# CS523 Complex Adaptive Systems Project 3: SARS-CoV Genetic Algorithm

1st Alana Chigbrow
*Dept. of Computer Science*
*University of New Mexico*
Albuquerque, USA
achigbrow@unm.edu

2nd Jason Stewart
*Dept. of Computer Science*
*University of New Mexico*
Albuquerque, USA
jastewart@unm.edu

***Abstract*—In this project we examine the role of information and randomness in a Parallel Terraced Search as employed by a Genetic Algorithm. The Genetic Algorithm evolves the SARS-CoV-2 Receptor Binding Domain (RBD) from a variety of other RBDs. It uses either a purely random mutation method or a weighted random choice based on the probability of amino acid change given a single mutation in the underlying genome. We find that the informed-random approach to searching performs worse by far in comparison to the purely random approach. We also find that a crossover of Bat and Pangolin coronaviruses performs best under both random and informed conditions with regard to generating the SARS-CoV-2 RBD.**

## I. Introduction

In this project, we explore the way in which Complex Adaptive Systems (CAS) process information. Specifically, we seek to understand the role of randomness in a parallel terraced scan (PTS). A PTS involves a random, breadth-first search followed by focused, deep exploitation of ”useful” data [6]. Ants perform PTS in their search for food. They begin by randomly exploring (walking about with no apparent pattern) outside their nest until they find a strong enough pheromone trail to follow and exploit the food data encoded in that trail [6]. Recent work by Stefan Popp and Anna Dornhaus found that the random-walk of the ant species *Temnothorax rugatulus* was not entirely random [10]. Specifically, they found that the ants they observed ”meandered” and employed a pattern of direction changes that was not wholly random and which helped the ants to avoid crossing their own tail [10]. So, how random is the randomness of a PTS? Can additional information help the PTS to perform better in its search for useful data?

To answer this question, we developed a Genetic Algorithm (GA) with both purely-random and informed-random elements. A GA performs a PTS by first generating a large population of random candidate solutions (breadth-first) and then extracting the most fit candidate solutions from that population to subject to further mutation testing (focused) [6]. Given the recent renewed debate over the origins of the SARS-CoV-2 virus [13], the fact that the origin of the virus has not yet been definitively identified, and some experience in our previous project, we chose to develop a GA that ”evolves” the Receptor Binding Domain (RBD) of the SARS-CoV virus, the Bat (RaTG13) and the Pangolin (PCoV_GX) coronaviruses into the RBD of SARS-CoV-2. We chose to use the SARS-CoV RBD in this project because, despite being fairly different when compared to SARS-CoV-2 and from different lineages, they both bind the ACE2 using the RBD [3]. Both the Bat and Pangolin coronaviruses have been noted for their genomic similarity to SARS-CoV-2 [11] with the Bat Coronavirus being most often cited as the likely origin due to its approximately 96% genomic match to SARS-CoV-2 [11]. We also include crossovers of the different RBDs based on the work of Pekar, et al. indicating the likelihood of multiple zoonotic crossovers as the origins of SARS-CoV-2 [9].

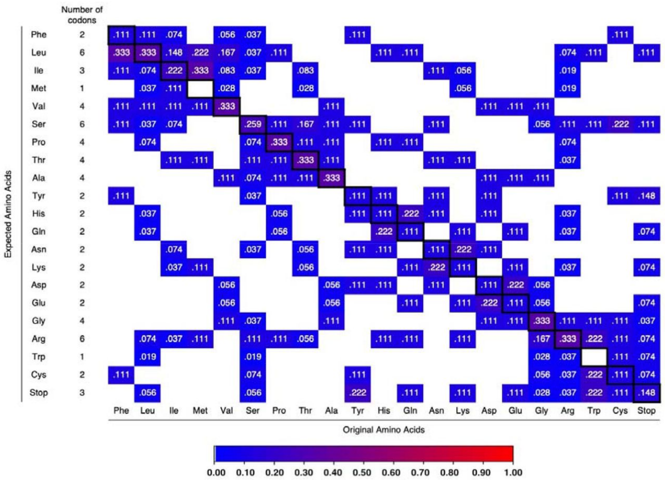

Fig. 1. Probability of Amino Acid Change [15]

During work on our previous project, we successfully predicted the percentage of neutral protein genomes within *at worst* 1% difference in the expected and actual percentages using the probability of amino acid self-preservation given a Single Nucleotide Substitution (SNS) [15]. We expand upon that idea here by using the full probability of amino acid change given an SNS (see Figure 1, rather than just the likelihood of self-preservation, in order to create an informed-random method for mutating a protein. We then compare the result of using a purely random approach to mutation and the informed-random method to determine whether or not additional information can make a PTS more efficient.

