Ubiquitous Genomics

Assignment #1 | Apr 29, 2022

Yoni Peleg, Eden Rosenthal, Ashley Newman, Adi Harush

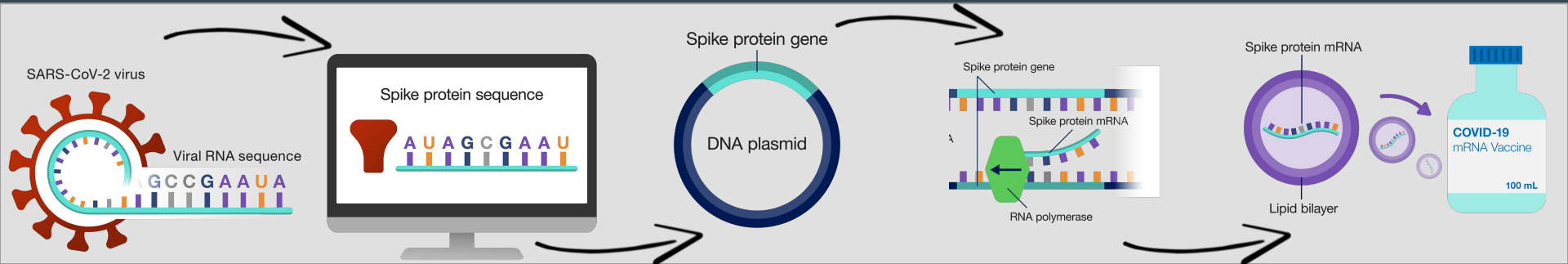
Agenda

- The Task
- Our Process
- Assignment Tasks:
  - Analyzing Pfizer and Moderna Vaccines
  - Comparison to the Wuhan Strain
  - BA.2 Protein Extraction
  - Building the mRNA Vaccine
- Key Takeaways
- References

THE TASK

The Task

1. Our goal was to develop a new vaccine for BA.2
2. How mRNA vaccines work:
   The spike protein is essential for the virus to attach to the host cell, thereby making it an effective antigen.
3. Isolate the spike protein from the sequence of BA.2
4. Use Pfizer's reference to determine the mRNA sequence, 5'UTR and 3'UTR, codon optimizations and the fatty lipid bilayer that protects it on the move.

