Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

INTRODUCTION:

A genetic algorithm (GA) solves both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. This application simulates the evolution of variants of a positive-sense single-stranded RNA virus i.e., SARS-CoV-2. It includes three phases of evolution: the initial phase, vaccination phase, and the end phase, where there might be a possible delta variant.

- **Aim:**
  To study the evolution of variants of SARS-cov-2 using Genetic Algorithm. To simulate reproduction
  of virus in series of hosts who are either
  (a). Naive (not infected, not Vaccinated)
  (b). Previously Infected
  (c). Vaccinated
  (d). Previously Infected and vaccinated.

- **Approach:**
  - To start the simulation, create a virus and infect one person in random with that virus.
  - The fitness of human is calculated by using the ascii value of their genotype.
  - The fitness of virus is calculated by based on the correct positioning of 10 genome.
  - If the infected person meets another person who is not infected, then spread
  - the virus and perform mutation of virus in each of the parent.
  - If the infected person meets another infected person, then perform recombination
  - People who are infected start fighting with the virus.
  - Chances of person falling sick, or death or recovery is based on the probability at which it effects their health
  - which is based on their genome. These probability values are stored in Hashtable for different variants
  - A new variant of virus is said to be formed when the fitness of virus found in person is greater by some value k than the virus
  - last stored in the hash table and add the new variant to the hash Table.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

PROGRAM

Data Structures and Classes

Data Structures used in this project are:

- **HashTable:** We have used hash table to store different variants of virus encountered and array list of values which indicate the probability or chances by which it effects different genotypes.
- **HashMap:** We have used hash map to store different Variants with its respective fitness value.
- **ArrayList:** We have used array list to store population of 1000 people.
  To store values which indicate the probability or chances by which it effects different genotypes.
- **Array:** We have used array to store and represent Virus gene structure and Human gene structure

The classes used in this project are:

- **SimulationJPanel Class:** In this class, simulate method is declared using which the application runs by making internal calls to different classes
- **RuntimeAttributes Class:** This class contains static variables that are used during different runtime phases.
- **Gene Interface:** This interface consists of abstract methods which are implemented by their childrenclasses that are HumanGene and Virus Gene. This interface also consists of nucleotides of Humans and Virus.
- **HumanGene class:** This class implements gene interface and is used to calculate fitness of hosts.
- **VirusGene Class:** This class also implements gene interface and is used to calculate virus fitness. Moreover, it has functionalities for mutation and recombination of the virus.
- **Person Class:** This class is used to represent hosts. It has objects of human gene and virus gene as instance variables. Moreover, it has functionality for fighting the virus. This class is also used to categorize population based on genotypes.
- **Population Class:** In this class an array list of persons data structure and is used to initialize host population.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

Algorithm

Step 1: Initialize population with 1000 persons.
Step 2: Initialize a virus variant and add it to the hash table. Calculate the probability at which the initialized virus variant affects different host Genome and add those values
to the hash table.
Step 3: Select one person in Random from the population and infect that person with the created virus.
Step 4: While number of days reaches 731 do:
1. for each person in the population,

- If person is infected, then mutate the virus in the host and person starts fighting virus.
- If person fights virus for 4 weeks, then mark person as recovered and increase fitness of the person by 5000 units to indicate the development of antibodies against the virus.

2. If number of days > 365 then start vaccination in people. select 10 people per day and mark them as vaccinated and increase fitness of person by 5000uts.

3. Select two persons (A, B) in random

- If A & B are infected and come in contact, perform recombination of two virus present in A and B and create a progeny. Infect the progeny in any one of the parents

- If only A is infected, then mark B as infected and mutate virus in A and B. Hosts A, B start fighting virus. Check for a new Variant in A, If present add it to the fitness table along with the probability at which the new variant affects different host Genome and add those values to the hash table.

- If only B is infected, then mark A as infected and mutate virus in B and A. Hosts A, B start fighting virus. Check for a new Variant in B, If present add it to the fitness table along with the probability at which the new variant affects different host Genome and add those values to the hash table.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

Algorithm to check for new Variant in a host.

If the current fitness of virus present inside the host is greater than the fitness of last added variant in the hash table by a factor of k then, update Fitness Table.