## II. Methods

### A. Data Preparation

We elected to work with the spike proteins of each of the coronaviruses directly rather than mutating the underlying genomes. The reason for this is that the proteins for all of the viruses were readily available through the National Library of Medicine online database [8]. The exact residue locations of the RBD is annotated in the FASTA files found in that database, making it much easier to identify the correct part of the amino acid chain to mutate. We experienced difficulty in finding the corresponding nucleotide positions for the RBDs. We deemed the direct mutation of the protein to be sufficient for the purposes of this project.

All RBDs contained 223 amino acids except for the SARS-CoV RBD, which contains 222. For ease of calculation of fitness when comparing it and its mutated offspring to the SARS-CoV-2 RBD, we expand the SARS-CoV RBD by 1 additional amino acid. The method for doing this was to randomly select a location in the RBD, randomly select an amino acid, and then insert that amino acid at the selected location. This change was made with a purely-random approach since we were inserting, rather than swapping, an amino acid.

The probability of amino acid change given an SNS included the probability that the amino acid would become a stop codon [15]. We choose to omit the stop codons from our GA. The probability that an amino acid would become a stop codon was divided equally by the number of amino acids it could become with greater than 0% probability and then added to those probabilities.

### B. Crossover

Crossover was used in two senses: first in generating the initial population of RBDs with more than 1 source RBD (e.g. Bat crossed with Pangolin coronavirus) and second as part of the process of creating new mutations of each successive generation. We implemented a 2-point crossover that resulted in 1 child RBD for each set of parent RBDs. The parents were divided into 3 parts each and those 6 parts were divided between the children. 80% of each generation was subjected to crossover rather than mutation.

While the focus of our work was on mutation, we did want to mention results we experienced when finding the optimal crossover percentage. In their work on performing computations using cellular automata, Mitchell, Crutchfield, and Hraber found that turning crossover off completely resulted in only 44% of their initial population reaching epoch 3 in their GA [5], thus indicating that crossover is important to the success of a GA. We too found this to be true. We originally set the percent of the population to be subjected to crossover at a conservative 0.30. Under this setting the GA was only able to achieve at best between 0.919 and 0.942 fitness after 3000 generations. We arrived at an optimal crossover percentage of 0.80 and were able to achieve perfect matches from every candidate RBD under random mutation.

### C. Mutation: Choosing Amino Acids

Mutation was used both to generate the initial candidate population of a single source RBD iteration and to mutate 20% of the top 50% most fit genomes in each generation to produce the next generation. We implemented two approaches to this mutation. The first approach is mutating the protein purely randomly. (Or, as random as we can get with our computers) In this approach, a residue position is randomly selected and then a new residue is likewise randomly selected from the 20 available amino acids to fill that position. The second approach to mutating is an ”informed randomness”. The position selection is still purely random, but the choice of the new residue incorporates probabilities regarding the likely outcome of what an amino acid will become given a Single Nucleotide Substitution (SNS) as found by Chan, et al. in [2]. We generate a weighted random choice of amino acid changes based on the outcome of an SNS that is specific to the residue being mutated and use that list to generate its replacement. In Section III, we discuss the relative performances of both approaches. We mutated 2 amino acids per round of mutation based on one of the two approaches described above.

### D. Calculating Fitness

The fitness of each new RBD generated by the GA was determined by a simple residue by residue comparison with the SARS-CoV-2 RBD. The number of residues that were shared across both RBDs was divided by the total residues (223) to calculate a percent match between the two strings.

### E. Implementation Details of the Genetic Algorithm

#### 1) Running our code

The code for our GA is contained in the *src* directory of our GitHub repository [14]. The driver script for our GA is located in *run_sars_ga.py*. At the beginning of the main function there are five parameters that can be modified, stored in the *params* dictionary. These parameters are as follows:

- informed: True or False, whether to perform informed mutations. For more details, see Section II-E2.
- gen_size: The number of amino acid strings in each generation
- gen0: The RBD(s) to seed the first generation. The amino acid sequences of the RBDs are found in *fasta/sequence.fasta*. We obtained these sequences and the locations of the RBD from the National Library of Medicine [8]. The links to the amino acid sequences are documented in the README file in our GitHub repository [14].