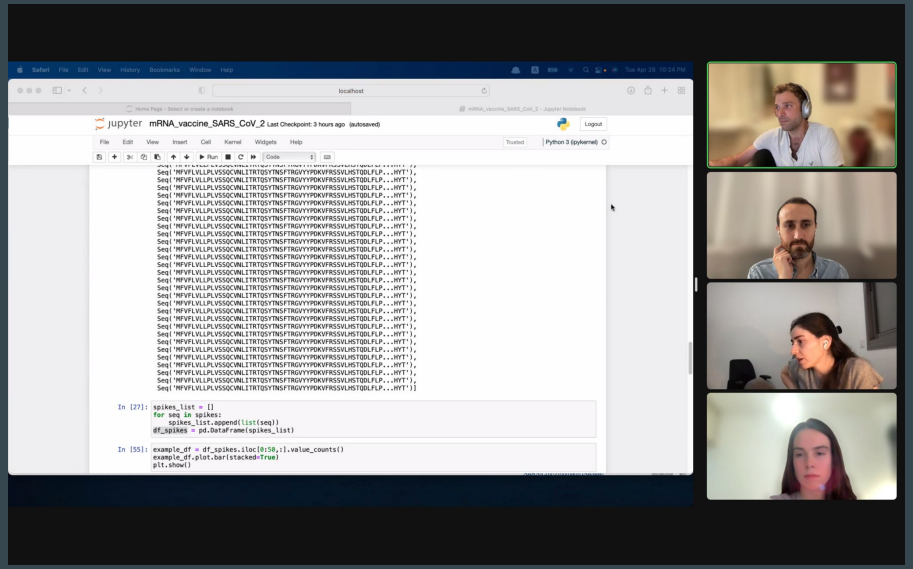


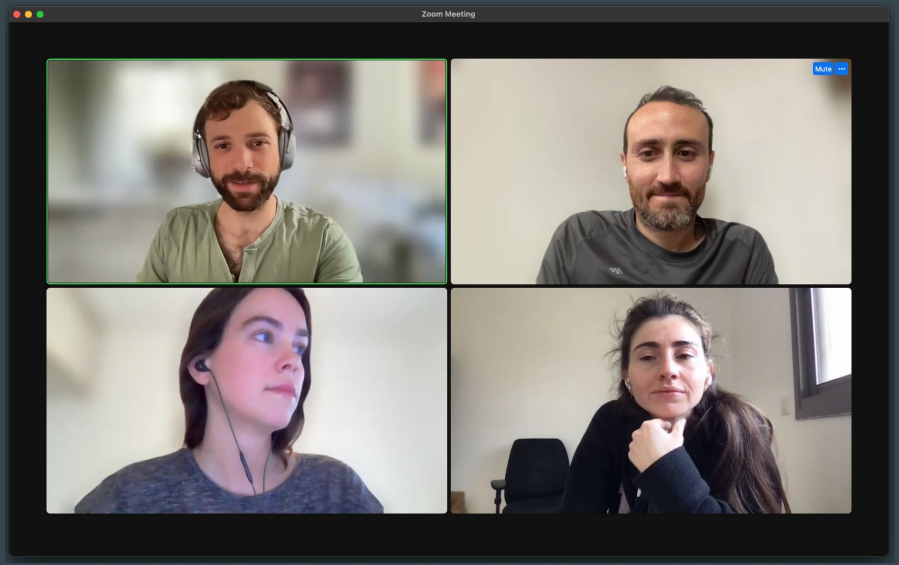
OUR PROCESS

Our Process

1. Started with reading the assignment
2. Got a bit scared :)
3. Went through the lectures, researched and gained confidence
4. Collected the relevant data
5. Got comfortable with BioPython
6. Data exploration
7. Researched more into GC-content, Spike lengths and additional terms
8. Wrapped the new vaccine 💉

And Of Course :)





