Prediction of Virus Mutant Based on Biological, Mathematical Analysis

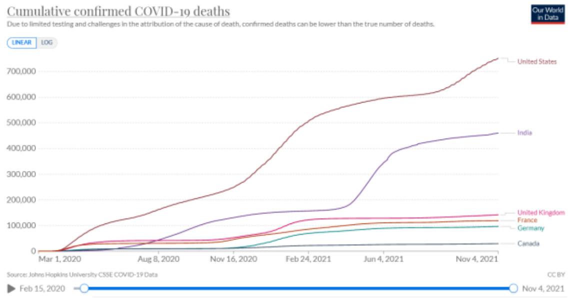
Oh Yugyung
Korean Minjok Leadership Academy

Abstract

As the physical distance between viral hosts shrunk, the frequency of pandemics caused by novel virus mutations increased. In the case of the influenza virus, vaccinations are manufactured and distributed prior to the appearance of the mutant. Firstly, from a biological perspective, the most effective strategy to target a virus would be to target spike proteins, which function as the virus's common weakness. Multiple mutations in spike proteins are typically responsible for the formation of new mutant viruses. A significant alteration in the form of the spike protein renders the viral cell invasion process itself impossible. This is because the structure of spike proteins is one of the most crucial factors in the process. Therefore, the spike protein of the virus varies only subtly, and by comparing the form of the spike proteins and cell receptors, we may predict the shape of the next mutant. Consequently, unlike other qualities that necessitate individual analysis, targeting spike proteins is remarkably efficient. Second, from a mathematical perspective, the capsid of the virus will have the structure of a truncated polyhedron with T=7. As viruses construct their capsids with a restricted number of genes, they use repeated units to form a regular polyhedron as their basic structure. In addition, since the solid figure that occupies the greatest volume with the smallest surface area is a sphere, the viral capsid will have a shape that is most similar to the sphere. To get closer to the sphere, the figure is truncated, and the triangulation number determines this fine structure. When T=7, evolutionarily speaking, the structure closest to the sphere in terms of stability and efficiency would be favored. The amino acid sequences of spike proteins and virus capsids dictate their respective shapes. ProteinBERT, a deep learning language specific to proteins, demonstrated the modeling of virus protein mutation in the spike proteins and virus capsids, as anticipated by the two perspectives on the virus prediction. However, there were also other variances that were not the result of natural mutations. Hence, a viral prediction would be far more accurate if these additional factors were included in the research. However, it is essential to predict the shape of the virus surface protein, which is the most crucial element in the virus-cell invasion. If it is possible to determine the genetic change of the spike protein by considering the shape of the cell receptor and the structure of the virus's capsid as it reaches a tetrahedron, an appropriate vaccine could be developed.

1. Introduction

Humans occupy the highest position in the ecosystem, but scientists assert that even they could become extinct, with infectious diseases triggered by viruses being one of the primary factors that could lead to their demise. In addition, there are numerous movies about the end of humanity due to viruses. However, people have already felt endangered by the global pandemic since the onset of Covid-19. According to [Figure 1], there have been 249,395,694 Covid-19 cases and 5,046,071 deaths around the planet. [1] Nevertheless, scientists expect that more dangerous and vital viruses will emerge and that the frequency of the disease will increase. This idea is not wrong, as the distance between humans and livestock is lessening and viruses have greater opportunities to transmit diseases from animals to humans as well as between humans. This phenomenon is known as a zoonotic spillover, and it can come in several ways, as indicated in [Figure 2]: directly, through a vertebrate or invertebrate intermediary host, or through the environment. [2]



Cumulative confirmed COVID-19 deaths

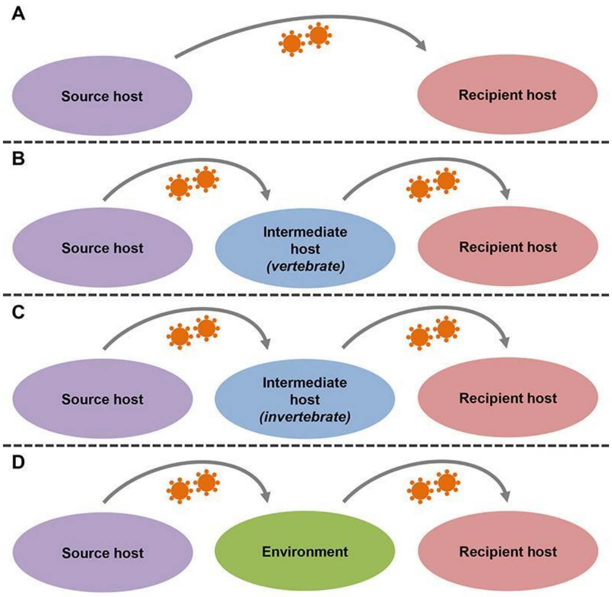
Due to limited testing and challenges in the attribution of the cause of death, confirmed deaths can be lower than the true number of deaths.

Source: Johns Hopkins University CSSE COVID-19 Data

CC BY

[Fig. 1] Cumulative confirmed deaths from COVID-19

It displays the total number of deaths caused by Covid-19 from February 15, 2020 to November 4, 2021.



