mRNA Sequence Design Using Optimization Techniques

Sarah Cross

Keystone High School

Alamo Regional Science & Engineering Fair

February 20-22, 2021

Abstract

Recently, the treatment and prevention of disease has become dominated by a relatively new and revolutionary technique: mRNA. Although initially used to prevent cancer, its potential applications in the medical industry have expanded drastically, as it has massive advantages over conventional DNA techniques such as LAV(live attenuated viruses) by being much faster and safer to develop, providing precision in antigen design, and allowing researchers the ability to adapt vaccines to virus variations quickly and to synthesize fully customizable treatments for patients.

Technology is now at the point where we have the ability to form mRNA vaccine candidates using computational biology models, creating full working prototypes before testing in the lab. By applying state-of-the-art techniques, bioinformaticians can encode for specific antigens such as the SARS-CoV-2 spike glycoprotein and compute vaccines for a virus and its potential mutations. A process that once required years of work and a lab now can be solved computationally in a matter of days.

Constrained discrete optimization will be used to minimize a fitness function balancing both GC(guanine-cytosine) content and utilization of common codons in the human body to maximize mRNA levels and the formation of antibodies in the cell. The researcher will show that these results come very close to encoding for the BNT162b2 vaccine, with a 78.1% codon similarity and 90.7% nucleotide similarity. This discrete optimization method will be compared against those produced by that of a widely accepted codon mapping approach; however, the approach illustrated here enables greater levels of customization in vaccine design.

Table of Contents

Introduction

Past vaccines for infectious diseases came from the process of LAV, or live attenuated vaccines, in which disease-causing pathogens were weakened under laboratory conditions; however, this process is slow, time extensive, can have many different limitations, and can pose a danger to immunocompromised people. mRNA vaccines provide a more efficient technique that saves money, could be applied to many different illnesses and infectious diseases, is much safer, faster to implement, customizable, provides precision in antigen design, good tolerability, and broad immune responses with a highly scalable manufacturing platform. There are two forms of mRNA vaccines: conventional and self-amplifying. Conventional mRNA vaccines, such as the Pfizer vaccine, contain nucleoside-modified mRNA which are modified to force an immune response from the cell. Self-amplifying mRNA are derived from positive strand RNA viruses that can be directly translated into proteins rather than viral RNA complementary to viral mRNA. It amplifies vaccine-encoding transcripts, resulting in higher antigen levels and inciting the immune response to form antibodies. Utilizing discrete optimization, a method with promising results in non-differential problems and the mRNA sequence design field, we will encode for promising antigens within several different viruses, taking into account codon optimization, GC content, the frequency of codons produced in the human body, hairpin structures, and modified nucleosides. The first trial will involve utilizing already-known solutions to optimize the nucleotide sequence within the BNT162b2 vaccine, also known as the Pfizer vaccine encoding the perfusion stabilized membrane-anchored SARS-CoV-2 full length spike such as simple codon mapping to increase the GC content. The second trial will implement a discrete optimization algorithm to find the best fitness of the non-differentiable function measuring GC content as well as codon rarity, and compare its performance with that of the true BNT162b2 vaccine nucleotide sequence, codon sequence, and protein sequence. The third trial will apply the most effective tool to different viruses and visualize the results. Finally, if the following trials elicit positive results, a fourth trial will be conducted to design the mRNA sequence using self-replicating techniques deriving from self-replicating single-stranded RNA viruses such as the Sindbis virus, the Semliki Forest virus, or the Kunjin virus.

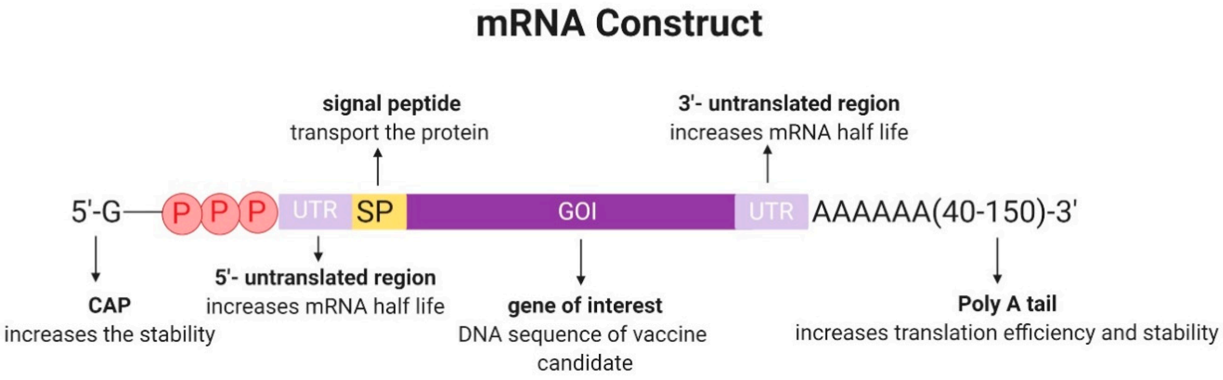
Background

The Pfizer vaccine contains tozinamaran, and is injected into the body by means of an intramuscular injection. The vaccine contains nucleoside-modified mRNA, and must be stored between -90 and -60 degrees Celsius until 5 days before vaccination because mRNA is much more fragile than DNA as a single-stranded nucleic acid; however, there have been recent attempts to create thermostable vaccine using freeze-dried mRNA with trehalose or naked mRNA, which in clinical trials expressed high levels of proteins and was associated with higher immunity levels in newborns and the elderly, or using protamine-encapsulated vaccine with oscillating temperatures between 4 and 56 degrees Celsius, which appeared not to change levels of immunogenicity, or using cationic liposome and cell penetrating peptide(CPP) tools, which protected mRNA from degradation by RNase. The Pfizer vaccine is made of mRNA, lipids to protect the mRNA and provide a “slippery” exterior to help mRNA enter cells(((4-hydroxybutyl)azanediyl)bis, (hexane-6,1-diyl)bis(2-hexyldecanoate), 2 [(polyethylene glycol)-2000]-N, N-tetradecyl acetamide, 1,2-Distearoyl-sn-glycero-3-phosphocholine, and cholesterol), salts to balance the acidity within the body(potassium chloride, monobasic potassium phosphate, sodium chloride, dibasic sodium phosphate dihydrate, and sucrose). The Pfizer vaccine studies were multinational, placebo-controlled, observer-blinded, pivotal efficacy trials without people under the age of 16 or with immuno-compromising conditions.