  The amino acid sequence for SARS-CoV-2, which is used to determine the fitness of each generation, is the second amino acid sequence in *fasta/sequence.fasta*. The RBD is from amino acid positions 318 to 541. Seeding Options:
  - **1**: SARS coronavirus Tor2 (SARS-CoV-1) — The first sequence in *fasta/sequence.fasta*. The RBD is from amino acid positions 305 to 527.

  - **b**: Bat coronavirus RaTG13 — The third sequence in *fasta/sequence.fasta*. The RBD is from amino acid positions 318 to 541.
  - **p**: Pangolin coronavirus — The fourth sequence in *fasta/sequence.fasta*. The RBD is from amino acid positions 316 to 539.
  - **bp**: Bat/Pangolin crossover — A 50/50 crossover between the Bat and Pangolin RBDs.
  - **b1**: Bat/SARS-CoV crossover — A 50/50 crossover between the Bat and SARS-CoV-1 RBDs.
  - **p1**: Pangolin/SARS-CoV crossover — A 50/50 crossover between the Pangolin and SARS-CoV-1 RBDs.
- max_gens: How many generations to evolve.
- ga_runs_per_setting: How many times to run the genetic algorithm with the settings above, to observe the variance.

Running *run_sars_ga.py* will generate a single CSV file in the data directory that concatenates the results of all *ga_runs_per_setting* simulations. These results can be plotted by running *run_sars_analysis.py*, which produces the figures in Section III. The figures are saved in either *figures/draft* or *figures/final* depending on the setting of the *DRAFT* variable at the beginning of *run_sars_analysis.py*.

### *2) How proteins are evolved from one generation to the next*

A new generation is created with the following steps:

1) The bottom 50% least-fit genomes are removed.
2) The remaining 50% of amino acid sequences are split into two groups:
   - Sequences that will be crossed-over with each other to generate part of the next generation. (80% of the remaining amino acid sequences). The crossover algorithm is described in Section II-B.
   - Sequences that will mutated, either randomly or informed, to generate part of the next generation. (20% of the remaining amino acid sequences) The crossover algorithm is described in Section II-C.
3) Using the methods in 2), amino acid sequences are generated until the generation size is met.
4) The maximum and average fitness is calculated for the generation. If the maximum fitness reaches 1 (which means a amino acid sequence has evolved into the SARS-CoV-2 sequence), the algorithm terminates.

## III. Results

In this section, we will discuss our results from GAs seeded with either a SARS-CoV, Bat, Pangolin, Bat/Pangolin crossover, Bat/SARS-CoV crossover, or Pangolin/SARS-CoV crossover RBD. In all figures, the ”True” line is for informed mutations and the ”False” line is for random mutations.

All results were generated with GAs that have 500 proteins per generation and evolve 3000 generations at maximum.

Figures 2, 3, 4, 5, 6, and 7 show the results of random versus informed mutations for GAs seeded with different

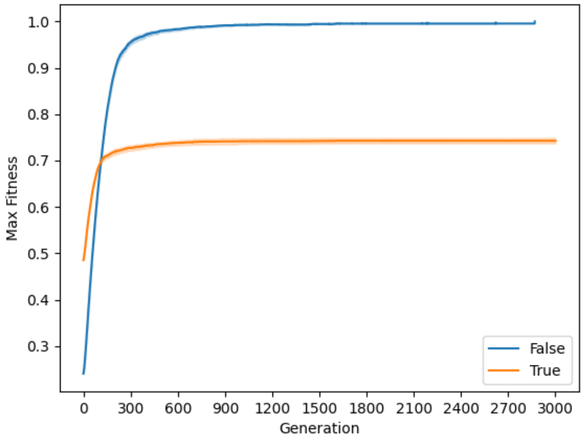

Fig. 2. Comparison of SARS-CoV seed RBD with random and informed-random mutations