[Fig. 2] Cases of zoonotic spillover

Pathogens overcome barriers between species in different manners. It can be spread directly or indirectly via vertebrate or invertebrate intermediary hosts and also the environment.

In addition, the virus's rapid evolution increases the frequency of outbreaks. Viruses constantly change and evolve into new types of viruses, especially RNA viruses. They grow more rapidly than DNA viruses because their genetic material is RNA, which is more unstable than DNA. And if this mutation increases the fitness of the virus, it will eventually result in the virus' evolution.

The Covid-19 virus, which is presently wreaking havoc, is also a coronavirus of the same type as the well-known SARS and MERS viruses. Then why was a prompt medical response to this novel coronavirus not possible? Although it has nearly the same structure as previous viruses, the viruses have mutated, demanding new research by scientists. The same holds true for the new influenza strain that emerges each year. There are numerous types of influenza viruses, but they may all be classified into four categories. This is also caused by flu virus mutations. [3]

Accordingly, the WHO estimates the level of mutation in the virus, predicts the influenza virus which will be prevalent that year, and uses this information to make vaccinations. In order to forecast the next phase in the development of the influenza virus, the WHO maps the evolution as phylogenetic trees and antigenic cartography. This information is then combined using statistical learning models. [4]

However, apart from evolutionary connections, what other factors could influence the evolution of viruses? The following passages discuss the directions of viral evolution based on biological and mathematical evidence. Mathematically, viruses are predicted to evolve toward a structure resembling an icosahedron. And biologically, the common tactic of viruses was found in their spike proteins. Thus, studying the structure of spike proteins in conjunction with their corresponding human cell receptor might allow us to forecast the next virus resulting from gene mutations.

2. Body

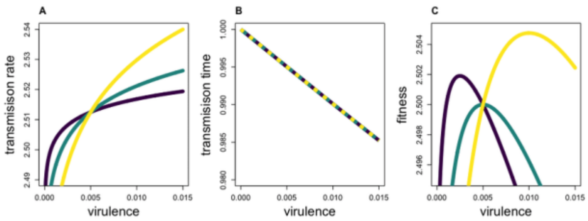
2.1 Biological Analysis

2.1.1 Characteristics requiring case-by-case analysis - Virulence

As the characteristics of viruses vary so widely, a case-by-case analysis is typically used to anticipate the next likely virus. Usually, these case-by-case studies are based on phylogenetic trees and "evolutionary trade-off," one of evolution's key concepts. The term "evolutionary trade-off" refers to the fact that organisms cannot benefit in all directions from evolution and must suffer losses in some areas. For instance, in the case of virulence, as depicted in [Fig. 3], the evolutionary trade-off between virulence and transmissibility, high virulence increases the transmission rate and duration of infection, but it is ineffective for the transmission cycle in hosts because high virulence can cause the host to die before the virus has sufficiently spread. [5]

[Fig. 3] Virulence-transmission trade-off

(a) Virulence-Transmission rate graph



As virulence increases, so does the transmission rate, however, the pace of growth in transmission rate reduces gradually.

(b) Virulence-Transmission time graph

As virulence increases, transmission time decreases, because hosts with severe virulence have a limited amount of time to transmit the virus before they die.

(c) Virulence-Fitness graph

Each virus has an optimal point of virulence for maximizing its own fitness. Too much or too little virulence could be detrimental to their survival or reproduction.

There are three possible scenarios for virulence evolution. The virulence could rise, decrease, or stay the same. Average virulence is likely to decrease for efficient transmission across hosts because 'reproduction' is the most critical factor and it is fatal if the host dies prematurely.

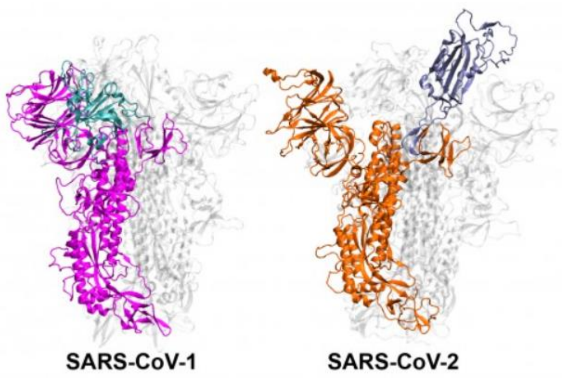
Yet, there are cases in which virulence increases. HIV, which nearly invariably results in AIDS, myxoma virus (MYXV), which killed 99 percent of infected rabbits, malaria, and rabbit hemorrhagic disease virus (RDHV) are examples of viruses with heightened pathogenicity. These viruses share the ability to overcome the disadvantages of their high virulence. HIV has a prolonged incubation period, allowing hosts to spread the disease before becoming aware of it. Similarly, MYXV, malaria, and RDHV are transmitted by mosquitoes, fleas, and blowflies. These vector-mediated infectious diseases could be transmitted even after the host has died because the host is not necessary for a subsequent infection; instead, the 'vector' is essential. Particularly for RDHV, the vector blow flies feed on animal carcasses, therefore the host is not required to be alive. [6]

2.1.2 Targeting the spike protein receptor

It will be far more efficient to target the common weaknesses of viruses, notably spike proteins, rather than conducting case-by-case analyses for every type of bacteria.

