Analysis of SARS-CoV-2 Mutational Networks

Avatar Jaykrushna
ajaykrus@asu.edu
Arizona State University
Tempe, Arizona

Ankush Tale
avtale@asu.edu
Arizona State University
Tempe, Arizona

ABSTRACT

Humanity has been through eight major pandemics since the early 1700s and is valiantly fighting the COVID-19 at this moment. While the scientific fraternity is focused on finding a cure, it is also important to study the evolution of the SARS-CoV-2 virus which would give a critical insight in not only it's origin but it's future possible mutations. In this paper, we will explore the evolution aspect of SARS-CoV-2 (Wuhan-Hu-1: NC045512[1]) and impact of different mutation rates on the phenotypic and genotypic changes in the virus. We theoretically examine the probability of evolution in the current strain of SARS-CoV-2 which matches with observed data. We also estimate the timeframe for evolution of the current strain to a previously known coronaviruses with higher fatality rate.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; **Computational genomics**.

KEYWORDS

Mutation Networks, Coronaviruses, Genomics

ACM Reference Format:
Avatar Jaykrushna and Ankush Tale. 2020. Analysis of SARS-CoV-2 Mutational Networks. In *Proceedings of CSE 598 (Bio-inspired Computing)*. Arizona State University, Tempe, AZ, USA, 5 pages. https://doi.org/10.1145/nnnnnnn

1 INTRODUCTION

The virus SARS-CoV-2 has undeniably shook all of mankind and disrupted lives of all around the globe. As of April 30, 2020, there are 3.3 million reported cases of COVID-19 disease with roughly one-thirds in North America. As per report from Center for Infectious Disease Research and Policy[5], the disease is expected to last for 18 to 24 months as herd immunity gradually develops in the human population. While the pandemic can be controlled over time with mitigation efforts, the socio-politico-economic ramifications would be around for far more. Currently, the scientific community is actively studying the origin, transmissibility and treatment of the virus SARS-CoV-2. What makes SARS-CoV-2 unique is the presence of 6 amino acids in receptor-binding domain (RBD) from spike protein which has found to have high affinity of binding with ACE2 receptors in humans[3]. ACE2 is an enzyme that attaches to the outer surface of cells in lungs, arteries, heart, kidney and intestines [2]. The high affinity of ACE2 to the RBDs in spike protein of SARS-CoV-2 gives it it's zoonotic nature. Out of the 6 amino acids, 5 differ between SARS-CoV-2 and SARS-CoV from 2002, a more deadlier form of coronaviruses.

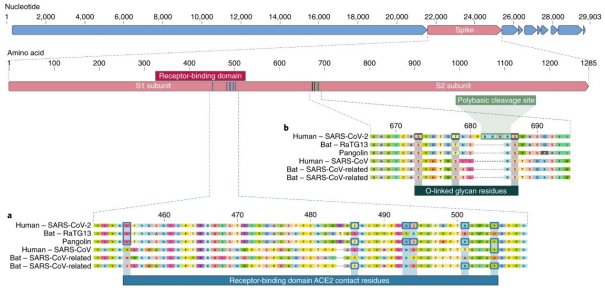

Figure 1: Features of spike protein in SARS-CoV-2 and comparision of its nucleotides with other known coronaviruses [3]

The Wuhan-Hu-1 genome sequence collected from an early patient in Wuhan on Dec 19 is considered a baseline of SARS-CoV-2. With numerous transmissions, the genome sequence occasionally undergoes mutations (single letter typos) and thus forms a new variant of the virus. This mutation cycle continues throughout the pandemic transmission thus creating a network of mutated genome sequences originating from a single base sequence. In Figure 1, you can see the similarity between Bat-SARS virus and SARS-Cov-2 virus leading to the popular belief that the virus transmitted from Bats. In this paper we analyze this mutation pattern.

2 MUTATIONAL NETWORKS

Coronaviruses are RNA genome based and as all RNAs are made up of 4 types of nucleotides a, c, u and g. A sequence of 3 nucleotides forms a codon. Thus, a sequence of 3 nucleotides each with 4 options result in 64 possible combinations of codons. These 64 codons help corresponding to 20 amino acid and the codon wheel[6] can be referred for the mapping. For the sake of simplification in this study, we disregard the importance of Open reading Frames(ORF) which consider the start and stop codon. ORFs are used to mark areas of RNA used for protein synthesis. In this paper we present our theoretical analysis addressing the following questions.

