Q4:

The free start and end gaps improved the alignment in several significant ways:

1.Increased alignment score: The score rose from 1075 in #2 (fixed gap penalty) to 1711 in #3 (free end gaps). This substantial increase of 636 points indicates that free end gaps allowed the algorithm to find a more optimal alignment solution.

2. Focus on the Core Coding Region:

Free start and end gaps allowed the algorithm to focus on aligning the most critical part of the sequences—the core coding region—where the spike protein is encoded. Without being forced to align non-coding or less relevant regions at the ends, the algorithm could concentrate on optimizing the alignment in the central, functionally important part of the sequences.

This approach led to longer stretches of consecutive matches in the core regions, improving the alignment quality and reducing mismatches.

3.Enhanced matching quality in the middle region: By not forcing alignment at the ends, the algorithm could focus more on aligning the central coding regions. This potentially led to:

- More base matches
- Longer stretches of consecutive matches

4.Improved biological relevance:

By allowing free gaps at the sequence boundaries, the algorithm avoided forcing alignments in non-coding regions, such as the 5' UTR and 3' UTR in the Pfizer mRNA sequence, which are not directly related to the spike protein. Ignoring gaps in these non-coding regions ensured the alignment focused on the biologically meaningful parts of the sequences, specifically the spike protein-coding region.

This resulted in a more accurate representation of the functional similarity between the viral spike protein sequence and the Pfizer mRNA vaccine sequence.

5. Reduction of Gaps in the Internal Regions:

In alignment #2, where start and end gaps were penalized, many gaps were introduced within the internal regions of the sequence to avoid adding gaps at the ends. This resulted in more fragmented alignments, spreading gaps throughout the core sequence, leading to more mismatches (represented by 'x').

In contrast, in alignment #3, with free start and end gaps, the algorithm was able to introduce gaps at the beginning and end of the sequences without penalty. This allowed the core region to be aligned with fewer gaps, reducing the number of internal mismatches. By minimizing unnecessary internal gaps, the alignment focused more on biologically relevant core regions.

Q5:

The Pfizer mRNA vaccine sequence contains additional non-coding regions at both ends (5' UTR and 3' UTR) that are not present in the spike protein gene sequence. These regions are crucial for the mRNA's function in the cell, such as regulating translation efficiency and mRNA stability, but do not correspond to any part of the spike protein gene.

The mRNA vaccine has a specific structure:

5' UTR -> Start codon -> Signal peptide -> Spike encoding region -> Stop codon -> 3' UTR -> polyA tail

In contrast, the SARS-CoV-2 spike protein gene likely only contains the coding region. The signal peptide in the vaccine mRNA, while not part of the mature protein, is essential for proper protein processing and cellular localization.

By allowing free start and end gaps, the alignment algorithm can effectively 'skip' the non-coding regions and focus on aligning the actual spike protein-encoding part of the mRNA with the spike protein gene sequence. This approach prevents forcing biologically irrelevant alignments between the UTRs of the vaccine sequence and parts of the spike protein gene.Moreover, the polyA tail, which is critical for mRNA stability and efficient translation but not present in the viral gene, can be accommodated by the free end gaps without affecting the alignment of the coding regions.

In summary, this alignment strategy, which ignores start and end gaps, enables a more biologically meaningful comparison by focusing on the functionally equivalent parts of both sequences - the actual spike protein-encoding regions. This approach recognizes the structural differences between the engineered mRNA vaccine and the viral gene while emphasizing their functional similarity in encoding the spike protein.

Q6:

There are 1054 mismatches between the real spike protein and the Pfizer version in the coding portion of the RNA sequences, counting gaps as mismatches. In the reference material given by the teacher, the stop condon is TGATGA, so the result is 1054. If only one TGA is counted, the result should be 1051

Q9:

There are 2 mismatches in total between the two amino acid sequences (including gaps if any). Two consecutive amino acid differences were found in the middle of the protein sequence: a) Position 986: Lysine (K) in the SARS-CoV-2 spike protein is changed to Proline (P) in the Pfizer version. b) Position 987: Valine (V) in the SARS-CoV-2 spike protein is changed to Proline (P) in the Pfizer version. These two differences are located in the middle part of the protein sequence, not at the beginning or end. The rest of the sequence, including the start and end portions, are perfectly matched, as expected.

Q10:

The two amino acid changes (K986P and V987P) found in the Pfizer vaccine sequence are intentional modifications known as the "2P mutation." This technique, developed by scientists Jason McLellan and Barney Graham prior to the COVID-19 pandemic, was initially created for other coronavirus vaccines.

The 2P mutation introduces two proline amino acids—the most rigid of the 20 amino acids—at a crucial joint of the spike protein. This modification serves to stabilize the spike protein in its "prefusion" conformation, which is critical for two reasons:

1. The prefusion form is the primary target for neutralizing antibodies that can prevent viral infection.

2. Without stabilization, the spike protein readily transitions from its prefusion to postfusion form, the latter being less effective at inducing protective antibodies.

This seemingly minor alteration significantly enhances vaccine effectiveness by:

- Enabling the immune system to develop antibodies against the most relevant form of the spike protein.

- Potentially boosting the production of neutralizing antibodies, thereby increasing the vaccine's protective capacity.

The 2P mutation has been widely adopted in COVID-19 vaccine development, not only by Pfizer but also by other major pharmaceutical companies such as Moderna and Johnson & Johnson. This widespread use underscores the importance of this innovative approach in modern vaccine design.

Q11:

Vaccine manufacturers introduce numerous synonymous mutations instead of directly copying the spike protein sequence to optimize mRNA vaccine performance. The Pfizer mRNA vaccine's GC content (56.84%) is 19.53% higher than the original SARS-CoV-2 spike protein sequence (37.31%), offering multiple benefits:

1. Enhanced mRNA stability:

    - Higher GC content increases mRNA thermal stability, slowing in vivo degradation.

    - Optimized sequences form more double-stranded structures, further enhancing stability.

2. Codon usage optimization:

    - Human cells prefer high-GC codons; this "codon bias" optimization improves protein expression.

    - The LinearDesign algorithm, optimizing both stability and codon usage, significantly increases protein expression.

3. Improved mRNA secondary structure:

    - Increased GC content alters mRNA secondary structure, potentially benefiting translation.

    - Optimized sequences show higher stability and expression efficiency in vitro and in cells.

4. Avoidance of harmful sequences:

    - Synonymous mutations can avoid sequences that might trigger immune responses or affect functionality.

    - The design avoids long double-stranded regions to reduce potential immunogenicity.

5. Production process optimization:

    - Higher GC content may improve in vitro transcription efficiency.

    - Rapid optimization (11 minutes) facilitates quick vaccine development and production.

6. Enhanced immunogenicity:

    - Optimized sequences produced up to 128-fold higher antibody responses in mice.

- This demonstrates that sequence optimization significantly improves vaccine immunogenicity.

In conclusion, these changes enhance vaccine efficacy, stability, and safety while preserving the spike protein's amino acid sequence. This approach exemplifies the fine-tuning in mRNA vaccine design, optimizing both the encoded protein and the mRNA's characteristics for improved in vivo performance.