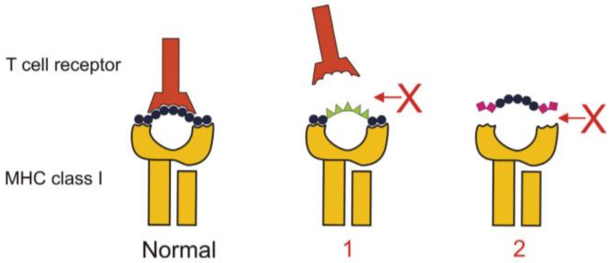
Viruses mutate faster than host genomes, with RNA viruses mutating quicker than DNA viruses because RNA viruses lack the ability to self-correct errors. Influenza viruses mutate very quickly, with an error rate of 0.5 nucleotide positions per genome per cell infection. Conversely, coronaviruses mutate at least four times slower than the influenza virus, which is still much quicker than the human mutation rate. [7]

The SARS-CoV-1 virus has undergone only 4 to 10 mutations to become the SARS-CoV-2 (Covid19) virus, but this small shift has already caused significant damage worldwide. This is because these tiny alterations usually happen in the sites where host cells and viruses combine, which are crucial to the life cycle of viruses. Prior to penetrating host cells, viruses must interact with the receptors present on the surface of host cells. Spike proteins play a prime role in this process. Spike(S) proteins are protrusions of the viral capsid that resemble knobs. Two domains, S1 and S2, are seen in spike proteins. S1 domain is responsible for binding to the host cell receptor, while S2 is responsible for membrane fusion. This spike protein is the primary difference between SARS-Cov-1 and SARS-Cov-2 viruses. Both spike proteins attach to ACE2 receptors, which are predominantly found on human nasal and pharyngeal cells. By mutating two amino acids in the spike protein, however, the SARS-Cov-2 viruses are now able to bind to human ACE2 receptors more effectively. As a result, SARS-Cov-2 viruses can spread more swiftly and easily among humans. [7],[8]



[Fig.. 4] SARS-CoV spike protein structure The overall structures of SARS-CoV-1 and SARS-CoV-2 spike proteins are similar, and the only difference is the colored part in the above figure. [9]

By interfering with molecular recognition or the phagocytosis process, viruses may also adapt to evade the human immune system. Influenza viruses are common examples. They are able to dodge the human immune system by mutating annually. There are three influenza strains: A, B, and C. Influenza A has 18 different haemagglutinin (HA) and 11 different neuraminidase (NA) subtypes (neuraminidase). HA is necessary for cell attachment or membrane fusion with host cells. But, when mutations occur in this region, it is crucial to make gradual changes so as not to lose the receptor-binding capability of the virus while avoiding antibody attachments.



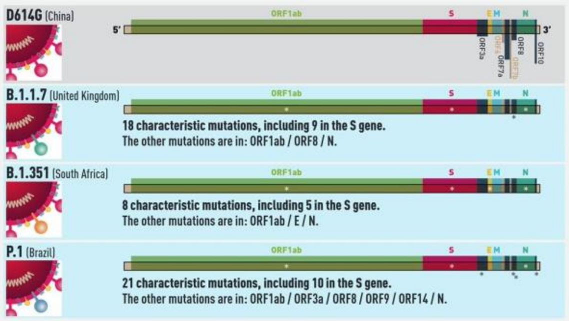
[Fig. 5] Influenza virus bypassing the immune system
T cell receptors can connect to altered influenza virus epitopes, thereby evading the host's immune system.

The changed conformation of epitopes prevents T cell receptors from recognizing the MHC class I protein of the virus.

(2) T cell receptors cannot recognize the MHC class I protein of the virus because mutations in the binding sites of epitopes and MHC class I render the epitopes incompatible with the virus. [10]

Therefore, the study of human cell surface receptors will be valuable for predicting the virus's mutation (evolution process in short term). The spike protein structure of the virus must not deviate greatly from the form of the human receptor for the virus to enter, allowing people to predict the next spike proteins that viruses will contain.

Four groups of SARS-CoV-2 variants have previously been identified: alpha variant (B.1.1.7), beta variant (B.1.351), gamma variant (P.1), and delta variant (B.1.1.7) (B.1.617.2). When comparing the DNA sequences of each alpha, beta, and gamma variation, there were 18, 8, and 21 characteristic mutations, with 9, 5, and 10 mutations occurring in the spike protein gene, respectively. And as a result of these modifications, the original SARS-CoV-2 virus could undergo functional alterations, such as increased transmission efficiency, reduced antibody binding and immune production, and diminished vaccine efficacy.


D614G (China)
ORF1ab
S
E M
N
5'
3'
ORF3a
ORF6
ORF7a
ORF7b
ORF8
ORF10
B.1.1.7 (United Kingdom)
18 characteristic mutations, including 9 in the S gene.
The other mutations are in: ORF1ab / ORF8 / N.
B.1.351 (South Africa)
8 characteristic mutations, including 5 in the S gene.
The other mutations are in: ORF1ab / E / N.
P.1 (Brazil)
21 characteristic mutations, including 10 in the S gene.
The other mutations are in: ORF1ab / ORF3a / ORF8 / ORF9 / ORF14 / N.