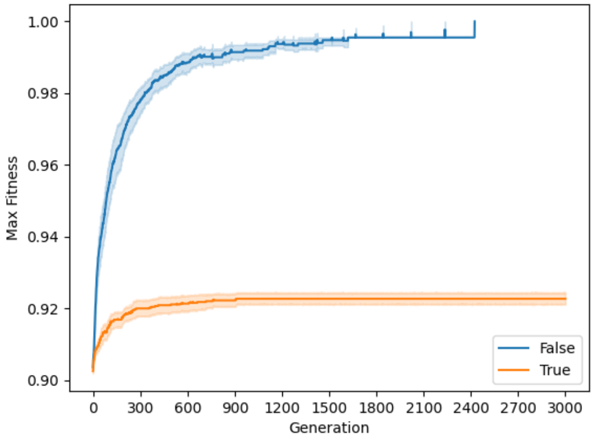

Fig. 3. Comparison of Bat seed RBD with random and informed-random mutations

TABLE I
GA PERFORMANCE WITH DIFFERENT GENERATION 0 SEEDS AND INFORMED MUTATIONS

| Gen 0 Seed | Best fitness after 3000 generations |
|---|---|
| SARS-CoV | 0.778 |
| Bat | 0.933 |
| Pangolin | 0.937 |
| SARS-CoV X Bat | 0.933 |
| SARS-CoV X Pangolin | 0.946 |
| Bat X Pangolin | 0.955 |

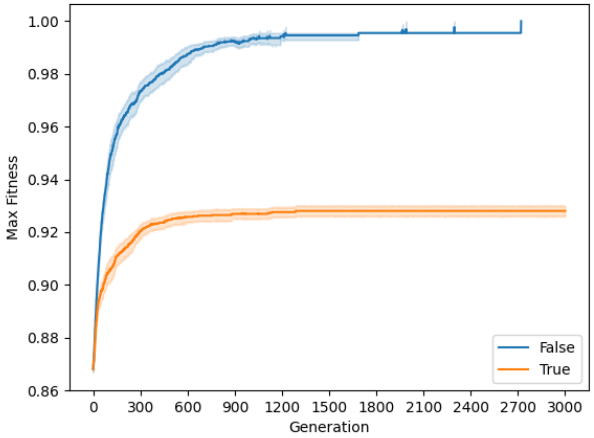

Fig. 4. Comparison of Pangolin seed RBD with random and informed-random mutations

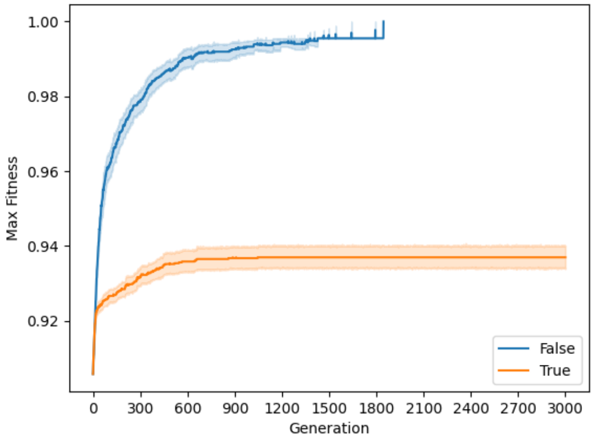

Fig. 5. Comparison of Bat/Pangolin crossover seed RBD with random and informed-random mutations

TABLE II
GA PERFORMANCE WITH DIFFERENT GENERATION 0 SEEDS AND RANDOM MUTATIONS (GENERATIONS TO REACH FULL FITNESS)

| Gen 0 Seed | Min | Avg | Max |
|---|---|---|---|
| SARS-CoV | 744 | 1619 | 2869 |
| Bat | 568 | 1237 | 2424 |
| Pangolin | 793 | 1296 | 2720 |
| SARS-CoV X Bat | 248 | 1282 | 2983 |
| SARS-CoV X Pangolin | 322 | 1230 | 2628 |
| Bat X Pangolin | 488 | 1200 | 1845 |

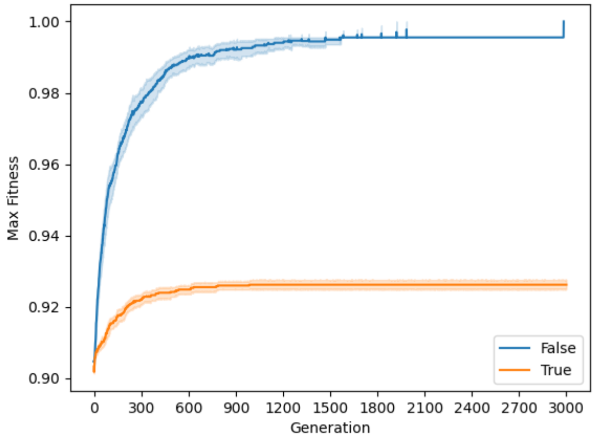

Fig. 6. Comparison of Bat/SARS-CoV crossover seed RBD with random and informed-random mutations