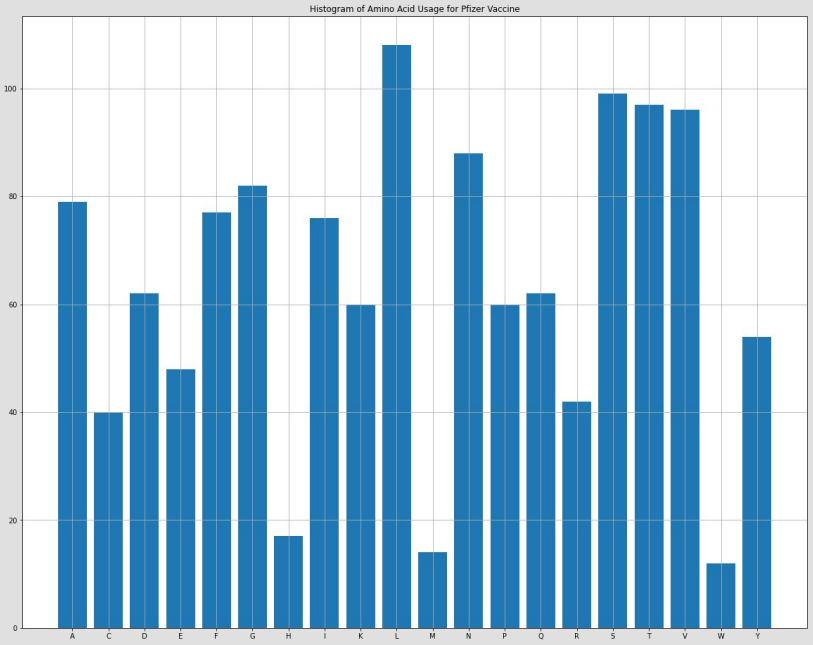
ASSIGNMENT TASKS

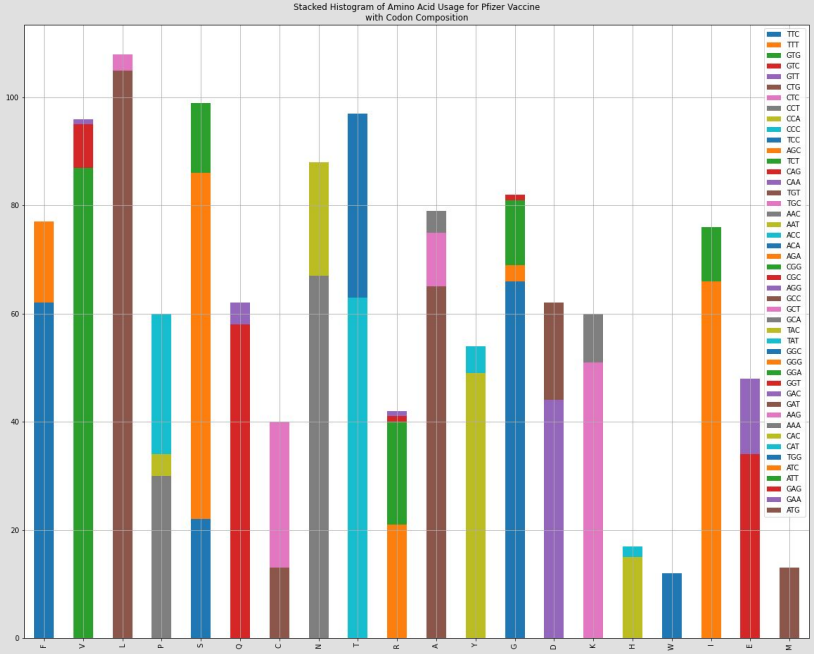
Part 1

Analyzing Pfizer and Moderna Vaccines

	Pfizer	Moderna
Initiation Codon	54	57
Termination Codon	3873 UGA	3876 UGA
Translated Length	1273	1273
Proteins	Identical protein level	

Pfizer Vaccine Distributions





Part 2

Comparison to the Wuhan Strain

Genome Length	29,903
S Gene Start	21,562
S Gene End	25,383

	Wuhan	Pfizer
GC-content	38%	57%
Protein Differences	985: K 986: V	985: P 986: P
Change Motivations	1. Higher G-C count converts more efficiently to proteins! 2. Proline substitution leads to a more rigid spike	

Wuhan Protein Sequence

Wuhan's protein's sequence is:

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQS
LLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIG
INITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFA
SVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPF
ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQ
TLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYT
MSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
SKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFN
SAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCG
KGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHT
SPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Wuhan AUGUUUGUUUUU...
Pfizer AUGUUCGUGUUC...

```
{'AAA': {'AAA', 'AAG'},