2.1 Possible RNA strands one mutation away

The SARS-CoV-2 virus is extraordinary long (29872) RNA genome. For the sake of this paper, we assume its length to be 30000 nucleotide long. To start with, let's find out the total number of possible

Bio-inspired Computing, Spring 2020,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn

RNA strands of the same length that are 1 missense mutation away from the given sequence.

2.2 Neutral mutations

Since not all 64 combinations of 3 nucleotides (codons) do not directly map one-on-one with amino-acids, there is a lot of redundancy. For any mutation in a codon, it is probable that the resultant codon encodes to the same amino acid. For example, if the third nucleotide mutates from TCT to TCG, the codon still encodes from S. This is termed as Neutral mutation. In continuation to the previous scenario in 2.1, let's find out the fraction of neutral mutated RNAs from the total mutated RNA strands. This helps in determining the fraction of mutations that probably would cause phenotypic change.

2.3 Spike Protein mutations

Hereon we only focus on RBD ACE2 contact residue section (Figure 1) of the spike protein responsible for high affinity of the virus to humans. As mentioned before, all the virus strands can be traced back to SARS-2002 and all are mutated derivatives of SARS-2002 as seen in Table 1. We'll try to find the total number of combinations possible with 15 nucleotides of SARS-2002. Code for the same is available in Github repo[1].

SARS-2002	Y	L	N	D	T
Civet-2002	Y	L	K	D	S
Bat-2013	S	F	N	D	N
SARS-2	L	F	Q	S	N
position	1	2	3	4	5

Table 1: Accession in Coronaviruses' RBD ACE2 contact residue

2.4 Cyclic mutations in spike protein

Here we try to find the number of possible genomes that are up to 15 mutations away from the original length 15 genome mentioned above. The objective is to prove the vastness of possible mutations from a current genome sequence as small as of length 15. Each mutation leads to a new branch of mutations. This is referred to as mutation network.

2.5 Mutate from SARS-CoV-2 to SARS-2002

As an extension to the above objective, we can assume that any one of the possible mutations can be same as that of SARS-2002 which is associated with a far more fatality rate (close to 10%) compared to SARS-CoV-2.

2.6 Benign Mutations

Here we introduce another assumption of Benign mutations. Consider that b% of non-neutral mutations (change in amino acid) are benign for the evolution of virus. Rest (1-b)% will perish and cut off it's corresponding branch from the mutation network. We comment on which observed variants provide a possible benign path from SARS-CoV-2 to SARS-2002.

[1]https://github.com/ankushtale/mutation_network_SAR2COV_COVID19

3 ANALYSIS

In this section we will be discussing the solutions of the problems stated in the Section 2 along with the necessary assumptions we encapsulated in our solutions.

3.1 Possible RNA strands one mutation away

Here, we are assuming that the sequence is given excluding the stop and start protein strand which implies that mutations can occur at any given position of the sequence. Here, we are also not considering neutrality in the mutation. The mutations are assumed to take place at a given position and substitute the given nucleotide with another nucleotide in giving rise to a mutated sequence. Other types of mutations like Insertion and deletion are not considered for this calculations.

while we do not exclude neutral mutations which cause no change in the amino acids, we will consider missense mutation (one change in nucleotide) as a valid mutation. It means that given a nucleotide at a position, a mutation at that position will replace the nucleotide with any other nucleotide except the current one. Given all the above assumptions, there is a possibility of each nucleotide being replaced by 3 other nucleotides. So, the total number of possible sequence combinations which are 1 mutation away will be 3^{30000}.