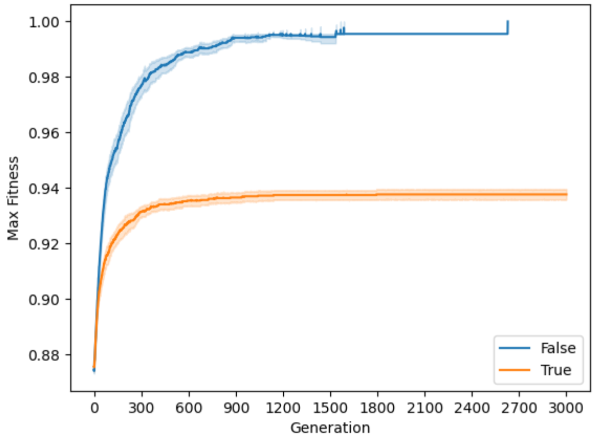

Fig. 7. Comparison of Pangolin/SARS-CoV crossover seed RBD with random and informed-random mutations

RBDs. Random mutations performed much better than informed mutations across all RBDs. In fact, as shown in Table I, GAs performing informed mutations were never able to reach a fitness of 1 after 3000 generations. On the contrary, as shown in Table II, every RBD was able to evolve to the SARS-CoV-2 RBD within 3000 generations if random mutations were performed. Another interesting observation is that GAs performing informed mutations reached its fitness plateau earlier than GAs performing random mutations across all simulations. This may be because since the amino acid substitutions are constrained to probabilities, it takes less time for the RBD sequences to converge to a dynamic equilibrium than RBDs that are mutated randomly.

We analyzed the top performing RBD in the informed-random mutation run that was generated using the Bat X Pangolin crossover. This RBD had 10 amino acid positions that differed from the SARS-CoV-2 RBD. Table III shows the

TABLE III

RESIDUE MISMATCHES BETWEEN THE BAT X PANGOLIN RBD AND THE SARS-COV-2 RBD AND THE PROBABILITY THAT GIVEN AN SNS, THAT THE CANDIDATE AMINO ACID WILL BECOME THE SARS-COV-2 AMINO ACID.

| Position | Candidate AA | SARS-CoV-2 AA | Probability |
|---|---|---|---|
| 324 | D | E | 0.111 |
| 346 | T | R | 0.056 |
| 372 | T | A | 0.111 |
| 445 | T | V | 0.000 |
| 483 | Q | V | 0.000 |
| 486 | L | F | 0.111 |
| 490 | Y | F | 0.111 |
| 493 | Y | Q | 0.000 |
| 494 | W | S | 0.111 |
| 519 | N | H | 0.111 |

positions, amino acids, and probability of matching the SARS-CoV-2 amino acid at the given position. The top performing candidate would need more than one additional round of mutation before it would become the SARS-CoV-2 RBD.

In light of how close the Bat X Pangolin crossover came to the SARS-CoV-2 RBD using informed-random mutations, we decided to test adjustments to other parameters to see if we could get a better match. We included the SARS-CoV RBD in the same experiments to compare the performance of both.

First, we increased the total number of generations to 5000 to see if the GA could arrive at an informed-random solution given more time (see Figure 8). It did not. The Bat X Pangolin seed hit a max fitness of approximately 0.924 fairly early in the generations, but the fitness did not increase from there. There were only 5 different max fitness values across 20 iterations of 5000 generations each.

The SARS-CoV RBD performed worse than the Bat X Pangolin crossover over 5000 generations, but still better its performance over 3000 generations with informed-random mutations. The SARS-CoV RBD managed to achieve a max fitness of approximately 0.83. This may be due to the ”luck of the crossover” in the the initial population — it had a much higher max fitness at generation 1 for the 5000-generation run, averaging 0.726.

We also adjusted the number of mutations per candidate RBD from 2 to 6. We chose a value of 6 because the completely random selection of a new amino acid corresponds to 3 mutations in the codon. That means that when we randomly mutate 2 amino acids, we are mutating 6 nucleotides. The underlying logic of our informed-random mutation is that a mutation of 1 nucleotide per codon mutating 2 amino acids corresponds to mutating 2 nucleotides. While not a perfect logical match, we wanted to see what would happen if we increased the number of sites that were mutated per RBD. Once again, the Bat X Pangolin crossover very quickly hit a 0.92 max fitness and did not increase from there. The SARS-CoV-seeded GA starts generation 1 at approximately 0.29 fitness and maxes out at approximately 0.686 fitness. These results are shown in Figure 9.