On the tenth of January, 2020, the full SARS CoV-2 genome was released by the Chinese Center for Disease Control and Prevention, with one feature noted within the GenBank file: the S structural glycoprotein. This structural protein has been the target of the majority of vaccines being currently developed, as it is the method by which the RNA sequence of the virus manages to make its way into the cells, be read by the ribosomes, assembled into proteins, and subsequently take over the cell. The Pfizer vaccine, also known as BNT162b1, encodes secreted trimerized SARS-CoV-2 receptor binding domain, a part of the virus on the spike domain that allows the virus access into the cell. By placing messenger RNA by intramuscular injection into the body, the mRNA is accepted into cells and forms copious amounts of the spike protein, forcing the immune system to recognize the spike as a threat and form antibodies to suppress the spike protein. Although the mRNA simulates the same shape and structure of the spike protein, it is not identical to the true spike protein; it contains a higher percentage of guanine and cytosine, two nucleotides which in past observations have proven to raise mRNA levels in mammalian cells, and it also contains two different protein substitutions to Proline such that the S glycoprotein can keep the same structure despite not being connected to the entire virus(without the protein substitutions, the vaccine collapses). Surrounding the mRNA are lipid nanoparticles, chosen primarily because they are good carriers, protect the mRNA construct as it travels through the bloodstream, and helps it cross cell membranes and get from the bloodstream into the cell. While DNA offers higher stability, RNA can enter and encounter its site of action significantly faster, as it simply needs to cross the cell membrane, while DNA must cross both the cell membrane and the nucleus; DNA vaccines also offer their own dangers as well, as they can be mistakenly incorporated into the cell’s own genome. RNA never enters the nucleus, and it is too fragile to remain in the cell when the cell is being replicated, so mRNA has far less long-term consequences on the cell other than helping T and B cells form antibodies. The vaccine starts producing SARS-CoV-2 Spike proteins in large enough quantities that the immune system immediately is forced to respond, and gives it enough signs that the cells have been taken over through the start cap GA, tricking the ribosome into thinking the mRNA is coming from the nucleus, and replacing the nucleotide Uracil with Pseudouridine which calms the immune system while being accepted as Uracil in relevant parts of the cell.

However, there is potential for mutations in the SARS-CoV-2 genome, in which case mRNA solutions become more crucial than ever before as they are capable of being adaptive and computationally designed very quickly. The Spike N501 substitution, for example, poses a major problem, as it is a mutation within the S glycoprotein and is located in the viral receptor binding site for cell entry, which increases the binding to the receptor, angiotensin converting enzyme 2, which is an enzyme attached to the cell membranes of cell located in the lungs, heart, kidney, and intestines.

Conventional mRNA vaccines are formed by the cap(a header to inform the ribosomes that the mRNA came from the nucleus in order to trick the ribosome, which takes mRNA and emits a strand of amino acids which fold into proteins, to replicate the sequence), a 5’ UTR(untranslated region), or the five-prime beginning that increases the stability of the mRNA vaccine, the gene of interest encoded and purified to increase translation efficiency and the chance of the immune response responding and forming antibodies, a 3’ UTR(which increases mRNA half life), and a Poly A tail containing 40-150 Adenosine nucleotides to increase translation efficiency and stability(as mRNA degrades very quickly, it preserves the end codon from being degraded and ensures that the mRNA simply produces the vaccine properly). The UTRs give the ribosomes metadata about how much translation should happen and how much; although UTRs are still relatively undiscovered, researchers have used machine learning algorithms to help select the best preexisting UTRs to select for vaccine construction. In the case of the Pfizer vaccine, a slightly modified alpha globin gene was used for the UTRs. In addition, researchers have found that replacing nucleotide bases with molecularly modified versions of the nucleotide base; for example, Ψ, or Pseudouridine, is used in the Pfizer vaccine as a replacement to Uracil, working to calm the immune system while also being accepted as a "U" in relevant parts of the system.



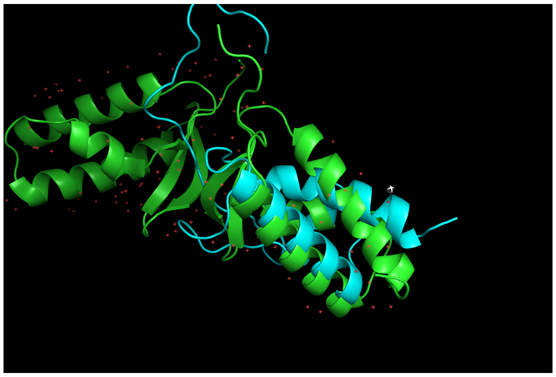
mRNA Vaccine Construction

Source:

mdpi.com/vaccines/vaccines-07-00122/article_deploy/html/images/vaccines-07-00122-g002.png

Ultimately, the main goal of the vaccine is to teach the immune system what the viral proteins that hijack the cell look like so as to allow the immune system to form antibodies and vigorously attack the actual virus if it were to enter the body. The vaccine makes the ribosome form copious amounts of the protein, forcing the immune system to respond to the threat by forming antibodies.

AlphaFold is a machine learning tool which can solve the infinite folding protein problem using neural networks. Because functions of proteins are defined by their 3D structures(for example, Spike protein's structure allows virus DNA/RNA to enter cells), it is essential to come up with a computational model to simulate this process. Previously, this process required a lab and the utilization of x-ray crystallography, nuclear magnetic resonance, or cryo-electron microscopy, all of which are both financially and time expensive. Another relatively new type of solution to the protein folding problem is observing structures of proteins with similar amino acid sequences and referencing already known structures using MSA; however, there are still many different ways in which those clusters of proteins can fit together, and typically this process must be undertaken by supercomputers. AlphaFold makes folding proteins quick, inexpensive, and highly accessible to the general public.



Blue is the predicted structure for a piece of the green protein(PDB ID 5chy) using AlphaFold.

Materials

- Computer - the researcher used an HP ENVY Laptop 13-aq0044nr with an Intel Core i7 processor and NVIDIA GeForce MX250 GPU; running VMD for rendering requires a pentium class machine with at least 512MB of RAM and a 16-bit color video, although the full-featured graphics-enabled mode of VMD used in this project requires an OpenGL-capable graphics accelerator with up-to-date drivers.

- OS - The researcher tested and ran the project on both Ubuntu 20.04 and Windows Home 10.0.19042 Build 19042 with Python 3.7.9 without issues.

- Libraries(can be installed by **py -m pip install** in Windows, or **pip3 install** in Ubuntu):

  - BioPython and SeqIO(to read GenBank and Fasta files)

  - json(a built-in Python library to process .json files)

  - absl(to input customizable file paths into the program)

  - matplotlib(to graph the fitness function and GC content over the genomic sequence)

Note: If being run on Linux or on WSL, the repository contains a bash file, configure.sh, to quickly install these libraries on Python 3. The bash file can be run within the directory on the CLI by **sh ./configure.sh**.

Procedure

1) Implement the codon mapping technique and apply it to the S glycoprotein spike of SARS-CoV-2 genome.