3.2 Neutral mutations

In this problem, we have considered the SARS-CoV-2 sequence to be around 30000 nucleotides long. Considering the redundancies in creation of amino acids from codons, it is possible for a mutation to result in no change in the amino acid at the location. This relies on the initial state of the codons. For example, it is possible for all 10000 codons to be M (ATG) or W (TGG). In this case, any mutation in any location will result in a non-neutral mutation. So the lower limit of the fraction of neutral mutated RNA strands is 0. Consider a scenario where all the nucleotides are ACG in sequence. Now, if the mutation happens only in third place of every codon, then the total 27 possible mutations of a codon reduces to 9 (3*3 for first and second place). Then the total neutral mutations turn out to be 9^{10000} and the fraction is 3^{-10000}. If the mutation occurs on first or second place of every codon then there turn out to be no neutral mutations. Thus the neutral mutation cases depend on the initial state of mutations. A general derivation is that for codon groups of 4 on the third nucleotide (ACA, ACG, ACC, ACT), the total number of neutral mutations is 9^{10000}. For codon groups of 2 on the third nucleotide, the total number is 18^{10000}. We already discussed the case of 1, with the answer being 27^{10000} that is 3^{30000}. For mutation on second place in codon, it leads to no neutral mutations. For mutations on first place, it too leads to no neutral mutations except for the case of CG[A/G] and AG[A/G].

3.3 Spike Protein mutations

The RBD ACE2 contact residue section for SARS-2002 is Y442, L472, N479, D480, T487 and Y491 (the number denotes position in full genome sequence). Each of the amino acid mentioned here can be derived from multiple ways. For example, amino acid Leucine(L) can be encoded from CTT, CTC, CTG, CTA, TTA, TTG. If we consider all such possibilities of 5 amino acids we get 17 Adenine, 13 Thymine,

13 Cytocine, 5 Guanine bases in total. In this case, a recombination of all these nucleotides result in all 64 possible codons. On the contrary, if we assume just one possible encoding for every amino acid, then the number drastically reduces. For instance, let Y only be decoded from TAC, L from CTG, N from AAT, D from GAC and T from ACA. In this case, we have 6 Adenine, 3 Thymine, 4 Cytocine, 2 Guanine bases. This can be rearranged to form 19 amino acids 'K', 'N', 'N', 'K', 'I', 'I', 'M', 'T', 'T', 'R', 'F', 'F', 'L', 'S', 'S', 'W', 'P', 'P', 'R'. From this list, we can find 11 distinct amino acids.

3.4 Cyclic mutations in spike protein

Our analysis of the problem is two folds. Here, we are finding the possible number of genomes which are 1 to 15 mutations away from the SARS-2 spike protein. The sequence assumed for this was "TATTTGAACGATACC" where T, A, G and C refer to the nucleotides.

For our first analysis, we will not exclude any neutral mutations in the sequence. So, we will follow a combinatorial formula to calculate number of possible sequence which are n mutations away where n lies in the range [1,15].

Number of sequence n mutations away = $15^n * 3^n$

The above equation signifies that for each mutation there is a option of 15 nucleotides which can be mutated and out of which for each nucleotide, there are 3 options for each position. The plot for this equation in the given range of n is given in figure 1 where we plot value of n, i.e the number of mutations, on x axis and log scale of the possible number of sequences corresponding to the given number of mutations on y axis.

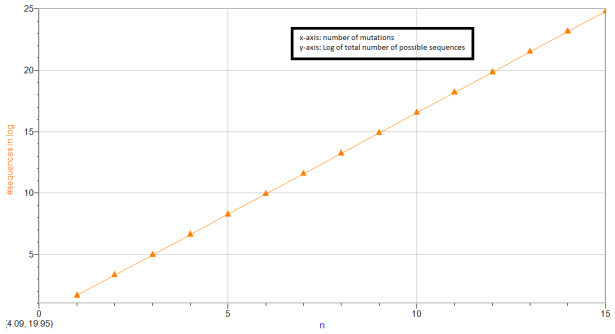

Figure 2: Log scale analysis of possible sequences which are n mutations away from SARS-2 spike protein where n lies in range [1,15]

For our second analysis, we have considered neutrality as a factor in deciding the relevance of the mutated sequence. In our approach, we find all the possible sequences for a particular number of mutation and we remove the neutral mutations i.e where the change in nucleotide does not change the amino acid. Here, we have also included duplication removal. For example, a necleotide mutation may change amino acid from L to F and a second mutation may in turn again change the F amino acid to L. Since the computation and memory usage of this process was very exponential, we were only able to calculate the number of mutation that are 6 mutations away. After that, we have used curve fitting to approximate the curve function and using that we have extracted the number of possible sequences uptil 15 mutations away.