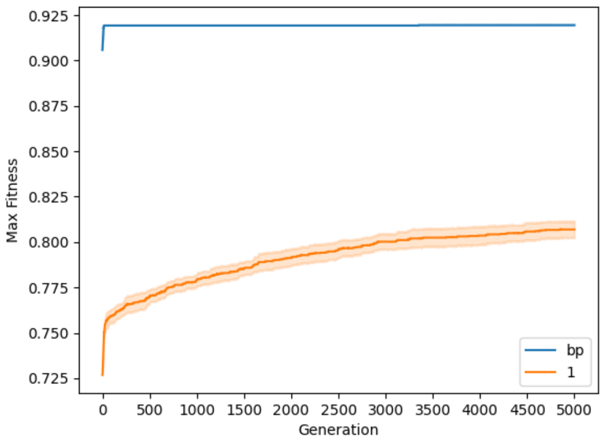

Fig. 8. Comparison of Bat X Pangolin crossover and SARS-CoV RBD seeds when both are using informed-random mutation over 5000 generations.

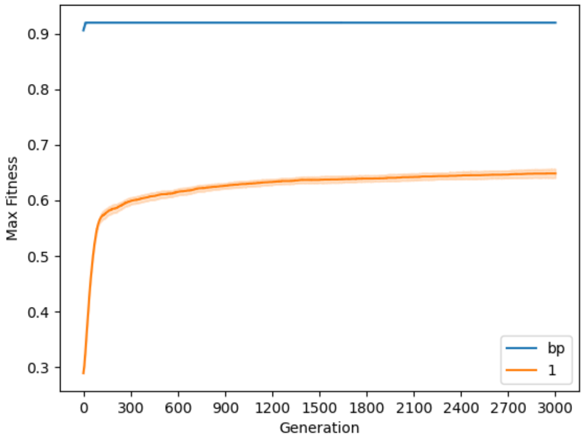

Fig. 9. Comparison of Bat X Pangolin crossover and SARS-CoV seeds under informed-random mutation with 6 total mutations.

## IV. CONCLUSION

### A. Discussion

First, we address our results in light of the ongoing debate as to the origin of SARS-CoV-2. We are in no way making a definitive claim as to its origin and acknowledge that the results produced by our experimental GA are not a basis for making any such claim. We do find the results interesting.

We found found that most of the available literature proposes a likely zoonotic origin for SARS-CoV-2 [9] [4] [12] as opposed to evolving from SARS-CoV. This makes sense given that the original virus was driven to extinction during the 2002-2003 pandemic it caused [4], it was itself a product of zoonotic crossover [4], and the fact that the two viruses are so dissimilar [3]. Our results align neatly with these facts as the SARS-CoV RBD as seed performed worst overall (taking much longer on average and max to match the SARS-CoV-2 RBD) and especially under informed-random mutation (reaching at best 0.778 fitness). These results are not entirely unexpected given that the SARS-CoV and SARS-CoV-2 RBDs have only 0.525 amino acids in common.

The bat and pangolin seed RBDs both perform as expected given their high initial fitness values (0.901 and 0.865 respectively) and the fact that both have been proposed as potential candidates for the origin of SARS-CoV-2 with an emphasis on the Bat Coronavirus has prime contender [11]. What is surprising to us was the fact that the Bat/Pangolin crossover performed best overall to include coming to within 1 amino acid difference from the SARS-CoV-2 RBD (0.955 fitness) under informed mutation. Further, the top 3 results under the informed-random mutation all involved the pangolin RBD as seed and the top 2 results were both crossovers. The better performance of the pangolin RBD seed (as crossover and as a single source in informed mutation) is interesting in light of the fact that pangolins have been proposed as an intermediate host for the virus because its RBD binds the human ACE2 receptor with stronger affinity that does that of the bat [1]. The fact that the top two performers in informed-random mutation and the top performer in random mutation were all crossovers also aligns well with the proposed multiple zoonotic crossover theory [9].