2) Implement the discrete optimization algorithm, using nucleotide, codon, protein, and GC content comparison as a fitness measurement and compare its performance to that of the true BNT162b2 vaccine as well as the first technique.

3) The class OptimizeSeq creates the following functions:

__init__(seq) - Initializes and inputs a sequence(the section of the virus we are encoding)

change(val) - Changes the selected sequence to val

gc_content() - Calculates the GC content of the particular sequence inputted(self.seq)

codon_freq() - Iterates through codons within the sequence inputted, finds their frequency in constant codonS(a lookup table of codon frequencies in the human body), and calculates the average codon frequency for the sequence.

fitness(alpha, expected_GC) - Calculates the fitness(or loss function) of the sequence through multiple combinations of the gc_content and codon_freq functions(as described in the table below). Alpha is a constant used in measuring the importance of GC content in the fitness function in relation to codon frequency.

discrete_descent() - Use discrete gradient descent to increase fitness. This method can differ based on the version being implemented, although in general this function iterates through the codons within the sequence, maps them to their particular amino acid, finds all possible codons that could encode for that particular amino acid through a lookup table(revCodonAmino), iterates through those codons and changes the sequence, calculates the loss function of the new sequence, and finds the best codon to change.

print_info(include header) - Print loss(fitness), GC content(gc_content), and amino acid frequency calcs(codon_freq) from the current sequence.

The following versions of the discrete optimization algorithm were implemented:

Discrete Optimization Version	Concept	Loss Function
0	For each codon, find the best optimized codon.	$\alpha * gcCont + (1 - \alpha) * codonFre$
1	Measures fitness within a specific frame(ex. of size 12), and for each frame finds the best codon to change.	$\alpha * gcCont + (1 - \alpha) * codonFre$
2	Measures loss for the entire sequence and finds the best codon to change.	$\alpha * \lvert d - gcCont \rvert + (1 - \alpha) * coc$
3	Same concept as Version 2	$e^{-(d-gcCont)^2/v^2} * codonFreq$

d = Desired GC content

v = Variation of Normal distribution

3) Apply the most effective tool to different viruses and render the results using the 3dRNA web server(http://biophy.hust.edu.cn/new/3dRNA/create), matplotlib to graph GC content and fitness reduction across iterations, and tables to compare the most effective parameters for the discrete optimization algorithm.

3a) For versions 0-2, adjust the values of αusing gradient descent and graphing to find the most optimized value of $\alpha$by which to weigh the gc_content and codon_freq functions(by graphing the true vaccine's fitness function and finding the α value that results in the smallest loss).

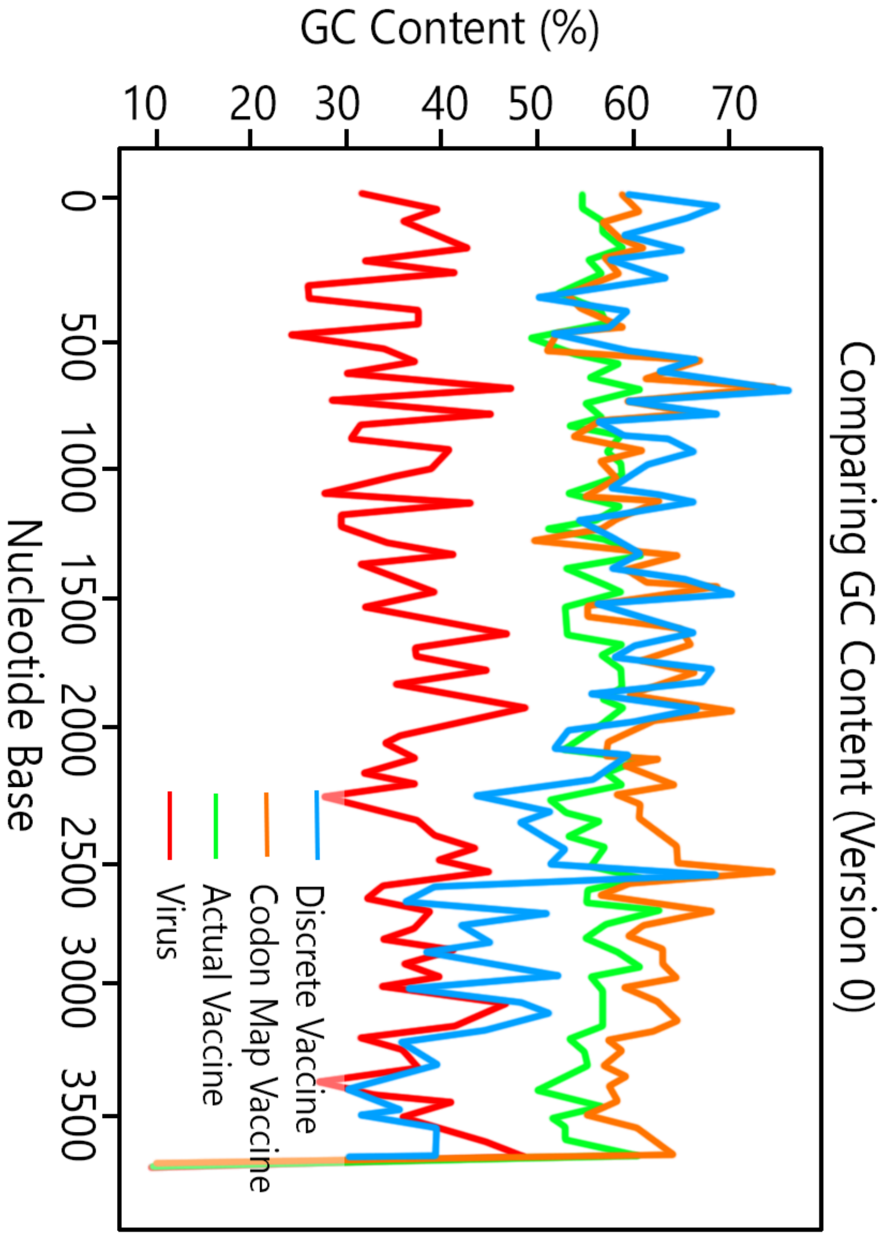
3b) For version 3, run different values of *v*, and find the best variation to result in the closest possible estimation to the true vaccine.