Procedure to update Fitness Table:

Step 1: Calculate the increase in virus fitness by finding the difference between the virus fitness inside the host and fitness of last added variant in the hash table.

Step 2: Calculate percentage increase in fitness of virus.

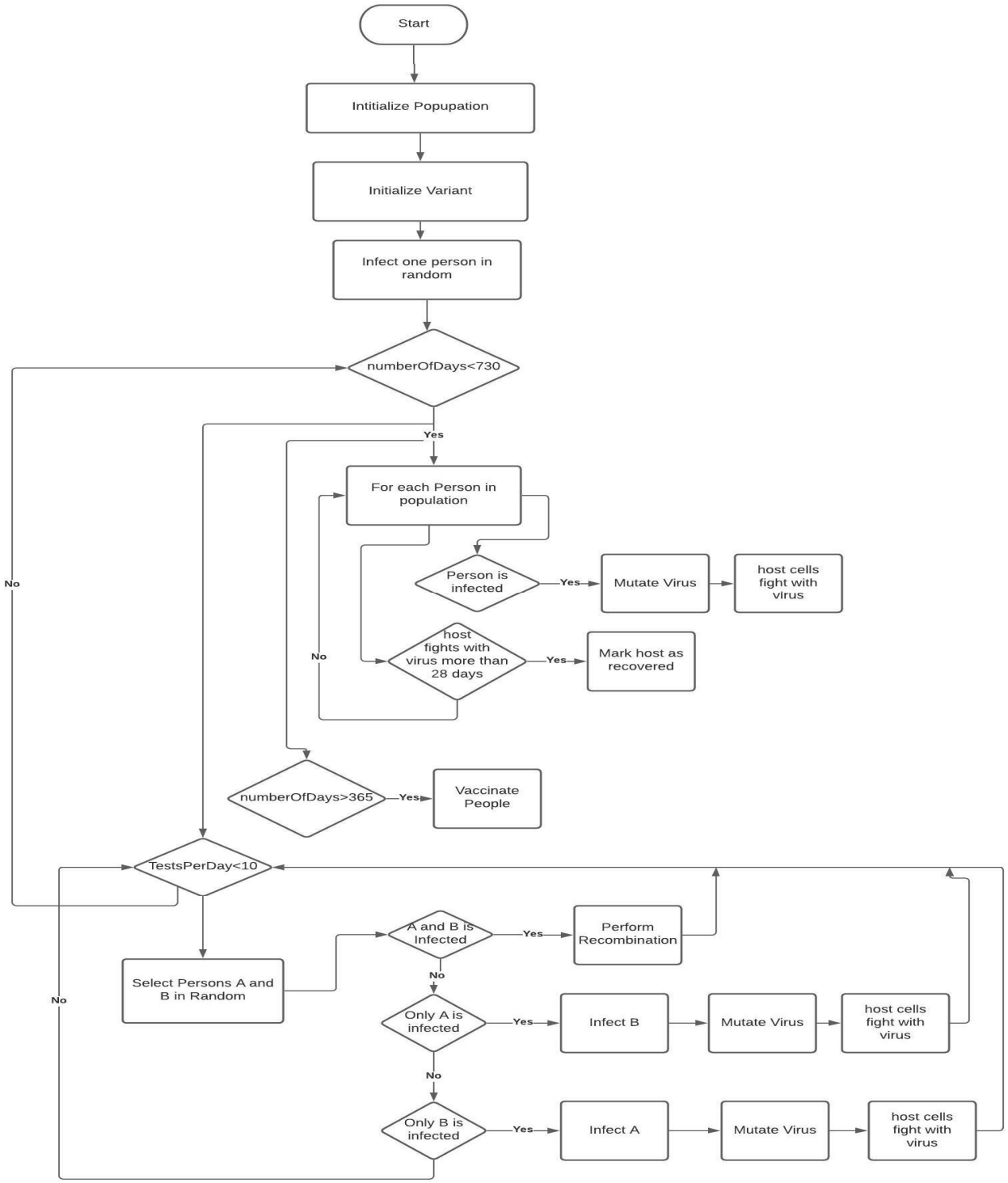
Step 3: Based on the calculated percentage, increase the probability at which the last added variant affects different host genotypes and add the new values to the hash table.