 'AAC': {'AAC', 'AAU'},
 'AAG': {'AAA', 'AAG'},
 'AAU': {'AAC', 'AAU'},
 'ACA': {'ACA', 'ACC'},
 'ACC': {'ACA', 'ACC'},
 'ACG': {'ACC'},
 'ACU': {'ACA', 'ACC'},
 'AGA': {'AGA', 'AGG', 'CGG'},
 'AGC': {'AGC', 'UCC'},
 'AGG': {'AGA', 'CGC', 'CGG'},
 'AGU': {'AGC', 'UCU'},
 'AUA': {'AUC', 'AUU'},
 'AUC': {'AUC', 'AUU'},
 'AUG': {'AUG'},
 'AUU': {'AUC', 'AUU'},
 'CAA': {'CAA', 'CAG'},
 'CAC': {'CAC', 'CAU'},
 'CAG': {'CAG'},
 'CAU': {'CAC', 'CAU'},
 'CCA': {'CCC', 'CCU'},
 'CCC': {'CCC', 'CCU'},
 'CCU': {'CCA', 'CCC', 'CCU'},
 'CGC': {'CGG'},
 'CGG': {'AGA', 'CGG'},
 'CGU': {'AGA', 'CGG'},
 'CUA': {'CUG'},
 'CUC': {'CUG'},
 'CUG': {'CUG'},
 'CUU': {'CUG'},
 'GAA': {'GAA', 'GAG'},
 'GAC': {'GAC', 'GAU'},
 'GAG': {'GAA', 'GAG'},
 'GAU': {'GAC', 'GAU'},
 'GCA': {'GCA', 'GCC', 'GCU'},
 'GCC': {'GCC', 'GCU'},
 'GCG': {'GCC'},
 'GCU': {'GCA', 'GCC', 'GCU'},
 'GGA': {'GGA', 'GGC', 'GGG'},
 'GGC': {'GGA', 'GGC'},
 'GGG': {'GGA', 'GGC'},
 'GGU': {'GGA', 'GGC', 'GGG', 'GGU'},
 'GUA': {'GUG'},
 'GUC': {'GUC', 'GUG', 'GUU'},
 'GUG': {'GUC', 'GUG'},
 'GUU': {'GUC', 'GUG'},
 'UAC': {'UAC', 'UAU'},
 'UAU': {'UAC', 'UAU'},
 'UCA': {'AGC', 'UCC', 'UCU'},
 'UCC': {'AGC', 'UCC'},
 'UCG': {'AGC', 'UCU'},
 'UCU': {'AGC', 'UCC', 'UCU'},
 'UGC': {'UGC', 'UGU'},
 'UGG': {'UGG'},
 'UGU': {'UGC', 'UGU'},
 'UUA': {'CUC', 'CUG'},
 'UUC': {'UUC', 'UUU'},
 'UUG': {'CUC', 'CUG'},
```