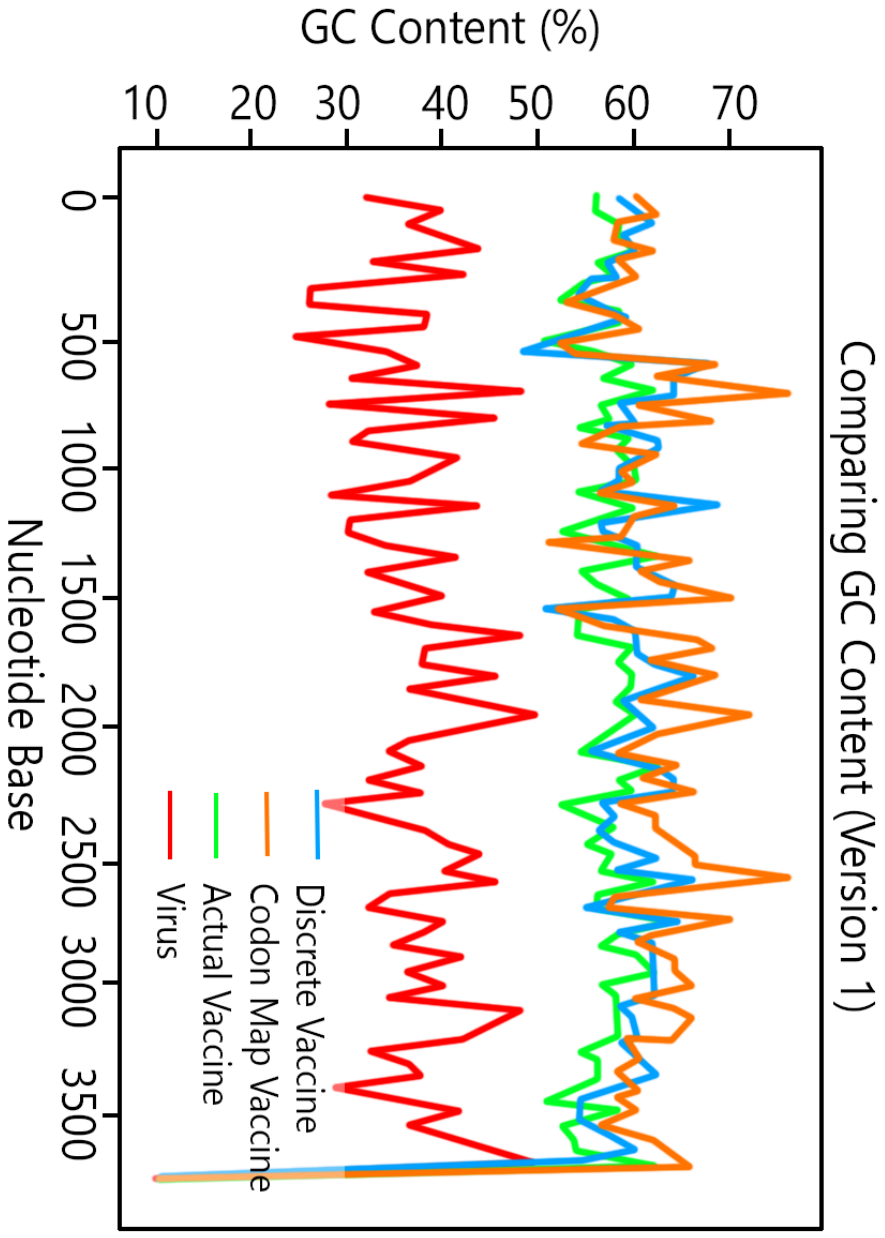
Data and Results

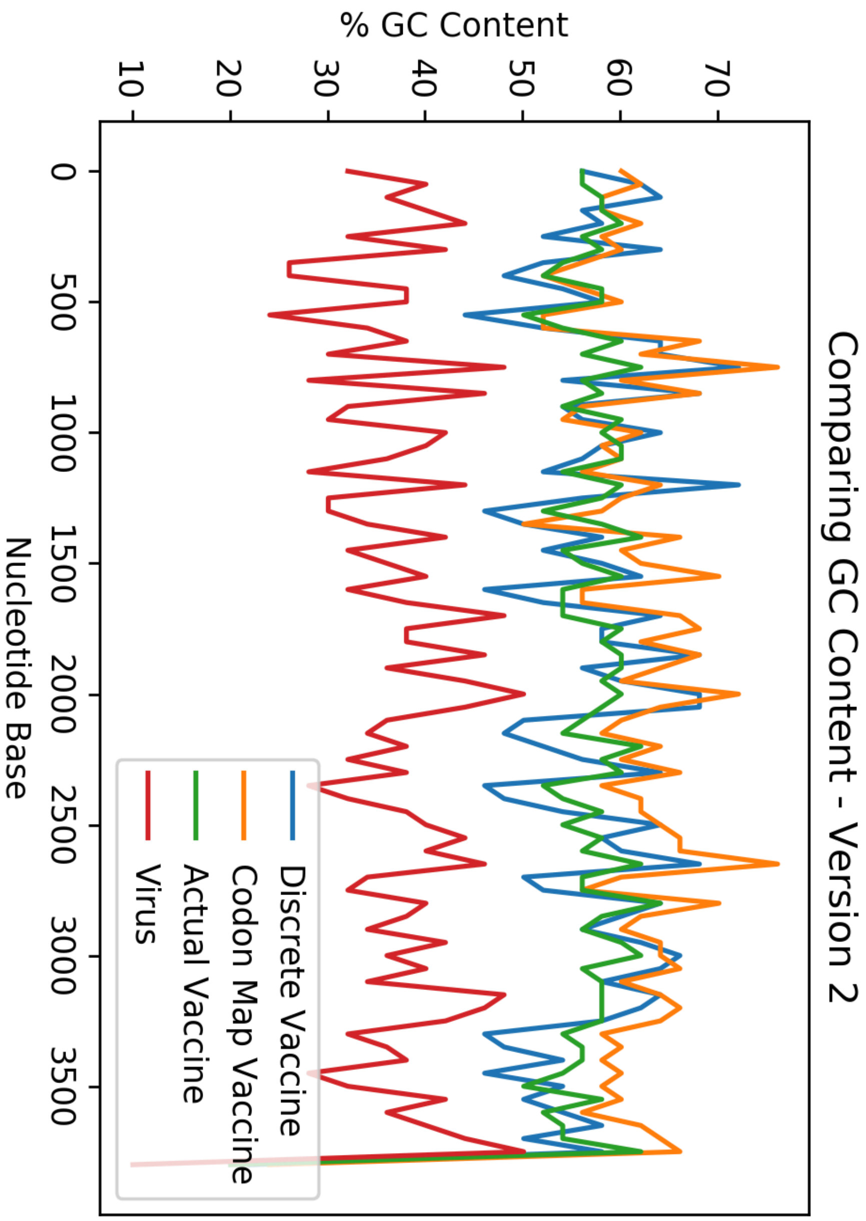
- Codon mapping - Mapping the codon values to a lookup table. Simple, quick, and high nucleotide and codon frequency matches.
- Discrete optimization version 0:
    - Iterates through and for each codon finds the best optimized codon.
    - Problems with high GC content at beginning and cutting down at end.
    - Slow, relatively good results but not state-of-the-art.
- Discrete optimization version 1:
    - Iterates through, measures fitness within a specific frame(size 12), and for each codon finds the best optimized codon.
    - GC content can get extremely close(within 0.1%) to the actual vaccine, at cost of major nucleotide and codon differences.
    - Fixes GC problems slightly(although sometimes avg GC content within a specific area might dip, this fixes it so it's not indicative of real vaccine)
- Discrete optimization version 2:
    - Iterates through, measures fitness for the entire sequence and finds the best codon to change.
    - Very good results - High nucleotide and codon % matches
    - Also high GC % and codon frequency %
    - Much slower than versions 0 and 1
- Discrete optimization version 3:
    - Same as version 2, but optimizes fitness function(see page 13 for fitness function)
    - Converges slightly faster
    - Fitness function normalized and doesn't require alpha value(which is a constant that isn't guaranteed to be the same across different viruses)