[Fig.6] DNA comparison of SARS-CoV-2 variants

Compared to the initial sequence, the DNA of the SARS-CoV-2 variants has only 18, 8, and 21 mutations. In addition, the majority of mutations have occurred in their spike proteins. This indicates that mutations in spike proteins are crucial for the emergence of a new viral mutant. [11]

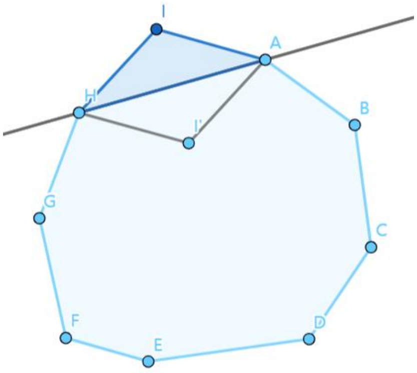
2.2 Mathematical Analysis

2.2.1 Evolution in Capsid Structure

A capsid is a protein shell that encases the viral genome. It safeguards the genome from a variety of chemicals, such as nucleic acid-degrading enzymes produced by the host cell or other substances. These capsids are comprised of microscopic particles, and diverse geometrical configurations are generated based on the arrangement of the particles; nonetheless, the majority of virus capsids are helical or icosahedral. [12], [13]

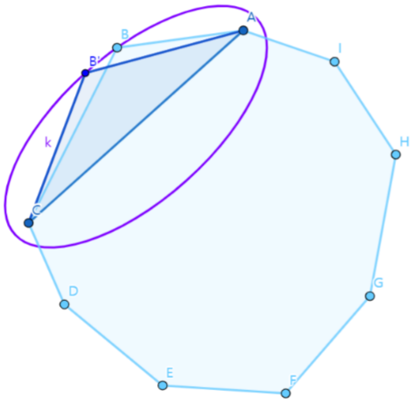
Under the same resource conditions, viruses with a nucleic acid size of 10kb will multiply ten times higher than viruses with a nucleic acid size of 100kb. Accordingly, viruses are likely to minimize their genomes and utilize their hosts' genomes to the greatest extent possible. As viruses cannot transport a large amount of nucleic acid, the amount of proteins they can generate is limited. Eventually, their capsids are composed of repetitive instances of one or more distinct types of proteins.

But the most essential objective is to maximize efficiency. In terms of physical stability and balance, it is preferable to have a symmetrical structure with relatively low free energy, which is usually observed in polyhedral formations. In accordance with the isoperimetric problem, the circle with a given circumference has the greatest area. In a concave polygon, reflecting the concave part outward would provide a polygon with the same circumference but a larger area, as depicted in Fig. 7. Therefore, concave polygons cannot have the maximum area.



[Fig.7] A concave polygon ABCDEFGHIJ and point K which point A is reflected. The area of convex polygon KBCDEFGHIJ is larger than the area of concave polygon ABCDEFGHIJ.

Draw an ellipse passing through point B with focal points A and C, as shown in Fig. 8, and it can be seen that the perimeter is constant regardless of how point B is positioned above the ellipse. However, the area is maximized when the lengths of sides AB and BC are equal, so polygons with all equal sides can have the largest area. Consequently, when any four points are picked, the four-point figure is at its largest when the four-point shape is a square that is inscribed in a circle, as illustrated in Fig. 9, based on Bretschneider's formula.



[Fig..8] A convex polygon ABCDEFGHIJ and point B' on the ellipse that passes through point B and takes points A and C as focal points The area of polygon AB'CDEFGHIJ is larger than the area of polygon ABCDEFGHIJ.

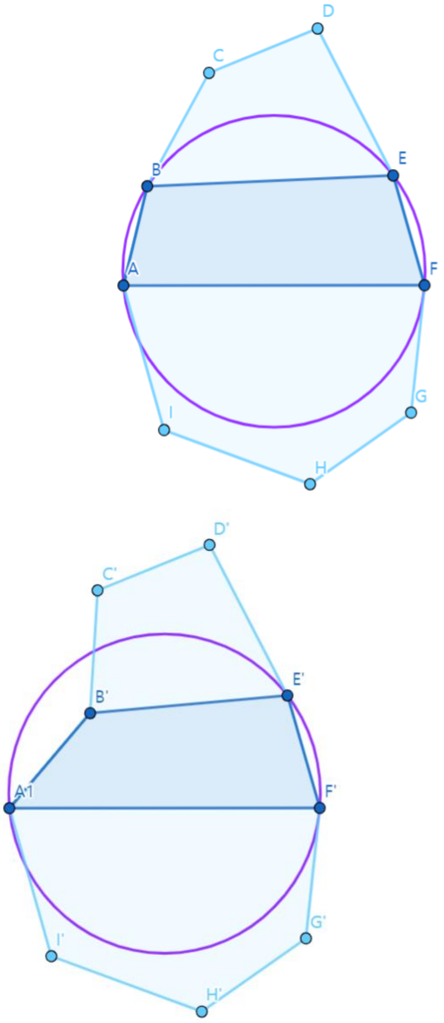
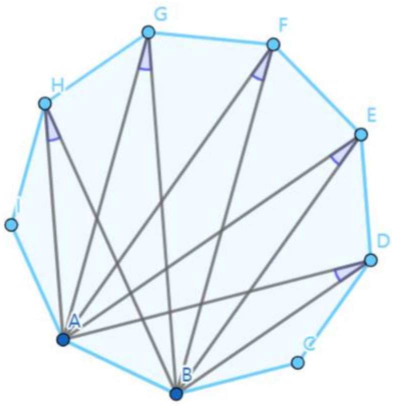


Fig. 9 Four points A, B, C, and D are not inscribed in a circle, and four points A', B', C', and D' are inscribed in a circle. The area of polygon ABCD is larger than the area of polygon A'B'C'D'.