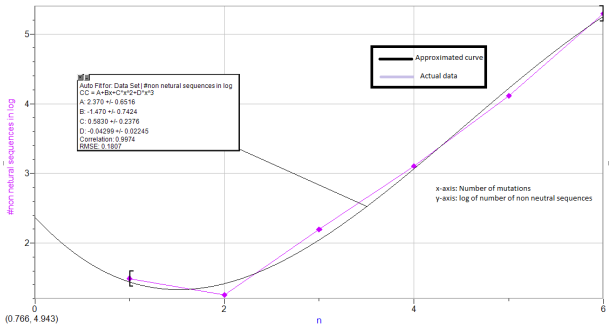

Figure 3: Log scale analysis of possible sequences which are n mutations away from SARS-2 spike protein where n lies in range [1,15]

As seen in figure 2, we were able to fit the given data to a cubic curve. The purple lines represent the original data while the black line represent the fitted curve. We have approximated the function as below,

$$y = 2.3 - 1.470 * x + .5830 * x^2 - 0.042 * x^3 \quad (1)$$

Equation 2 represents the approximated function which can give the approximate number of non neutral mutations which are n mutations away.

3.5 Mutate from SARS-CoV-2 to SARS-2002

In this problem, we are calculating the approximate time where SARS 2 virus will mutate into more lethal SARS 2002. For this, firstly we will need to assume the number of mutations required for SARS 2 to change into SARS 2002. Since there is no upper bound to how many mutations it will take, we are assuming the lower bound i.e. the minimum number of mutations for SARS 2 to roll back to SARS 2002. Table 2 shows the minimum number of mutations for each positional amino acid of SARS 2 to mutate into that of SARS 2002. So, the total number of mutations are calculated by summing the last column of Table 1. Hence, 8 minimum mutations are required for SARS 2 to become SARS 2002.

s2	s2_gene	**s1**	s1_gene	#min_mutations
L	CT[A/C/G/T], TT[A/G]	**Y**	TA[C/T]	2
F	TT[C/T]	**L**	CT[A/C/G/T], TT[A/G]	1
Q	CA[A/G]	**N**	AA[C/T]	2
S	AG[C/T], TC[A/C/G/T]	**D**	GA[C/T]	2
N	AA[C/T]	**T**	AC[A/C/G/T]	1

Table 2: Minimum mutations required for each SARS-2 (s2 column) amino acid to mutate into SARS-2002(s1 column) amino acid.

Now, we will delve into the calculations of number of days for this to happen. Our calculations are based on the assumptions

of Bedford's twitter thread [4], where we are assuming that one mutation happens in 7 days in the 30000 nucleotide long sequence. Now, since our spike protein is only 15 nucleotides long, we will find the probability of of spike protein nucleotide to be mutated. The calculations that we have conducted are in the following order.

- Probability of nucleotide in spike protein mutating is 15/30000 or 1/2000
- So, 1 in 2000 mutations occur in spike protein.
- Since, 1 mutation takes 7 days then 2000 will take 14000 days. Out of which 1 will be in the spike protein (one of the 15 nucleotides of spike protein) on an average.

So, for 8 minimum mutations calculated from Table 1, 8*14000 equals 112,000 days or 306 years approximately for SARS 2 to mutate to SARS 2002. Here, we are also assuming that the carriers of mutated strands (newly infected human beings) will live long enough to further transmit another mutated strand to a different host.

3.6 Benign Mutations

This problem is a continued version of previous problem but here we are assuming neutral networks and considering the beneficial paths for the spike protein of SARS 2 to change into SARS 2002.

Firstly, we will calculate the likelihood of SARS 2 to directly mutating to SARS 2002. As shown in Table 1, this will take 8 minimum mutations. Here, 8 amino acids are beneficial out of total of 20 amino acids. Therefore, we can assume that only 40 percent(8/20) of the mutations are beneficial. Thus, using the calculations from above problem we got 112000 days for SARS 2 to convert to SARS 2002. But, now only 40 percent of these mutations can be considered successful. Incorporating this into the total number of days gives 305000 days or 835 years approximately. This is the amount of time it will take for SARS 2 to mutate into SARS 2002 if only 40 percent of mutations are beneficial.