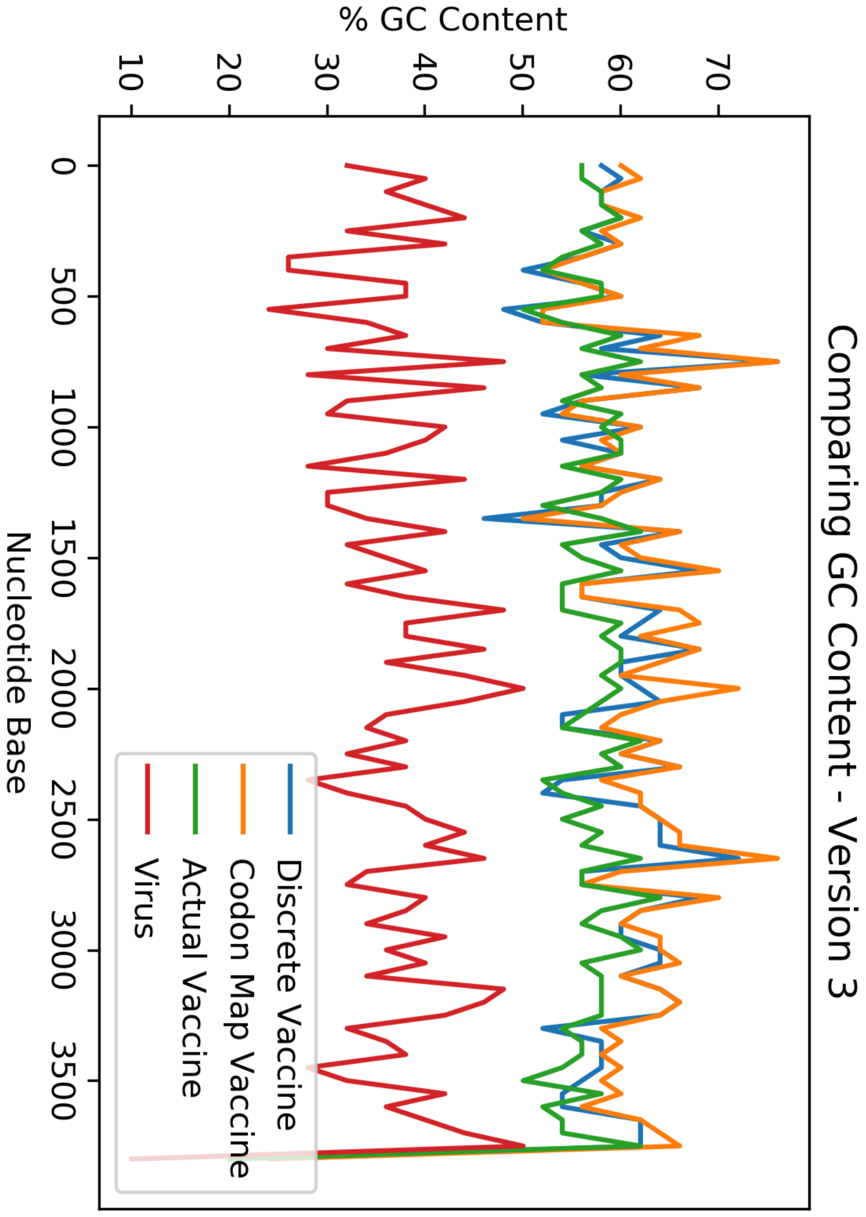
Comparison of Methods to Vaccine

Name of Method	Nucleotide % Similarity	Codon % Similarity	GC Content % (True GC Content 56.99%)	Codon Frequency % (True Codon Frequency 14.73%)
Original Virus	72%	28%	37.31%	11.31%
Codon Mapping	91%	79%	61.43%	15.32%
Version 0	71%	27%	54.47%	10.06%
Version 1	76%	40%	63.71%	11.92%
Version 2	88.4%	71.9%	57.01%	14.95%
Version 3(v=0.2)	87.6%	74.57%	57.67%	15.19%
Version 3(v=0.4)	**90.7%**	**78.1%**	**59.63%**	**15.29%**


Comparing GC Content (Version 0)
GC Content (%)
10
20
30
40
50
60
70
Nucleotide Base
0
500
1000
1500
2000
2500
3000
3500
Discrete Vaccine
Codon Map Vaccine
Actual Vaccine
Virus


Comparing GC Content (Version 1)
GC Content (%)
10
20
30
40
50
60
70
Nucleotide Base
0
500
1000
1500
2000
2500
3000
3500
Discrete Vaccine
Codon Map Vaccine
Actual Vaccine
Virus


Comparing GC Content - Version 2
% GC Content
10
20
30
40
50
60
70
Nucleotide Base
0
500
1000
1500
2000
2500
3000
3500
Discrete Vaccine
Codon Map Vaccine
Actual Vaccine
Virus


Comparing GC Content - Version 3
% GC Content
10
20
30
40
50
60
70
0
500
1000
1500
2000
2500
3000
3500
Nucleotide Base
Discrete Vaccine
Codon Map Vaccine
Actual Vaccine
Virus

CLI Sample Output

Format:

Amino acid of Vaccine

Vaccine sequence

Differences in sequence

Predicted Vaccine Sequence

Amino acid of Predicted Vaccine Sequence

```
 T E V P V A I H A D Q L T P T W

TACCGAAGTGCCCGTGGCCATTCACGCCGATCAGCTGACACCTACATGGC

!     !     !        !        !        !     !   !

CACCGAGGTGCCTGTGGCCATCCACGCCGACCAGCTGACCCCTACCTGGA

 T E V P V A I H A D Q L T P T W


R V Y S T G S N V F Q T R A G C L

GGGTGTACTCCACCGGCAGCAATGTGTTTCAGACCAGAGCCGGCTGTCTG

 !        !      !!!   !    !                 !

GAGTGTACTCTACCGGCTCTAACGTGTTCCAGACCAGAGCCGGCTGCCTG

R V Y S T G S N V F Q T R A G C L


 I G A E H V N N S Y E C D I P I G

ATCGGAGCCGAGCACGTGAACAATAGCTACGAGTGCGACATCCCCATCGG

     !                !                   !
```

```
ATCGGCGCCGAGCACGTGAACAACAGCTACGAGTGCGACATCCCTATCGG

 I G A E H V N N S Y E C D I P I G


  A G I C A S Y Q T Q T N S P R R

CGCTGGAATCTGCGCCAGCTACCAGACACAGACAAACAGCCCTCGGAGAG

  ! !            !   !  !!!  ! !

CGCCGGCATCTGCGCCAGCTACCAGACCCAGACCAACTCTCCTAGAAGAG

  A G I C A S Y Q T Q T N S P R R


A R S V A S Q S I I A Y T M S L G

CCAGAAGCGTGGCCAGCCAGAGCATCATTGCCTACACAATGTCTCTGGGC

                 !     !  !!!

CCAGAAGCGTGGCCAGCCAGAGCATCATCGCCTACACCATGAGCCTGGGC

A R S V A S Q S I I A Y T M S L G


 A E N S V A Y S N N S I A I P T N

GCCGAGAACAGCGTGGCCTACTCCAACAACTCTATCGCTATCCCCACCAA

              !         !    !

GCCGAGAACAGCGTGGCCTACTCTAACAACTCTATCGCCATCCCTACCAA

 A E N S V A Y S N N S I A I P T N


  F T I S V T T E I L P V S M T K
```

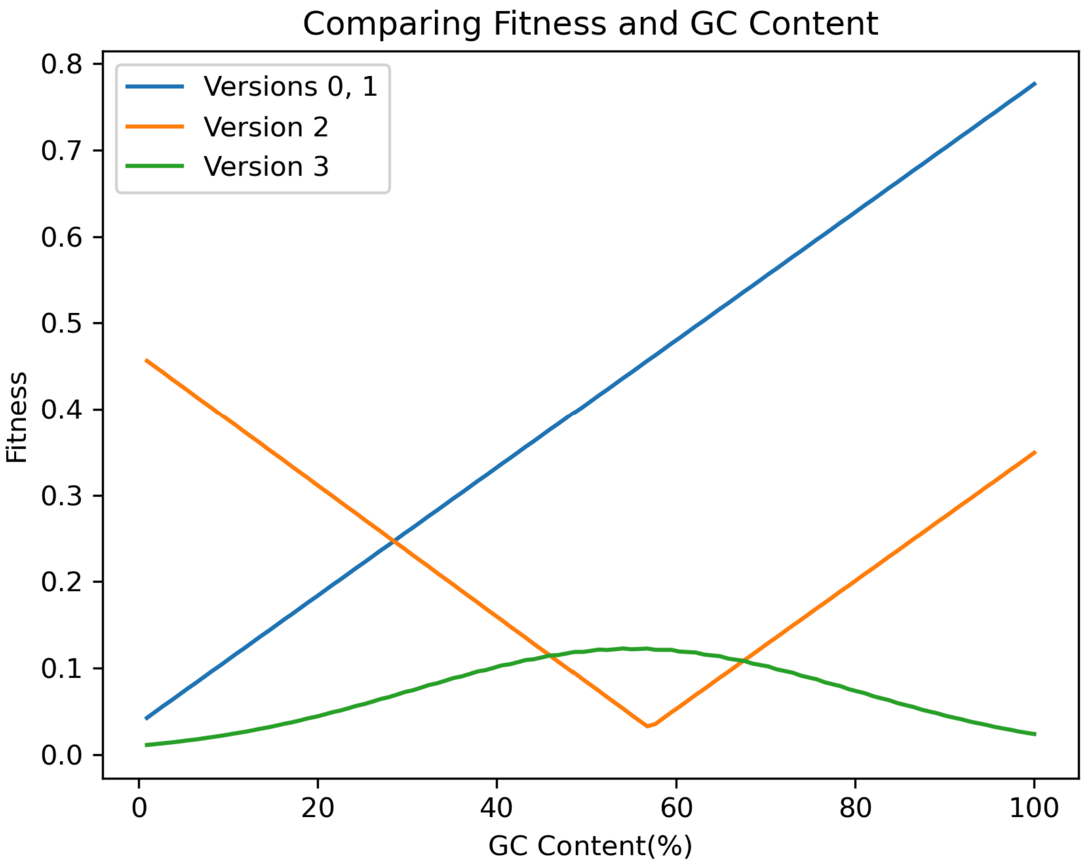
```
CTTCACCATCAGCGTGACCACAGAGATCCTGCCTGTGTCCATGACCAAGA

         !       !

CTTCACCATCAGCGTGACCACCGAGATCCTGCCTGTGTCTATGACCAAGA

 F T I S V T T E I L P V S M T K


T S V D C T M Y I C G D S T E C S

CCAGCGTGGACTGCACCATGTACATCTGCGGCGATTCCACCGAGTGCTCC

              !!!    !!

CCAGCGTGGACTGCACCATGTACATCTGCGGCGACAGCACCGAGTGCAGC

T S V D C T M Y I C G D S T E C S


 N L L L Q Y G S F C T Q L N R A L

AACCTGCTGCTGCAGTACGGCAGCTTCTGCACCCAGCTGAATAGAGCCCT

                 !

AACCTGCTGCTGCAGTACGGCAGCTTCTGCACCCAGCTGAACAGAGCCCT

 N L L L Q Y G S F C T Q L N R A L


 T G I A V E Q D K N T Q E V F A

GACAGGGATCGCCGTGGAACAGGACAAGAACACCCAAGAGGTGTTCGCCC

  ! !    !       !

GACCGGCATCGCCGTGGAGCAGGACAAGAACACCCAGGAGGTGTTCGCCC

 T G I A V E Q D K N T Q E V F A
```