Second, we address the results in light of our primary question: can additional information guiding the randomness of a parallel terraced search make the search more efficient? It constrains the search space and thus makes arriving at the solution more difficult. This constraint resulted in a worse performance of the informed-random mutation when compared to the purely random approach. This was not what we expected, however it made perfect sense once we saw the results of the experiment. If an amino acid can become any other amino acid under mutation, it is easier to arrive at some target protein. If you restrain the amino acid change to what is actually possible you place limitations on the search that make it less likely to succeed.

In reality, an amino acid cannot simply become any other amino acid given an underlying genomic mutation *unless* all three of the nucleotides in the codon are substituted (3NS). Our random mutation approach mimics the effects of a 3NS by randomly selecting a replacement amino acid. Genetic mutation has random elements to it but it is not purely random. That means that though the informed-random mutation performs worse than the random mutation, it is more reflective of the actual process involved. Our informed-random mutation is, however, overly constrained as it only accounts for an SNS when it is possible that there could be 2 or even 3 nucleotide substitutions (2NS and 3NS respectively) in the same codon.

Randomness should be informed *if* the goal is to generate an actual solution. The trick becomes identifying what information/constraints are necessary and sufficient to the task. Consider the Space Technology 5 (ST5) evolved antenna in Figure 10. The antenna design was evolved by a GA and, according to the NASA article, "most human antenna designers would never think of such a design" [7]. Why? Because we are constrained by our experience, by the information we have. This is not what an antenna looks like (says the human designer). This type of information places artificial constraints on the search space that are not conducive to arriving at the useful data that should be exploited. The designers of the GA that evolved the ST5 antenna design properly identified the necessary and sufficient constraints for the search space. This should be the goal of every designer of a GA and we should take that into account in our everyday lives as we try to find solutions for the problems we face using our own parallel terraced searches.

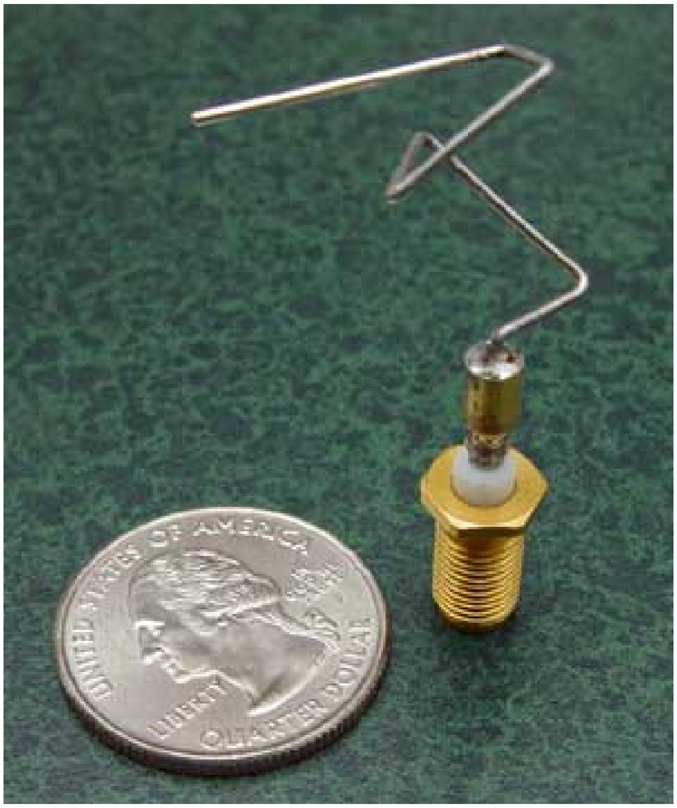

Fig. 10. ST5 Evolved Antenna [7]

### B. Future Work

The most interesting result that came about from the addition of the SARS-CoV RBD to the experiment is in the comparison of the random and informed-random mutation results and how they differ from the results of all of the other seed RBDs (both single source and crossover). Informed mutation of the SARS-CoV seed RBD produced candidate populations that very nearly approximated its original fitness when compared to SARS-CoV-2; the initial populations had a max fitness of approximately 0.485 compared to the 0.525 of the seed RBD. This stands out in stark contrast to the 0.241 max fitness of populations generated using random mutation of the SARS-CoV RBD (see Figure 2). No other seed RBD generated this sort of gap between the informed-random and randomly generated initial populations. Answering the question of why random mutation performed so much worse in the initial populations is a topic for future work. Though, given the fact that 1 of the 3 runs of informed mutations of the SARS-CoV RBD resulted in a lower initial max fitness, we think the answer may have something to do with the random magic of crossover.