Part 3

BA.2 Protein Extraction

The Path to the Spike:

- First we tried to find the BA.2 spike by aligning the entire BA.2 seq against the Wuhan spike seq.
- Then we decided to search on the translated seq instead of RNA/DNA.
- We searched for proteins that match the Wuhan protein (and spike protein in general) length.
- We used BioPython align to search the protein with the best align score for each record.
- We searched for the most common spike.

Part 3

BA.2 Protein Extraction

Final Method:

1. Translate mRNA sequence.
2. Align each protein sequence with the Wuhan spike.
3. Choose protein sequence with highest alignment score.
4. Repeat on multiple BA.2 sequences.
5. Choose the most common sequence.

Alignment scores calculated using BioPython
`pairwise2.align.globalxx`

Part 3

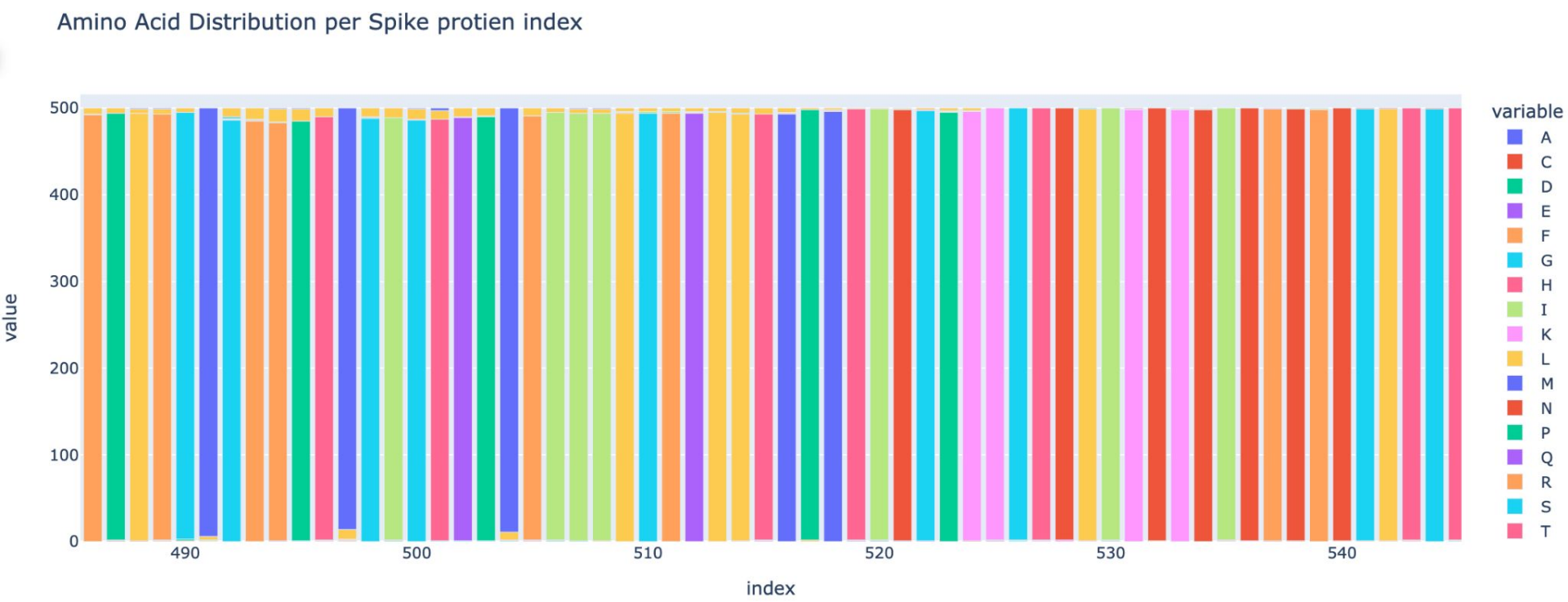
BA.2 Protein Extraction

Observations:

- We eventually decided to use the alignment score as it's a better measurement than length.
- We discovered that just by filtering proteins with the proper length we found the spike.
- We discovered shifts in the location of the spike protein between different BA.2 records.
- We also saw different variations of the spike protein between different records.

Alignment scores calculated using BioPython
`pairwise2.align.globalxx`

Comparison of BA.2 S Gene Amino-Acids



Part 4

Building the mRNA Vaccine

Method:

1. Use start-codon index in BA.2 S gene protein to extract S gene mRNA sequence
2. Use Pfizer lookup table to determine BA.2 codon mapping
3. Switch KV->PP based on index
4. Add 5’UTR and 3’UTR from Pfizer vaccine