Because squares comprised of any four points are embedded in circles, the circumference angles for each side will be identical. As all sides are of equal length, the circumference angles will have the same size, and all the internal angles of the figure will likewise be identical (Fig. 10). When a specific circumference is provided, the polygon will consequently have the greatest possible width. [14]



[Fig..10] In a polygon ABCDEFGHIJ, in which any four points are inscribed in a circle, the size of angles ACB, ADB, AEB, AFB, AGB, AHB, and AIB are all equal.

The following is a generalized formula for calculating the width of an equilateral polygon, which is associated with the center angle.

$s=(circumference\ of\ an\ equilateral\ polygon)$

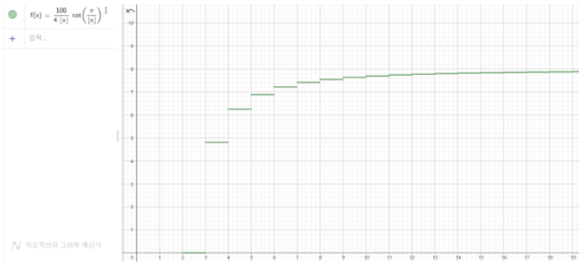
$n=(total\ number\ of\ sides)$

when the equilateral polygon is splited into triangles with each point at the center,

$h=(height\ of\ each\ triang\le s)$

$(Area\ of\ an\ equilateral\ polygon)=\frac{s}{2n}\times h\times n=\frac{s}{2n}\times\left(\frac{s}{2n}\times\cot\left(\frac{\pi}{n}\right)\right)\times n$

When a graph is drawn to compare a circle to a regular polygon by taking the variable n to its logical extreme in the above formula, it is clearly demonstrated that as n increases, so does the area (Fig. 5). In other words, a circle has the maximum width. When we stretch this isoperimetric problem into the third dimension, we find that the spherical solid figure has the greatest volume.



$f(x)=\frac{100}{4\,[x]}\cot\left(\frac{\pi}{[x]}\right)$

[Fig. 11] Side numbers and area of an equilateral polygon

The graph is generated by applying the formula presented above, converting n into [x], and then entering a random value of 10 into s. The number of sides, denoted by n, is plotted along the x-axis of the graph, while the area of an equilateral polygon is plotted along the y-axis. When n heads infinity, the area is at its maximum value.

When compared to its surface area, the icosahedron, which is the polyhedron that is closest to the shape of a sphere, has the largest volume. The icosahedron, which has the maximum volume and consists of 20 triangles, 12 vertices, and 30 edges, provides the virus with multiple advantages. Because viruses with icosahedron capsids would take up less volume and surface area than other solid figures when holding the same amount of genome, this makes them more efficient in their use of materials and able to easily withstand external shocks while simultaneously minimizing their external exposure.

Truncated icosahedrons, on the other hand, in which each vertex is cut off, have the potential to raise the number of acceptable genomes even further. The shapes of these truncated icosahedrons are determined by the number of pentamer and hexamer that may be counted by making use of the Triangulation number (T). As a result, in accordance with the natural selection theory, which states that evolution occurs in a direction that can optimize survival and reproduction, it is anticipated that the virus would evolve in the direction of having a structure that is similar to an icosahedron.

2.2.2 Triangulation Number and Virus Characteristics

As said before, the triangulation number determines the detailed shape of a truncated icosahedron. Triangulation number of a virus capsid hence controls the virus's size and stability. Lower triangulation numbers are advantageous to the virus for two reasons.

Triangulation number (T) is the number of structural units that make up one side of an icosahedron. When the number of identical units on each face rises, the icosahedron structure of the capsid expands without the requirement for a new protein type. Similarly, as T decreases, the virus drops. As stated previously, viruses tend to reduce the size of their genomes, therefore big capsids are unnecessary. Moreover, the reduced surface area of capsids with a lower triangulation number reduces their weak point. Thus, viruses would prefer a lower triangulation number, which leads to a smaller capsid.

Taking stability into account, T=7 capsids are preferred. Capsids consist of 60 protein subunits, which assemble into 12 pentons and a variable number of hexons. Here, the hexon structure of the capsid is the determining factor of stability. A capsid of T≤7 contains exactly one type of hexon (excluding T=1, which does not contain any), whereas T≥9 capsids contain multiple types of hexons. T=7 capsids, for instance, have a single-pucker hexon conformation, but capsids have both winged and flat hexons. The distinct hexon shape of T≥9 capsids necessitates the presence of decorating proteins such as MCP for the virus's stability, demanding another source of the material. [15] The capsid with the closest three-dimensional shape to a sphere, between 1≤T≤7 capsids, is T=7.