s2	s2_gene	c	c_gene	s1	s1_gene	min_path	#min_mutations
L	CT[A/C/G/T], TT[A/G]	Y	TA[C/T]	Y	TA[C/T]	_C_=>_A_=>_A_	1
F	TT[C/T]	L	CT[A/C/G/T], TT[A/G]	L	CT[A/C/G/T], TT[A/G]	__[C/T]=>__[A/G]=>__[A/G]	1
Q	CA[A/G]	K	AA[A/G]	N	AA[C/T]	C_[A/G]=>A_[A/G]=>A_[C/T]	2
S	AG[C/T], TC[A/C/G/T]	D	GA[C/T]	D	GA[C,T]	AG_=>GA_=>GA_	2
N	AA[C/T]	S	AG[C/T], TC[A/C/G/T]	T	AC[A/C/G/T]	_A_=>_G_=>_C_	2

Table 3: Minimum mutations required for each SARS-2 (s2 column) amino acid to mutate Bat-2013(c column) and further into SARS-2002(s1 column) amino acid.

s2	s2_gene	c	c_gene	s1	s1_gene	min_path	#min_mutations
L	CT[A/C/G/T], TT[A/G]	Y	TA[C/T]	Y	TA[C/T]	_C_=>_A_=>_A_	1
F	TT[C/T]	L	CT[A/C/G/T], TT[A/G]	L	CT[A/C/G/T], TT[A/G]	__[C/T]=>__[A/G]=>__[A/G]	1
Q	CA[A/G]	K	AA[A/G]	N	AA[C/T]	C_[A/G]=>A_[A/G]=>A_[C/T]	2
S	AG[C/T], TC[A/C/G/T]	D	GA[C/T]	D	GA[C,T]	AG_=>GA_=>GA_	2
N	AA[C/T]	S	AG[C/T], TC[A/C/G/T]	T	AC[A/C/G/T]	_A_=>_G_=>_C_	2

Table 4: Minimum mutations required for each SARS-2 (s2 column) amino acid to mutate Civet-2002(c column) and further into SARS-2002(s1 column) amino acid.

Now, we will incorporate paths of Bat-2013 and Civet-2002. We will follow the same procedure as problem 5 but the likelihood of mutations will change. Table 2 shows the minimum number of mutations for each amino acid of SARS 2 to mutate into Bat-2013 amino acid and then to SARS 2 amino acid. Summation of the last column gives a total of 9 minimum mutations by following this path. The calculation of time taken for each mutation with the likelihood of beneficial mutations is given below.

- Probability of nucleotide in spike protein mutating is 15/30000 or 1/2000
- So, 1 in 2000 mutations occur in spike protein.
- Since, there are 8 amino acids beneficial to all 3 of these spike proteins, the likelihood of a mutation being beneficial is 8/20 or 04. So, 1 in 2000/0.4 or 5000 mutations can be considered to take place in spike protein.
- Since, 1 mutation takes 7 days then 5000 will take 35000 days. Out of which 1 will be in the spike protein (one of the 15 nucleotides of spike protein) on an average.

So, for 9 minimum mutations calculated from Table 2, 9*35000 equals 315,000 days or 1029 years approximately for SARS 2 to mutate to SARS 2002.

Similarly, we will calculate the likelihood of SARS 2 to mutate back into Civet-2002 and then into SARS 2002.Table 3 shows the minimum number of mutations for each amino acid of SARS 2 to mutate into Civet-2002 amino acid and then to SARS 2 amino acid. Summation of the last column gives a total of 9 minimum mutations by following this path. The calculation of time taken for each mutation with the likelihood of beneficial mutations is given below.

- Probability of nucleotide in spike protein mutating is 15/30000 or 1/2000
- So, 1 in 2000 mutations occur in spike protein.
- Since, there are 9 amino acids beneficial to all 3 of these spike proteins, the likelihood of a mutation being beneficial is 9/20 or 045. So, 1 in 2000/0.45 or 4446 mutations can be considered to take place in spike protein.
- Since, 1 mutation takes 7 days then 4445 will take 31115 days. Out of which 1 will be in the spike protein (one of the 15 nucleotides of spike protein) on an average.

So, for 9 minimum mutations calculated from Table 2, 9*31115 equals 280,035 days or 767 years approximately for SARS 2 to mutate to Civet-2002 and then to SARS 2002. This leads us to infer that not only SARS 2 is very unlikely to mutate back into SARS 2 but if we consider that only a small portion of non neutral mutations are beneficial then this likelihood further decreases. Also, the likelihood of SARS 2 mutating directly into SARS 2002 is more than it following a path from Civet-2002 or Bat-2013. Hence, the neutral paths leading from these spike proteins are least likely, according to our calculations. To further add, we can consider this scenario as worst case because we have considered the minimum number of mutation in the spike protein, which is hardly ever the case in real world scenarios, and also we are assuming that each transmitted genome continues to live which is very unlikely because the life expectancy of human transmitters is way less than the above calculated time for the mutations to be successful.