In future work, the model can and should be expanded to increase its probabilistic realism. Specifically, we would like to:

- Identify a general mutational rate for RNA viruses and use this combined with the total number of nucleotides in the RBD to set the number of amino acid mutations per candidate RBD.
- Generate a probability state table that calculates the probability of amino acid change given 2NS.
- Identify the respective probabilities of SNS, 2NS, and 3NS.
- Use the above information to update the informed-random mutation by selecting the number of SNS, 2NS, and 3NS per candidate and using the appropriate amino acid selection based on how many nucleotides were swapped.

## V. Contribution Statement

The following statements are very broad. We both assisted each other with support functions and debugging and so there is very little work in our repo that was not touched by the both of us. Jason Stewart was primarily responsible for setting up the repo and the overall structure of the project, the Genetic Algorithm, the GAProtein Class, and the running of the Genetic Algorithm. Alana Chigbrow was primarily responsible for the GAGeneration Class, the utilities, and the analysis of the results.

The same statement regarding cooperative work applies to the paper as well. Jason Stewart set up the document, wrote the implementation details of the methods, and the bulk of the results. He also performed most of the editing to ensure readability and consistency. Alana Chigbrow wrote the introduction, subsections A-D of the methods, the results of the extended studies, and the conclusion.

## References

[1] P. Xu L. J. Calder A. Borg C. Roustan S. R. Martin P. B. Rosenthal J. J. Skehel A. G. Wrobel D. J. Benton and S. J. Gamblin. “Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2”. In: *Nature Communications* 12 (2021).

[2] Kwok-Fong Chan et al. “Probability of Change in Life: amino acid changes in single nucleotide substitutions”. In: (2020). DOI: 10.1101/729756.

[3] A. Fell. “Understanding the Evolution of SARS and COVID-19 Type Viruses”. In: *UC Davis News* (Feb. 2021).

[4] S. Alberti G. Poli I. Pagani S Ghezzi and E. Vicenzi. “Origin and Evolution of SARS-CoV-2”. In: *Springer Nature - PMC COVID-19 Collection*. 2nd ser. 138 (Feb. 2023).

[5] J. P. Crutchfield M. Mitchell and P. T. Hraber. “Evolving Cellular Automata to Perform Computations: Mechanisms and Impediments”. In: *Physica D* 75 (1994), pp. 361–391.

[6] Melanie Mitchell. *Complexity: A Guided Tour*. USA: Oxford University Press, Inc., 2009. ISBN: 0195124413.

[7] NASA. *Evolved Antenna*. https://www.jpl.nasa.gov/nmp/st5/TECHNOLOGY/antenna.html.

[8] *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/protein/.

[9] Jonathan E. Pekar et al. “The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2”. In: *Science* 377.6609 (2022), pp. 960–966. DOI: 10.1126/science.abp8337.

[10] S. Popp and A. Dornhaus. “Ants combine systematic meandering and correlated random walks when searching for unknown resources”. In: *iScience* 26 (2 Feb. 2023). DOI: 10.1016/j.isci.2022.105916.

[11] J. Yu J. Zeng S. Shan L. Tian J. Lan L. Zhang S. Zhang S. Qiao and X. Wang. “Bat and Pangolin coronavirus spike glycoprotein structures provide insights into SARS-COV-2 evolution”. In: *Nature News* (Mar. 2021).

[12] D. Singh and S. V. Yi. “On the origin and evolution of SARS-CoV-2”. In: *Experimental Molecular Medicine* 53 (2021).

[13] Cecelia Smith-Schoenwalder. *U.S. Agencies Divided Over COVID-19 ‘Lab Leak’ Origin Theory*. https://www.usnews.com/news/health-news/articles/2023-02-27/u-s-agencies-divided-over-covid-19-lab-leak-origin-theory. Feb. 2023.

[14] Jason Stewart and Alana Chigbrow. *CS523-project3-GA GitHub Repository*. https://github.com/JStewart28/CS523-project3-GA (private) zipped file of repository included in submission.

[15] Y. K. Mohanta T. K. Mohanta A. K. Mishra and A. Al-Harrasi. *Virtual 2D mapping of the viral proteome reveals host-specific modality distribution of molecular weight and isoelectric point*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8553790/. Oct. 2021.