Consequently, in the perspective of triangulation number, viruses tend to have capsids with T=7.

3. Heuristic suggestion with proteinBERT

3.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a 'pre-training language model' as opposed to LSTM, CNN, and ensemble models that perform tasks such as object name recognition and text classification. If there is sufficient data, Embedded has a substantial impact on the model's ability to do a specific task. Embed words that effectively convey their meaning as vectors will naturally perform well during the training process. In this embedding procedure, BERT is utilized, and it can be stated that BERT is a language model that can enhance the performance of a specific task through pre-training embedding prior to executing it. BERT is composed of three distinct components: Input, Pre-training, and Transfer Learning. Token Embedding, Segment Embedding, and Position Embedding are necessary for the input part. In the Token Embedding process, each Char unit is embedded using the Word Piece embedding method, and the most often occurring sub-word of the longest length is converted into a single unit. Words that appear seldom are reformed into sub-words. This can help tackle the OOV problem, which worsened model performance by 'OOV' processing all words that had not previously appeared frequently. Segment Embedding is synonymous with Sentence Embedding. It involves recombining the tokenized words into a sentence. In BERT, two sentences are separated by a separator ([SEP]) and specified as a single segment. This segment is limited to 512 sub-words by BERT. The author of BERT previously released the Transformer model, which employs the Self-Attention model instead of CNN and RNN. BERT uses simply Transformer's encoder and decoder. To be more specific, Self-Attention takes into account the location information of the input token but does not consider the location of the input token itself. Therefore, Transformer models use Position Encoding using Sinusoid functions, and BERT follows suit.

As encoded data are ready, we are currently conducting pre-training. Previous methods often predict the following word by either learning sentences from left to right or by examining the left and right context of the word to be predicted. Masked Language Model(MLM) and Next Sentence Prediction(NSP) are utilized by BERT for superior training, though. MLM discards the token at random from the input sentence (Mask) and then matches the token to proceed with the learning. NSP predicts the sequence when provided with two sentences. To fine-tune NLI and QA, where the relationship between the two sentences should be considered, we proceed with learning the relationship between the two sentences.

Last but not least, Transfer Learning is the process of transferring the learned language model and carrying out the actual NLP Task. This is where performance is observed. Prior to the introduction of BERT, if I wanted to solve the NER problem, I had to devise an algorithm or language model, and if I wanted to address the QA problem, I had to create an algorithm or language model separately. Using the BERT language model, it has been demonstrated, however, that performance is enhanced by learning the transfer and performing the desired task. In many aspects, BERT appears to be an excellent model. Previously, the part that makes a language model was semi-supervised learning that labels itself; however, transfer learning is supervised learning in which labels are given. Transfer Learning is the process of constructing a second model for NLP Tasks within the BERT language model.

3.2 proteinBERT

BERT has been designed for NLP, as described earlier. Protein sequences can be represented as an array of amino-acid characters. However, there are two key distinctions between human language and protein sequence. In the first place, unlike natural language, the protein sequence cannot be divided into words and sentences. Second, the length of the protein sequence is less predictable. In 2021, Nadav Brandes et al. developed a revolutionary deep-learning model for protein sequences, dubbed 'proteinBERT', that resolved the aforementioned difficulties. [16]

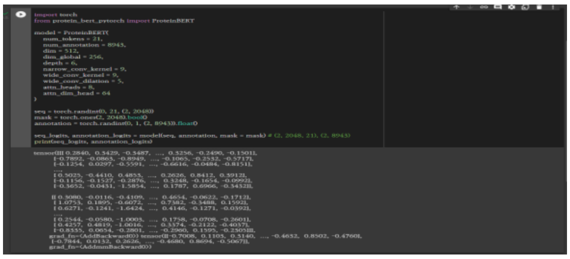
In the process of proteinBERT, they pre-trained on the entire tree of life proteins derived from the UniProtKB/UniRef90. The protein sequence was encoded as a sequence of integer tokens. They employed 26 unique tokens to represent 20 standard amino acids, selenocysteine (U), undefined amino acids (X), additional amino acids (OTHER), and three other tokens (START, END, PAD). To help the model interpret proteins longer than the selected sequence length, START and END tokens were added to each sequence before the first and after the last amino acids, respectively. Without a START or END token, the model can recognize that only a portion of the sequence has been received.

By periodically switching the encoding length of the protein sequence, they could avoid the risk of overfitting the model to a set length. After fine-tuning the data, the proteinBERT model emerged with six transformer blocks, each containing four global attention heads. The model is independent of the length of the sequence being processed and can be adapted to sequences of any length without changing the parameter length.

3.2 Heuristic method

The proteinBERT learns potential protein sequences in the same way that it learns to fill in the blanks through mask filling. Replacing and predicting amino acid residues in the spike protein sequence with mask-flagged tokens allows structurally realistic mutations in proteins.