UBI GROUP BA.2 VACCINE

GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGUUUGUCUUCCUGGUGCUGCUGCCUCUGGUCUCCAGCCAGUGCGUCAAUCUGAUCACCCGGACACAAUCCUACACCAAUUCCUU
CACACGGGGUGUCUAUUACCCUGAUAAGGUCUUCAGAAGCAGCGUGCUCCAUUCCACACAGGACCUGUUCCUCCCCUUCUUUUCCAAUGUCACCUGGUUUCACGCAAUCCACGUCAGCGGAACCAACGGCACCAAAAGAU
UCGACAACCCUGUUCUGCCUUUUAAUGAUGGUGUCUAUUUUGCUUCCACAGAGAAGUCUAACAUCAUCAGAGGAUGGAUUUUCGGCACCACCCUCGACAGCAAAACACAGUCCCUGCUGAUUGUGAACAACGCUACCAAC
GUCGUCAUCAAGGUGUGUGAGUUCCAGUUCUGCAAUGAUCCCUUUCUCGACGUGUAUUACCAUAAGAACAAUAAAAGCUGGAUGGAAUCUGAAUUCAGAGUCUAUUCUUCUGCCAAUAACUGUACCUUUGAGUAUGUUUC
UCAGCCCUUCCUGAUGGAUCUGGAAGGGAAGCAGGGAAAUUUUAAGAAUCUGAGAGAAUUUGUCUUCAAAAAUAUCGAUGGGUAUUUCAAGAUUUAUUCCAAACACACCCCCAUUAACCUGGGACGGGAUCUGCCACAGG
GUUUCUCUGCUCUCGAGCCCCUCGUGGAUCUGCCUAUCGGCAUCAAUAUUACCCGCUUUCAGACCCUGCUGGCCCUCCAUAGGAGCUAUCUGACACCAGGUGAUUCUAGCAGCGGUUGGACAGCCGGUGCCGCCGCAUAU
UAUGUCGGAUACCUGCAACCACGCACAUUCCUGCUCAAAUAUAAUGAAAACGGAACAAUCACAGACGCCGUGGAUUGCGCACUGGAUCCACUGUCUGAGACCAAAUGUACCCUGAAAAGCUUCACAGUGGAAAAGGGCAU
CUACCAGACCUCUAACUUUAGAGUCCAACCUACCGAGUCCAUCGUCAGGUUUCCAAAUAUUACAAACCUCUGUCCCUUCGACGAGGUGUUCAAUGCCACAAGAUUUGCAUCUGUGUAUGCCUGGAAUAGAAAAAGGAUUU
CCAAUUGCGUCGCAGACUACUCCGUGCUGUAUAAUUUUGCACCUUUCUUUGCCUUUAAGUGUUACGGCGUGAGCCCAACAAAGCUGAACGAUCUGUGCUUCACAAACGUGUACGCUGAUUCCUUCGUGAUCCGGGGCAAU
GAGGUCUCCCAGAUCGCUCCCGGACAGACCGGAAACAUCGCCGACUAUAAUUAUAAGCUGCCUGACGACUUUACCGGCUGUGUCAUCGCAUGGAAUUCCAACAAACUGGAUUCCAAAGUCGGCGGGAAUUACAAUUAUCU
GUACCGGCUGUUCAGAAAAAGCAAUCUGAAGCCCUUCGAGAGGGAUAUUAGCACCGAAAUCUACCAGGCCGGGAAUAAGCCCUGUAACGGCGUGGCUGGCUUUAACUGCUAUUUCCCACUCCGAUCCUACGGUUUCCGAC
CUACAUACGGCGUGGGACAUCAGCCCUACCGGGUGGUGGUGCUGUCCUUUGAACUGCUGCACGCUCCUGCUACAGUCUGCGGCCCAAAGAAGAGCACCAAUCUGGUGAAAAACAAAUGUGUUAACUUUAAUUUCAACGGG
CUGACAGGAACAGGUGUGCUGACAGAGUCUAAUAAAAAGUUCCUGCCAUUUCAGCAAUUUGGAAGAGACAUCGCCGACACAACCGACGCCGUGAGAGACCCCCAGACCCUGGAAAUUCUGGACAUUACCCCCUGCUCUUU
CGGGGGGGUCUCUGUCAUCACCCCCGGCACAAACACCAGCAACCAGGUGGCUGUCCUGUAUCAGGGGGUCAACUGUACCGAAGUUCCUGUGGCUAUUCAUGCUGACCAGCUGACCCCUACCUGGCGGGUCUACUCUACCG
GCAGCAACGUGUUCCAAACAAGAGCUGGAUGCCUCAUUGGCGCCGAAUACGUCAACAAUUCUUACGAAUGCGAUAUCCCCAUCGGGGCCGGAAUCUGCGCCAGCUAUCAGACCCAGACCAAAAGCCAUCGGAGAGCUCGG
UCUGUGGCAAGCCAAUCCAUCAUCGCUUACACCAUGUCCCUGGGUGCAGAGAAUUCCGUGGCUUAUUCUAACAACAGCAUUGCUAUCCCCACCAAUUUCACCAUCUCUGUCACCACCGAGAUUCUGCCUGUGUCUAUGAC
CAAAACCAGCGUGGAUUGUACAAUGUACAUUUGUGGGGACAGCACCGAGUGCUCCAACCUGCUGCUCCAGUACGGAAGCUUUUGUACCCAACUGAAAAGAGCCCUCACAGGAAUCGCAGUGGAGCAAGAUAAGAACACAC
AAGAAGUGUUCGCUCAAGUCAAACAAAUUUAUAAAACACCUCCUAUUAAGUACUUCGGGGGGUUCAAUUUCUCUCAAAUCCUGCCUGACCCCUCCAAACCUUCCAAGCGCUCCUUUAUUGAAGACCUGCUGUUCAAUAAA