Q V K Q I Y K T P P I K D F G G F

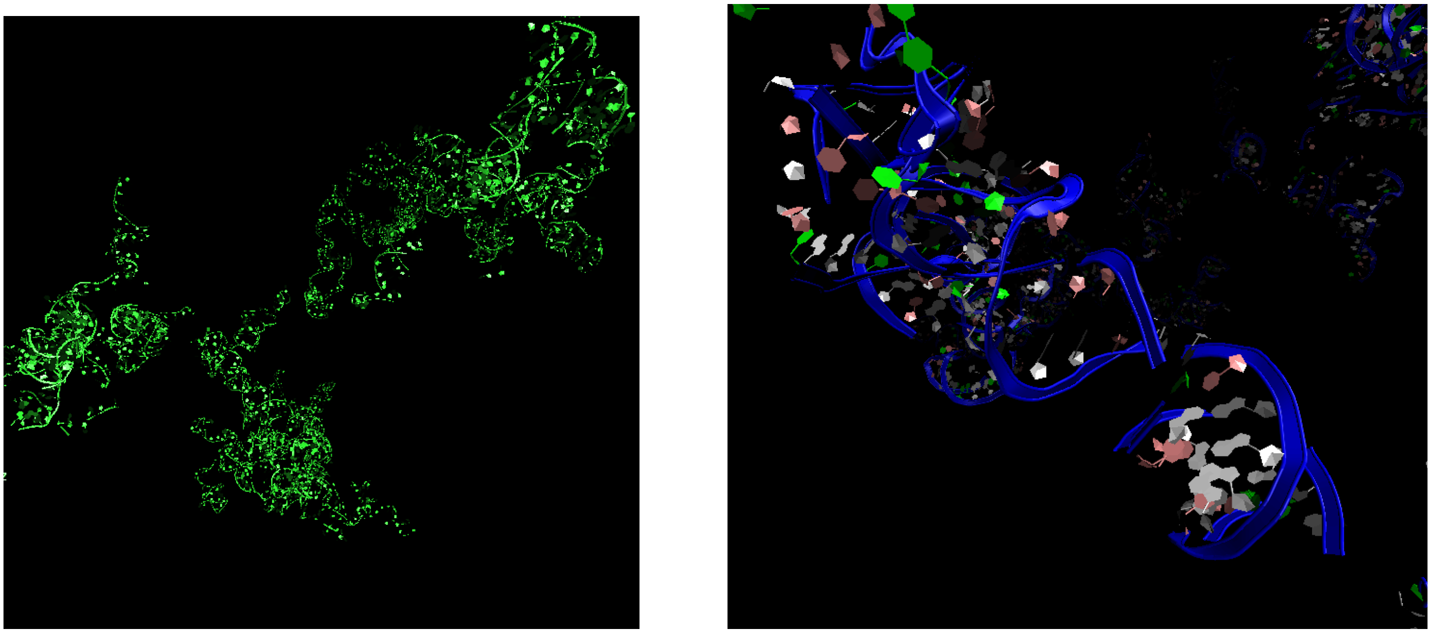
AAGTGAAGCAGATCTACAAGACCCCTCCTATCAAGGACTTCGGCGGCTTC

!

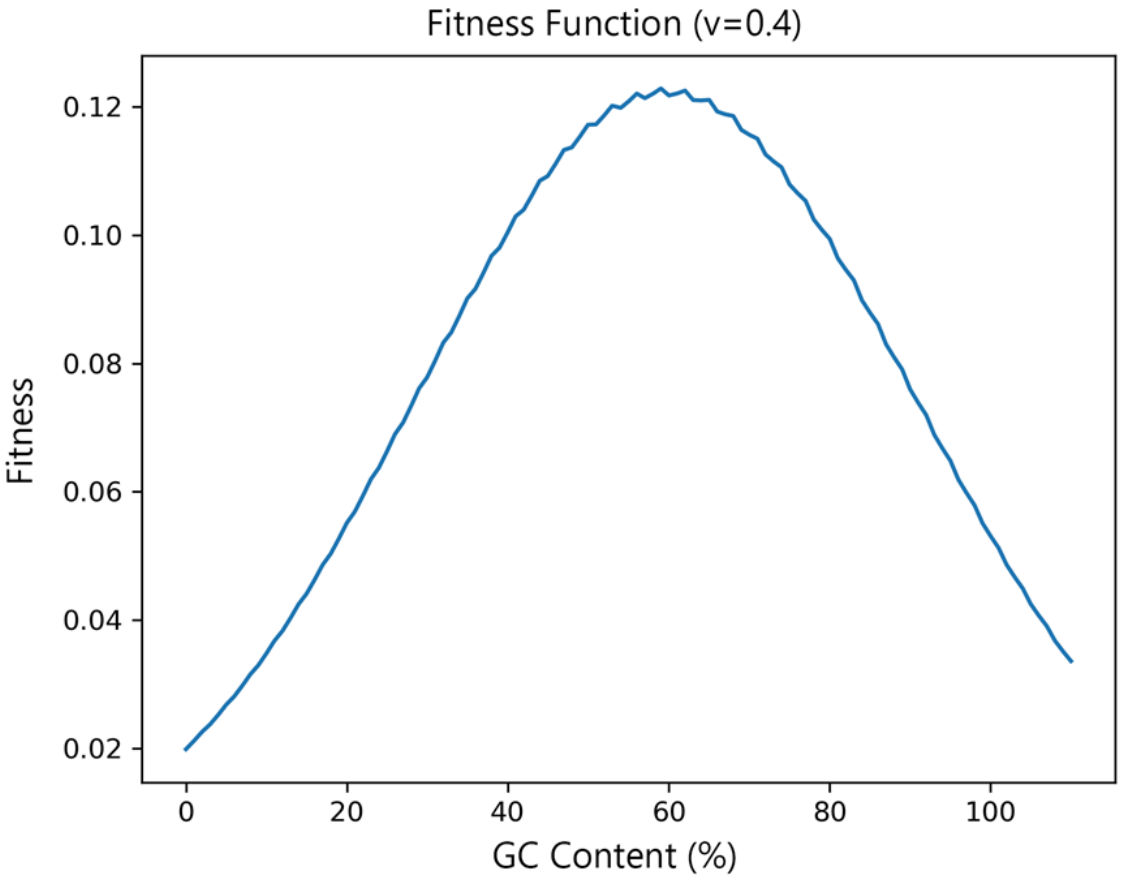
AGGTGAAGCAGATCTACAAGACCCCTCCTATCAAGGACTTCGGCGGCTTC


Comparing Fitness and GC Content
Versions 0, 1
Version 2
Version 3
Fitness
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80
100
GC Content(%)

Visualization of the True Vaccine(using 3dRNA Web Server and VMD. Left shows overall shape; right shows a zoomed in portion of the graph with colored nucleotides)



$$\varphi(C) = \sum_{1}^{N} P(C_n) \sum_{1}^{N} e^{-\frac{(GC_n - T)^2}{\sigma^2}}$$



Code

The code is available at https://github.com/html1101/Science-Fair-2020-2021/

Implementation

This project used a discrete optimization algorithm in order to encode the spike structural protein, with a focus on balancing GC content(the percent of guanine-cytosine base pairs within the particular sequence) as well as codon frequency(the average frequency of the codons within the sequence occurring within the human body; a higher level of codon frequency allows for faster protein folding). Data was taken from the complete genome of the SARS-CoV genome, available at the GenBank accession number MN908947, and the S spike structural protein was extracted using the locations within the CDS(Coding Sequence) section within the GenBank file using SeqIO(a package within the library Biopython). The discrete optimization algorithm was implemented using Python, as it is an object oriented language, allowing for the manipulation of the data with sufficient libraries available to manage biological files(although R has a larger selection of biological libraries, it is a procedural language, meaning that it is designed to carry out a series of computational steps rather than design objects to manage the manipulation of data). Several different versions of potential configurations of the discrete optimization algorithm fitness function and selection techniques were implemented and compared(see the second page for the details on the different versions).

