaps are and aren't penalized. The vaccine had 2 Proline (P) bases, which [mis]matched with a Lysine (K) and Valine (V) base respectively. (See question_9*_output.txt for alignment)

[10.] The findings from #9 made sense because of the 20 amino acids, proline (P) is the strongest, which prevents the release of bound area that would lead to fusion. In the MERS spike protein that they researched prior to the pandemic, there was a region with a small loop of amino acids towards the top of the spike that held two alpha helices together. When the spike protein binds to a human cell, said loop is released, and the two helices "straighten into one long helix that harpoons the human cell and pulls the virus and human membranes close together until they fuse". [Theoretically] The 2 prolines (2P) blocks the release of the loop/coil, allowing the immune system a chance to make antibodies to prevent infection. With the release of the SARS-CoV-2 genome, scientists were then able to compare SARS against MERS and "pinpoint the code corresponding to that bent spring" in the SARS virus and added the 2P mutation to "lock" the SARS-CoV-2 spike to prevent fusion and allow our body's immune system to make antibodies to prevent infection.

[11.] The reason as to why vaccine makers introduced many synonymous mutations into the vaccine was to safeguard against mutations that would occur in the virus. Synonymous mutations are changes in the DNA sequence that codes for aminos for amino acids in a protein sequence, but doesn't change the encoded amino acid (https://genomevolution.org/wiki/index.php/Synonymous_mutation). They are included to potentially optimize the stability or expression of the protein in a vaccine. In doing so, our immune system can be trained to recognize variants of the virus. As to how it rleates to GC content, GC content can promote the stability and expression of specific proteins and can affect codon usage in organisms -- certain organisms have biases towards specific codons due to their GC content (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4592985/). From what I can draw, this means is that manipulating GC content for synonymous mutations in vaccines allows our immune systems to train itself against variants of the virus (?).