GUGACACUGGCUGACGCAGGAUUUAUCAAGCAGUAUGGUGAUUGUCUGGGCGACAUUGCCGCCAGGGAUCUGAUCUGUGCUCAGAAAUUUAAUGGACUGACCGUCCUGCCCCCCCUGCUGACCGAUGAAAUGAUUGCCCA
GUAUACAAGCGCACUGCUGGCCGGCACCAUUACCUCCGGGUGGACAUUUGGGGCUGGUGCCGCUCUCCAGAUCCCUUUCGCUAUGCAGAUGGCCUACAGAUUUAACGGUAUUGGGGUCACCCAGAACGUCCUGUAUGAAA
AUCAAAAGCUGAUCGCCAAUCAAUUCAAUUCUGCUAUCGGCAAAAUUCAGGAUAGCCUGUCUAGCACAGCAAGCGCACUGGGCAAACUGCAGGAUGUGGUGAAUCACAACGCACAGGCACUCAACACCCUGGUCAAGCAG
CUGAGCUCCAAAUUUGGAGCUAUUAGCUCUGUGCUCAAUGACAUUCUGAGCAGACUGGAUAAACCUCCUGCUGAAGUGCAGAUCGACCGGCUCAUCACCGGAAGGCUGCAAUCUCUGCAGACCUACGUGACCCAGCAACU
GAUCAGGGCAGCUGAAAUUCGGGCAUCCGCUAACCUGGCUGCAACCAAGAUGUCCGAAUGCGUGCUGGGGCAGUCCAAGAGAGUCGACUUUUGCGGAAAAGGCUAUCAUCUGAUGAGCUUCCCUCAGUCCGCACCCCACG
GAGUGGUUUUUCUGCACGUGACCUACGUUCCCGCCCAGGAGAAGAAUUUUACAACCGCACCUGCCAUUUGUCACGACGGCAAGGCACACUUUCCUCGGGAAGGAGUGUUCGUGAGCAACGGCACACACUGGUUCGUGACC
CAGAGAAACUUUUACGAGCCCCAGAUCAUUACAACAGAUAAUACAUUCGUGAGCGGGAAUUGCGACGUCGUGAUCGGGAUUGUCAAUAACACAGUGUAUGAUCCCCUGCAGCCUGAGCUCGACUCUUUCAAAGAAGAGCU
CGAUAAGUAUUUCAAAAACCACACAUCUCCCGAUGUCGAUCUGGGAGACAUUUCUGGAAUCAACGCCUCCGUCGUGAACAUUCAAAAAGAGAUUGACCGGCUGAAUGAAGUGGCUAAAAACCUCAAUGAAUCUCUGAUUG
ACCUGCAAGAACUGGGCAAGUACGAGCAGUACAUUAAAUGGCCUUGGUAUAUUUGGCUGGGUUUUAUUGCCGGACUGAUCGCCAUUGUGAUGGUCACCAUUAUGCUGUGCUGCAUGACCAGCUGCUGCAGCUGCCUGAAA
GGAUGCUGUUCUUGUGGCUCCUGUUGCAAGUUUGACGAGGACGAUUCUGAGCCUGUGCUGAAGGGAGUCAAACUGCACUACACCUGAUGACUCGAGCUGGUACUGCAUGCACGCAAUGCUAGCUGCCCCUUUCCCGUCCU
GGGUACCCCGAGUCUCCCCCGACCUCGGGUCCCAGGUAUGCUCCCACCUCCACCUGCCCCACUCACCACCUCUGCUAGUUCCAGACACCUCCCAAGCACGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCC
ACGGGAAACAGCAGUGAUUAACCUUUAGCAAUAAACGAAAGUUUAACUAAGCUAUACUAACCCCAGGGUUGGUCAAUUUCGUGCCAGCCACACCCUGGAGCUAGCA

KEY TAKEAWAYS

Key Takeaways

1. Pfizer's lookup table is partially understandable (e.g GC-content maximization), but there is still a multiple-options dilemma
2. Solution based on large sample and choosing most common protein sequence
3. Small changes can be make a huge difference and be very impactful (PP, GC)
4. The fact that extremely valuable sequence databases (commercial vaccines included!) are open-source and available is mind-blowing and powerful

REFERENCES

References

- https://www.genome.gov/about-genomics/fact-sheets/COVID-19-mRNA-Vaccine-Production
- https://www.ncbi.nlm.nih.gov/nuccore/1798174254
- https://en.wikipedia.org/wiki/GC-content
- https://berthub.eu/articles/posts/reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5584442/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1463026/
- https://journals.asm.org/doi/10.1128/mBio.02648-20
- https://www.nature.com/articles/s41586-021-03275-y
- https://www.news-medical.net/health/What-are-Spike-Proteins.aspx

THANK YOU 🙏