3.7 Miscellaneous

Apart from this, we have also performed some further analysis which was not stated in the problem statement. Considering the rapid growth in number of Covid-19 cases and the current debates for and against the need for social distancing, we have analysed the

amount of time it will take for the entire population of the planet to be infected from the virus given the current trends in the increase in number of new cases. We have collected data of 118 days and predicted the amount of time it will take to spread across every person of the planet. The first day starts from 12/31/2019. We have used curve fitting to approximate the total number of cases on each day.

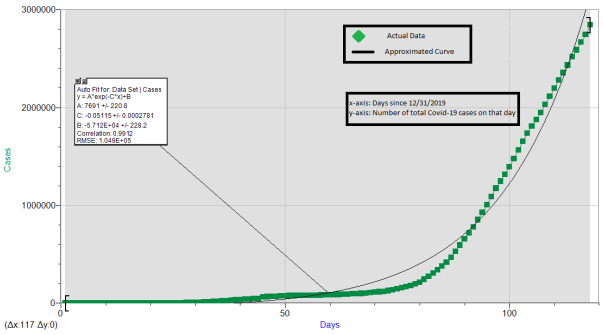

Figure 4: Total number of Covid 19 cases on each day from 12/31/2019 to 04/26/2020 i.e. 118 days

As shown in the Figure 4, the x axis depicts the day and y axis has the total number of cases on that particular day. The green line depicts the actual data while the black line depicts the approximate curve fitted to the data. Using the extracted equation we have estimated that if current trends continue, the entire population of 7.8 billion people will be infected in total of 270 days. We will like to make a note here that this is naive version of what the actual scenario depicts because we are not taking into account the current international border restrictions and the intensity of social distancing norms which are being followed. Nevertheless, our predictions can be used as a base worst case scenario to try and bring awareness about the spread of this infectious disease.

4 CONCLUSION

In this paper, we have analysed the effects of mutation on the strands of SARS-2002 and SARS-CoV-2 genome. There are a lot of factors which influence the transmission of the virus and its survival. Firstly, mutation itself is not a sole decision maker in the effects of the virus. We need to consider the neutral mutations as well as factor in the likelihood of mutation being non detrimental. We have also analyzed the current trend in the spread of this virus and how we are very close to a planet wide spread if the necessary precautions are not taken. Finally, on a positive note, we have depicted that the likelihood of this SARS-CoV-2 virus mutating into its more lethal predecessor, SARS 2002 is quite low. So, we can infer that a vaccine in the near future will be enough to stop the spread of this virus without worrying about mutations of the strand into more deadly forms. There are a lot of assumptions made in this paper and for future work we can incorporate latent variables of these underlying assumptions to create better model and predict more accurate results of our existing work.

REFERENCES

[1] [n.d.]. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, co - Nucleotide - NCBI. https://www.ncbi.nlm.nih.gov/nuccore/NC_045512
[2] 2020. Angiotensin-converting enzyme 2. https://en.wikipedia.org/wiki/Angiotensin-converting_enzyme_2
[3] Kristian G. Andersen, Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. 2020. The proximal origin of SARS-CoV-2. *Nature Medicine* 26, 4 (2020), 450–452. https://doi.org/10.1038/s41591-020-0820-9
[4] Trevor Bedford. 2020. A Thread on SARSCoV2 Mutations and What They Might Mean for the COVID19 Vaccination and Immunity, in Which I Predict It Will Take the Virus a Few Years to Mutate Enough to Significantly Hinder a Vaccine. 1/12. (2020). https://doi.org/trvrb/status/1242628550563250176
[5] Kristine Moore, Marc Lipsitch, John Barry, and Michael Osterholm. 2020. COVID-19: The CIDRAP Viewpoint. https://www.cidrap.umn.edu/covid-19/covid-19-cidrap-viewpoint
[6] Milton H. Saier. 2019. Understanding the Genetic Code. https://jb.asm.org/content/201/15/e00091-19