Discussion of Results

Version 0 of the discrete optimization algorithm, in which a single codon was replaced with the best codon for that particular spot that would code for the same amino acid, while encoding for a sequence with an almost exact GC content of the true vaccine, overcompensates in the beginning of the sequence, replacing codons with the codon that would produce the highest possible GC content because the fitness function analyzes the entire sequence; as a result, the initial spike protein, with a naturally low GC content, would do very badly in the fitness function overall, and the optimization algorithm, reading the sequence from the beginning to the end, would overcompensate, replacing codons with high GC content values until at a certain point in the sequence the GC content would drastically dip as the GC content had reached the expected value and replace codons at the end of the sequence with low GC content codons. As a result, the nucleotide and codon similarities between the true vaccine and version 0 were very low, although the GC content was within 4% of the actual GC content.

Version 1 of the discrete optimization algorithm, in which a single codon was replaced with the best codon within a specified window, aids in fixing the problems with version 0 by looking at the GC content and codon frequency within a specific region as opposed to the full sequence. Different window sizes were tested, and ultimately the most accurate window is of size 12; however, Version 1 can come much closer to the true GC content of the vaccine, at the cost of major nucleotide and codon differences(a window of size 6 could produce a GC content within 1% of the actual vaccine). Ultimately, version 1 performs markedly better than version 0, although it does not compensate for natural fluctuations in GC content within certain regions, holding it to the fixed GC content value of the true vaccine throughout each window of size 12.

Version 2 of the discrete optimization algorithm, in which the best codon to replace within an entire sequence is analyzed and changed, produces very good results, with high nucleotide and codon similarities as well as high GC content and codon frequency levels, at the cost of speed. This type of function also allows for the possibility of a non-convex loss function(preventing from being trapped in local minimas of the fitness function).

Version 3 of the discrete optimization algorithm is almost identical to Version 2; however, it normalizes(places within the range 0-1) and optimizes the loss function, improving nucleotide similarity by 2.5% and a 6.82% improvement in codon similarity.

Conclusion

While mRNA sequence design is a complex process, within the context of this problem the researcher has proven through experimentation that by utilizing a form of discrete gradient descent, a genomic sequence can be formed with a 90.7% similarity to the vaccine between nucleotides and a 78.1% similarity across codons to the BNT162b1 vaccine. The final optimization algorithm selected utilized the following fitness function:

$$f(x) = e^{(} - GC(x)^2/v^2) * freq(x)$$

For which:

$v =$*variance; the final product found 0.4 resulted in the most similar sequence*

$GC(x) =$*GC content of the string(as a decimal)*

$freq(x) =$*Average frequency of codons occurring within the human body*

The final solution took 5527 seconds to converge on a laptop utilizing an Intel Core i7-8565U CPU, or almost 2 hours. Although its results were relatively similar to that of the codon map, the discrete optimization method holds promise in encoding non-convex sequences, and its fitness function allowed for greater control over codon optimization levels and GC content, resulting in a GC content with 2.63% difference from the true vaccine compared to the codon map’s GC content with 4.2% difference from the true vaccine, and codon optimization levels within 0.56% of the true vaccine(although it is noteworthy that there is potential to optimize codons further than in the true vaccine, suggesting that there is either a limit to the optimization of codons or sections which cannot be optimized further so as to maintain the same hairpin structures as in the true vaccine).

Bibliography

“AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology.” Deepmind, deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

Anthony Komaroff, MD. “Why Are MRNA Vaccines so Exciting?” Harvard Health Blog, 19 Dec. 2020, www.health.harvard.edu/blog/why-are-mrna-vaccines-so-exciting-2020121021599.

“CASP RR Format.” Prediction Center, www.predictioncenter.org/casp13/index.cgi?page=format#RR.

“Exploring the Supply Chain of the Pfizer/BioNTech and Moderna COVID-19 Vaccines.” Exploring the Supply Chain of the Pfizer/BioNTech and Moderna COVID-19 Vaccines · Neubertify, blog.jonasneubert.com/2021/01/10/exploring-the-supply-chain-of-the-pfizer-biontech-and-moderna-covid-19-vaccines/.

Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics'' J. Molec. Graphics 1996, 14.1, 33-38.

Kudla, Grzegorz, et al. “High Guanine and Cytosine Content Increases MRNA Levels in Mammalian Cells.” PLoS Biology, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/pmc/articles/PMC1463026/.

“NCBI Taxonomy Homepage: SG1.” National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG1.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, Michiel J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, Volume 25, Issue 11, 1 June 2009, Pages 1422–1423, https://doi.org/10.1093/bioinformatics/btp163.

“Reverse Engineering Source Code of the Biontech Pfizer Vaccine: Part 2.” Articles, 31 Dec. 2020, berthub.eu/articles/posts/part-2-reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/.

“Reverse Engineering the Source Code of the BioNTech/Pfizer SARS-CoV-2 Vaccine.” Articles, 25 Dec. 2020, berthub.eu/articles/posts/reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/.

“Severe Acute Respiratory Syndrome Coronavirus 2 Isolate SARS-CoV-2/Hum - Nucleotide - NCBI.” National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/nuccore/MT072688.

“Safety and Efficacy of the BNT162b2 MRNA Covid-19 Vaccine.” McMaster Optimal Aging Portal, www.mcmasteroptimalaging.org/full-article/plus/safety-efficacy-bnt162b2-mrna-covid-19-vaccine-96137.

Walsh, Edward E., et al. “Safety and Immunogenicity of Two RNA-Based Covid-19 Vaccine Candidates: NEJM.” New England Journal of Medicine, 31 Dec. 2020, www.nejm.org/doi/10.1056/NEJMoa2027906.

Xie, Xuping, et al. “Neutralization of N501Y Mutant SARS-CoV-2 by BNT162b2 Vaccine-Elicited Sera.” BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2021, www.biorxiv.org/content/10.1101/2021.01.07.425740v1.full.

Sanchez-Trincado, Jose L., et al. “Fundamentals and Methods for T- and B-Cell Epitope Prediction.” Journal of Immunology Research, Hindawi, 28 Dec. 2017, www.hindawi.com/journals/jir/2017/2680160/.

Fu, Hongguang, et al. “Codon Optimization with Deep Learning to Enhance Protein Expression.” Nature News, Nature Publishing Group, 19 Oct. 2020, www.nature.com/articles/s41598-020-74091-z.

“Long-Term Storage of DNA-Free RNA For Use in Vaccine Studies.” Digital Object Identifier System, doi.org/10.2144/000112593.

Wang, J., et al., Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. Nucleic Acids Res. 2017. doi:10.1093/nar/gkx386.