[Fig. 12] is an illustration of the implementation of the proteinBERT. Users can select the number of tokens, etc. I assigned a random token array to the variable seq and the variable annotation. For masking, I generated a 1s array *(torch.ones())* and assigned it to the variable mask in order to retrieve all possible sequences.



```
import torch
from protein_bert_pytorch import ProteinBERT

model = ProteinBERT(
    num_tokens = 21,
    num_annotation = 8943,
    dim = 512,
    dim_global = 256,
    depth = 6,
    narrow_conv_kernel = 9,
    wide_conv_kernel = 9,
    wide_conv_dilation = 5,
    attn_heads = 8,
    attn_dim_head = 64
)

seq = torch.randint(0, 21, (2, 2048))
mask = torch.ones(2, 2048).bool()
annotation = torch.randint(0, 1, (2, 8943)).float()

seq_logits, annotation_logits = model(seq, annotation, mask = mask) # (2, 2048, 21), (2, 8943)
print(seq_logits, annotation_logits)
```

```
tensor([[[ 0.2840,  0.3429, -0.3487,  ...,  0.3256, -0.2490, -0.1501],
         [-0.7892, -0.0863, -0.8949,  ..., -0.1065, -0.2532, -0.5717],
         [-0.1254,  0.0297, -0.5591,  ..., -0.6616, -0.0484, -0.8151],
         ...,
         [ 0.5025, -0.4410,  0.4853,  ...,  0.2626,  0.8412,  0.3912],
         [-0.1156, -0.1527, -0.2876,  ...,  0.3248, -0.1654, -0.0992],
         [-0.3652, -0.0431, -1.5854,  ...,  0.1787,  0.6966, -0.3432]],

        [[ 0.3080, -0.0116, -0.4109,  ...,  0.4654, -0.0622, -0.1712],
         [ 1.0753,  0.1895, -0.6072,  ...,  0.7382, -0.3488,  0.1592],
         [ 0.6271, -0.1241, -1.6424,  ...,  0.4146, -0.1271, -0.0392],
         ...,
         [ 0.2544, -0.0580, -1.0003,  ...,  0.1758, -0.0708, -0.2601],
         [ 0.4257,  0.4819, -1.0016,  ...,  0.3374, -0.2122, -0.4037],
         [-0.8335,  0.0654, -0.2801,  ..., -0.2960,  0.1595, -0.2305]]],
       grad_fn=<AddBackward0>) tensor([[-0.7008,  0.1103,  0.3140,  ..., -0.4632,  0.8502, -0.4760],
        [-0.7844,  0.0132,  0.2626,  ..., -0.4680,  0.8694, -0.5067]],
       grad_fn=<AddmmBackward0>)
```

[Fig.12] The example of Implementing the proteinBERT

Since the SARS-CoV-2 responsible for COVID-19 has been identified, it is possible to set its default sequence using correlated integer tokens. The mask will then be used for the variations. However, it is anticipated that the number of variants would be enormous. At this point, pruning techniques can be considered. In the previous sections of this research, I suggested the potential structure of the virus' spike protein. There should be some protein structures that can never meet the proposed optimal structure. Once these sequences are removed, the total amount of outcomes and the computational time complexity will lessen. Thus, the proteinBERT may return the proper number of mutations. In addition, the history of varying processes from SARS-CoV-1 to SARS-CoV-2 provides valuable information for filtering the outcome.

4. Conclusion and Future works

Viruses frequently mutate, leading to fast mutation and evolution. This is a significant benefit for the viruses, but it poses a substantial danger to humans. To prepare for these "rapidly evolving" viruses, humans produce vaccinations based on their projected traits. These studies are often conducted by evaluating evolutionary relationships using phylogenetic trees and antigenic cartography. However, other biological and mathematical assessments suggested that the virus's capsid and spike proteins were responsible for its mutation. Factors such as virulence and incubation period could be the key to predicting the next virus.

As these characteristics require case-by-case analyses for each type of bacterium, it will be more efficient to target the common weaknesses of viruses, namely the spike protein. It is true that viruses undergo mutations fairly quickly, however, even the influenza viruses with the fastest mutation rate cannot reach 0.5 nucleotide positions per genome per cell infection. Furthermore, "shape" is the most significant factor in ligand-receptor interaction. Thus, the structure of spike proteins cannot be altered much. Subsequently, new spike proteins of viruses can be predicted by focusing on variants of the original structure while taking human cell receptors into account. This was confirmed through the examination of SARS-CoV-2 variants B.1.1.7, B.1.351, P.1, and B.1.617.2. Compared to the original DNA sequences of SARS-CoV-2, the majority of mutations occurred in the spike protein. And most of these mutations have resulted in more efficient transmissions and reduced antibody, immune production, and vaccine efficacy.

Next, the geometrical structure of a capsid tends to be an icosahedron and a sphere, which may be demonstrated by the extension of the isoperimetric problem. Due to the fact that viruses can only contain a small amount of genome, their primary objective is to construct a capsid structure that is as efficient and stable as possible with their limited genetic capacity, and this can be portrayed by a three-dimensional shape that resembles a sphere.

The subtle geometry of a truncated icosahedron is controlled by triangulation number, which is the number of structural units that make up one side of an icosahedron, and T=7 is the optimal number for the virus in terms of its stability and efficiency. First, a lower triangulation number reduces the genome's size and exposed areas, which is a tendency that viruses prefer. In addition, triangulation numbers below 7 are favored due to their simple hexon configuration, which contributes to increased stability. Therefore, T=7 is the optimal structure that is both stable and spherical.

In conclusion, when developing a vaccine against a forecasted virus, it is vital to take into account the exquisite shape of the spike protein receptors and capsids. First, biologically, spike protein receptors tend to exhibit only minor variations; therefore, identifying the spike protein variants would serve as an attack against the viruses' common weakness. Second, mathematically, the virus would evolve to have a capsid with a T=7 sphere if its efficiency and stability were factored into the equation when keeping the definition of evolution in mind. Therefore, when predicting the next virus mutation, the shape of spike protein receptors and capsids should be prioritized.

Lastly, the virus's protein sequence mutation was simulated using proteinBERT. A 1s array was assigned to the variable mask, and potential mutations were anticipated by filling in the masks. Currently, the amino acid sequences of coronavirus have been analyzed, allowing us to forecast the next mutant virus that would be derived from coronavirus and produce vaccines in advance, which is our ultimate goal. However, there are so many potential candidates that multiple pruning systems are required. Using preprocessing and varying the parameters in proteinBERT, a novel method known as "viral proteinBERT" may be applied to actual research. There appears to be a significant need for additional study in this area involving computational biology.

Nevertheless, it is glaringly obvious that it will be too late to research the virus and create a vaccine only after a new virus begins to spread. Damage could only be mitigated if the virus is foreseen and prepared for beforehand. The findings above provide more information about future viruses, which will enable us to guard against future viruses.

Once again, humans are at the apex of the ecosystem, and as long as we are prepared, viruses will one day be utterly conquered.

4. References

[1] COVID-19 CORONAVIRUS PANDEMIC, https://www.worldometers.info/coronavirus/

[2] Ellwanger J. H. & Cies J. A. B. Zoonotic spillover: Understanding basic aspects for better prevention, Genet. Mol. Biol. 44, (2021) https://www.scielo.br/j/gmb/a/TTzyffs6tcX37QwCr7fQqQp/?lang=en#ModalFigf2

[3] Petrova, V. N., & Russell, C. A., The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 2018, https://www.nature.com/articles/nrmicro.2017.118

[4] Geoghegan, J.L., Holmes, E.C. The phylogenomics of evolving virus virulence. *Nature Reviews Genetics* *19*, 2018, https://doi.org/10.1038/s41576-018-0055-5

[5] Miller Ian F. & Metcalf Jessica E., No current evidence for risk of vaccine-driven virulence evolution in SARS-CoV-2. medRxiv, 2020, https://www.medrxiv.org/content/10.1101/2020.12.01.20241836v1.full

[6] Elsworth Peter, Cooke Brian D., Kovaliski John, Sinclair Ronald, Holmes Edward C., Strive Tanja, Increased virulence of Rabbit Haemorrhagic Disease Virus associated with genetic resistance in wild Australian rabbits (Oryctolagus cuniculus), NCBI, 2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4598644/

[7] Ed Rybicki, Anna-Lise Williamson Don Cowan Stephanie G Burton, How the coronavirus mutates and what this means for the future of COVID-19, The Conversation, 2021, https://theconversation.com/how-the-coronavirus-mutates-and-what-this-means-for-the-future-of-covid-19-154499

[8] Shang Jian, Wan Yushun, Luo Chuming, Ye Gang, Geng Qibin, Auerbach Ashley, Li Fang, Cell entry mechanisms of SARS-CoV-2, PNAS, 2020, https://www.pnas.org/content/117/21/11727

[9] Biophysical Society, Why some coronavirus strains are more infectious than others may be due to spike protein movements, News Wise, 2021, https://www.newswise.com/coronavirus/why-some-coronavirus-strains-are-more-infectious-than-others-may-be-due-to-spike-protein-movements

[10] Sandt Carolien E. van de, Kreijtz Joost

H.C.M., Rimmelzwaan Guus F., Evasion of Influenza A Viruses from Innate and Adaptive Immune Responses, Viruses, 2012, https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.mdpi.com%2F1999-4915%2F4%2F9%2F1438%2Fpdf&psig=AOvVaw2_3Rv8E1HesIs60fxvGEms&ust=1636110014697000&source=images&cd=vfe&ved=0CAwQjhxqFwoTCKDwj_3G_vMCFQAAAAAdAAAAABAD

[11] Wanner Mark, The emergence of SARS-CoV-2 variants sparks concern, The Jackson Laboratory, 2021, https://www.jax.org/news-and-insights/2021/february/new-coronavirus-variants-spark-concerns

[12] Louten Jennifer, Virus structure and classification, NCBI, 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7150055/

[13] Dory Gascueña, The Secret Mathematics of Viruses, Openmind BBVA, 2020, https://www.bbvaopenmind.com/en/science/research/the-secret-mathematics-of-viruses/

[14] Gluck Herman, The isoperimetric problem, Math501, 2012, https://www2.math.upenn.edu/~shiydong/Math501X-5-Isoperimetric.pdf

[15] Parvez Mohammad Khalid, Geometric architecture of viruses, Baishideng Publishing Group, 2020, https://www.wjgnet.com/2220-3249/full/v9/i2/5.htm

[16] Brandes, Nadav, et al. "ProteinBERT: A universal deep-learning model of protein sequence and function." bioRxiv, 2021.