Invariants

- Threshold of fitness of Naive Humans: 8000
- Threshold of fitness of Recovered and vaccinated Humans: 4000
- K_Factor: 5
    - K factor is the difference between the virus fitness inside the host and fitness of last added variant in the hash table.
- Probability at which it effects different genomes:
    - A1: 2/20
    - A2: 1/20
    - B1: 5/20
    - B2: 12/20
- Decrease in probability of different Genomes being affected by 50% after being recovered as antibodies are developed inside them
- Decrease in probability of different Genomes being affected by 25% after being vaccinated.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

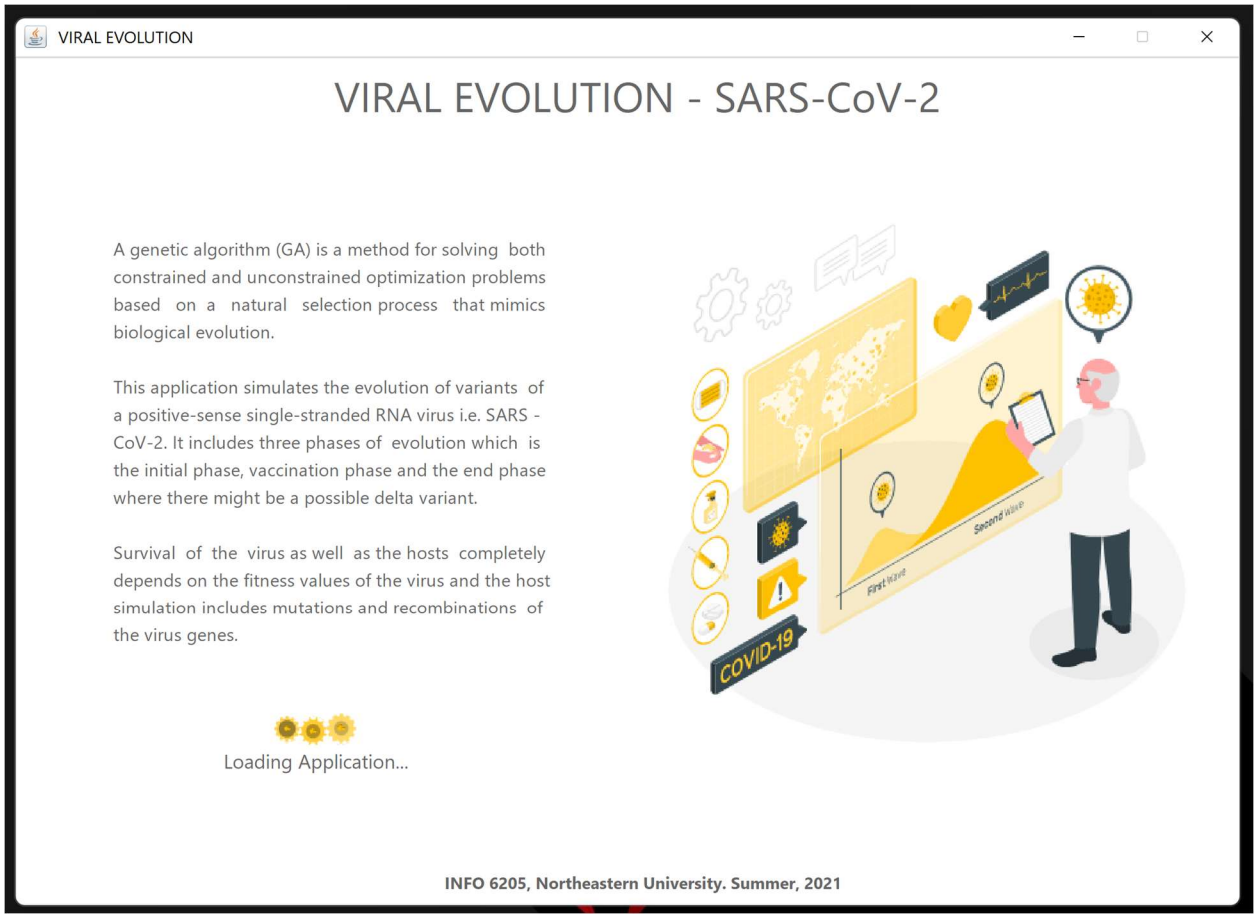
FLOWCHART



Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

UI Flow:

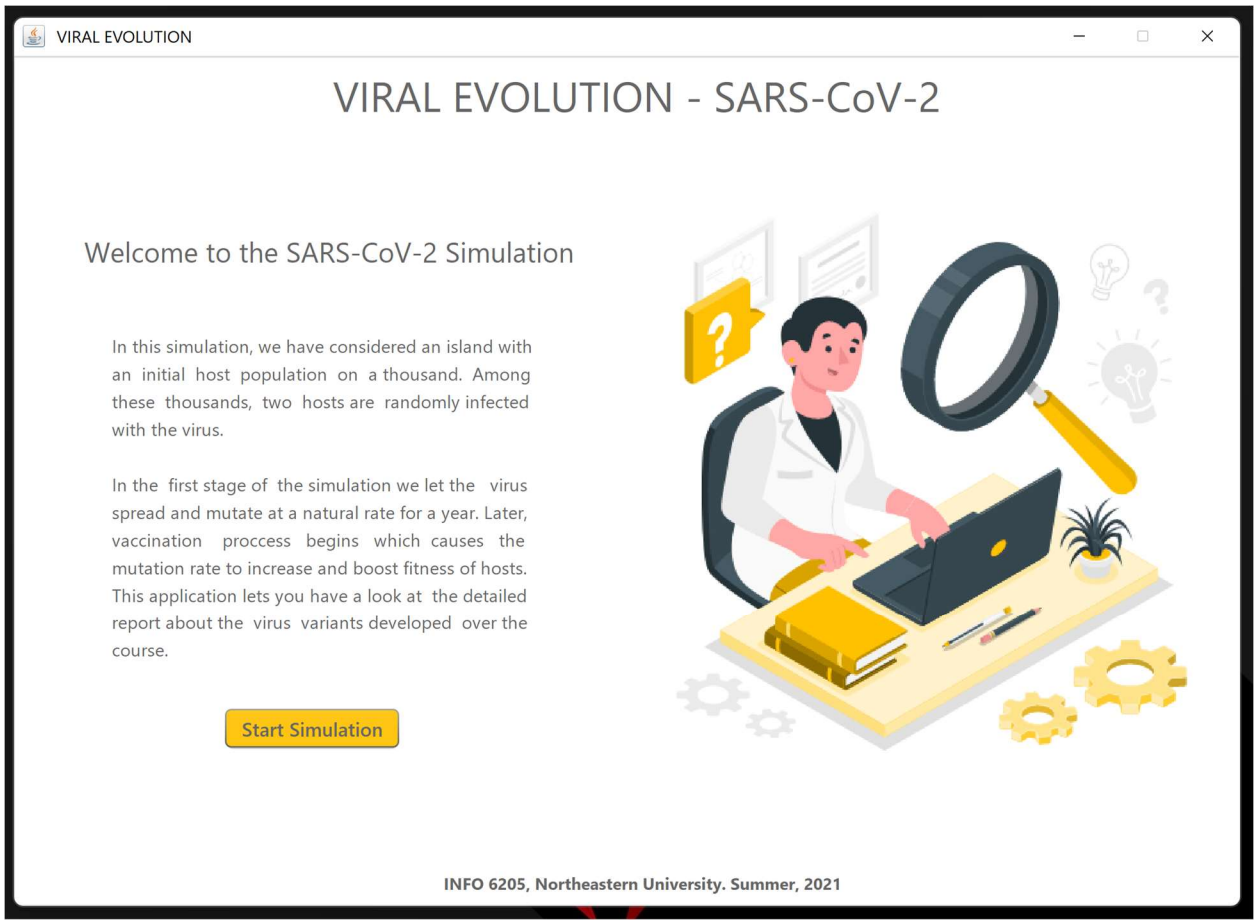
Landing Frame:



The loading page of our application runs for around 15sec to let the user to give an idea about genetic algorithm and gives brief introduction to our project.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

Introduction Frame:



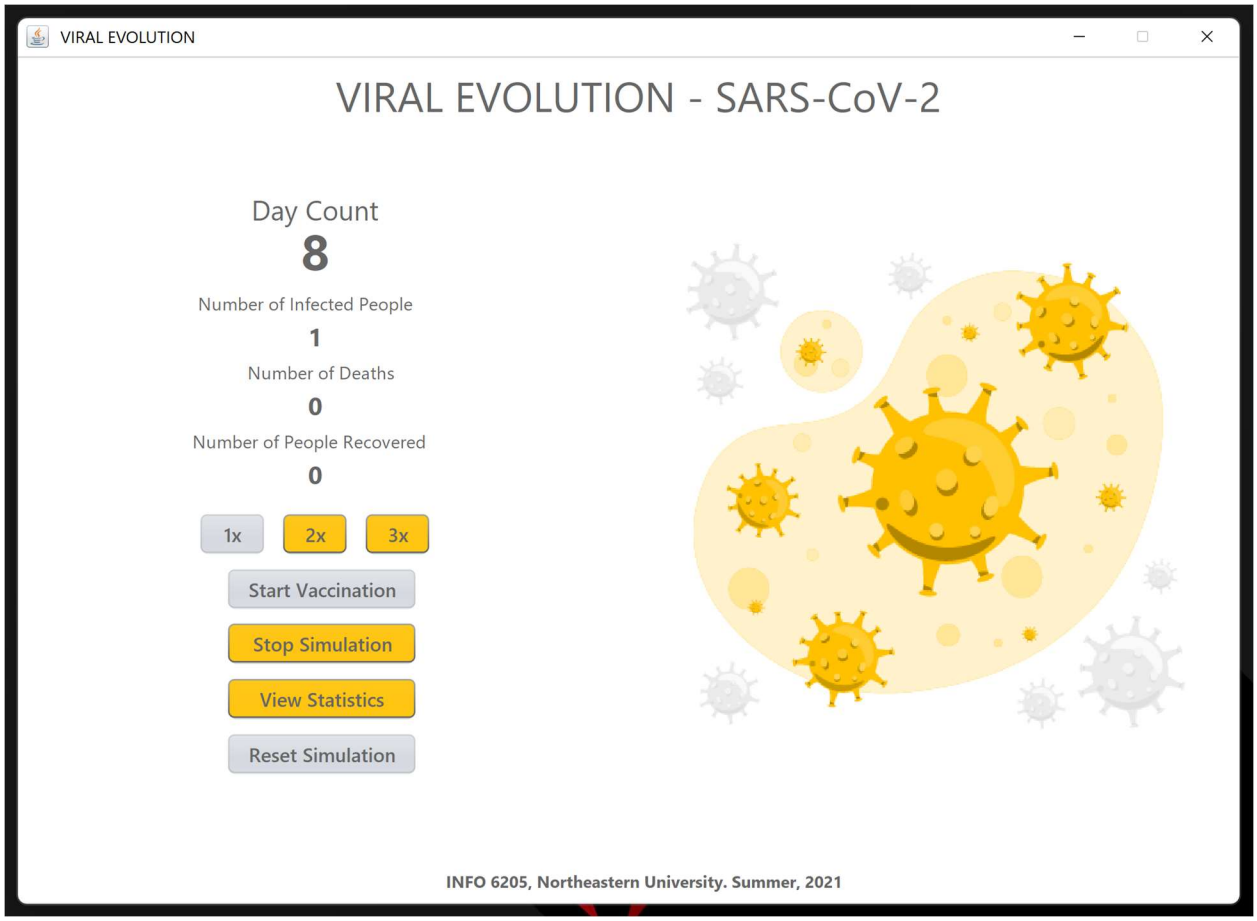
Gives an idea to the user about the scenario of our project

Buttons:

- **Start Simulation:** Starts simulation of the project.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

Simulation Frame:



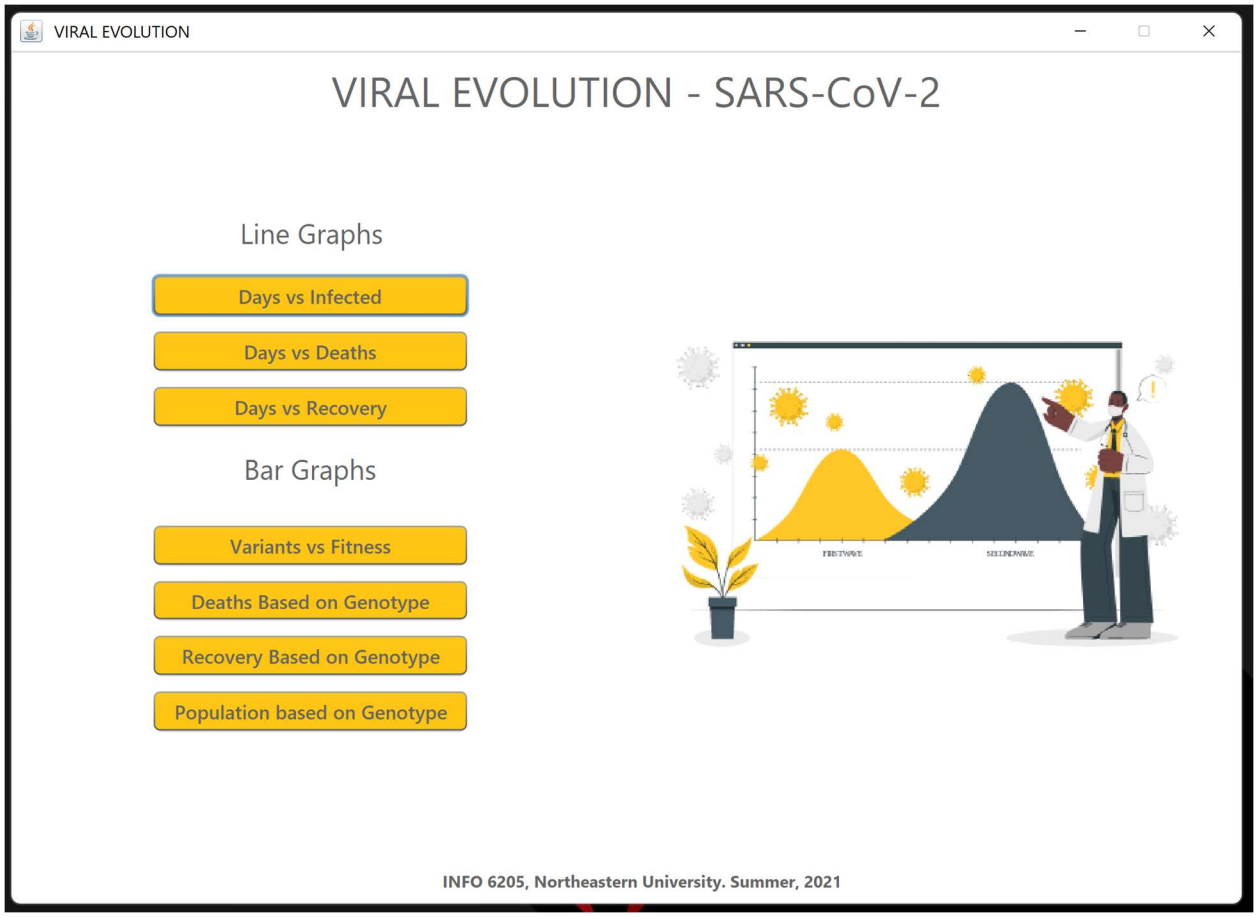
Consists of simulated information such as Day count, Number of people infected by the virus, Number of deaths due to virus, Number of people recovered from the virus

Buttons:

- **1x 2x & 3x:** To increase the speed of the simulation
- **Start Vaccination:** To start vaccination to people. This button is enabled after the day count reaches to 365 which is one year.
- **Stop Simulation:** To stop simulation
- **View Statistics:** View different statistical information incurred from the simulation
- **Reset Simulation:** Resets the simulation.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

Statistics Frame



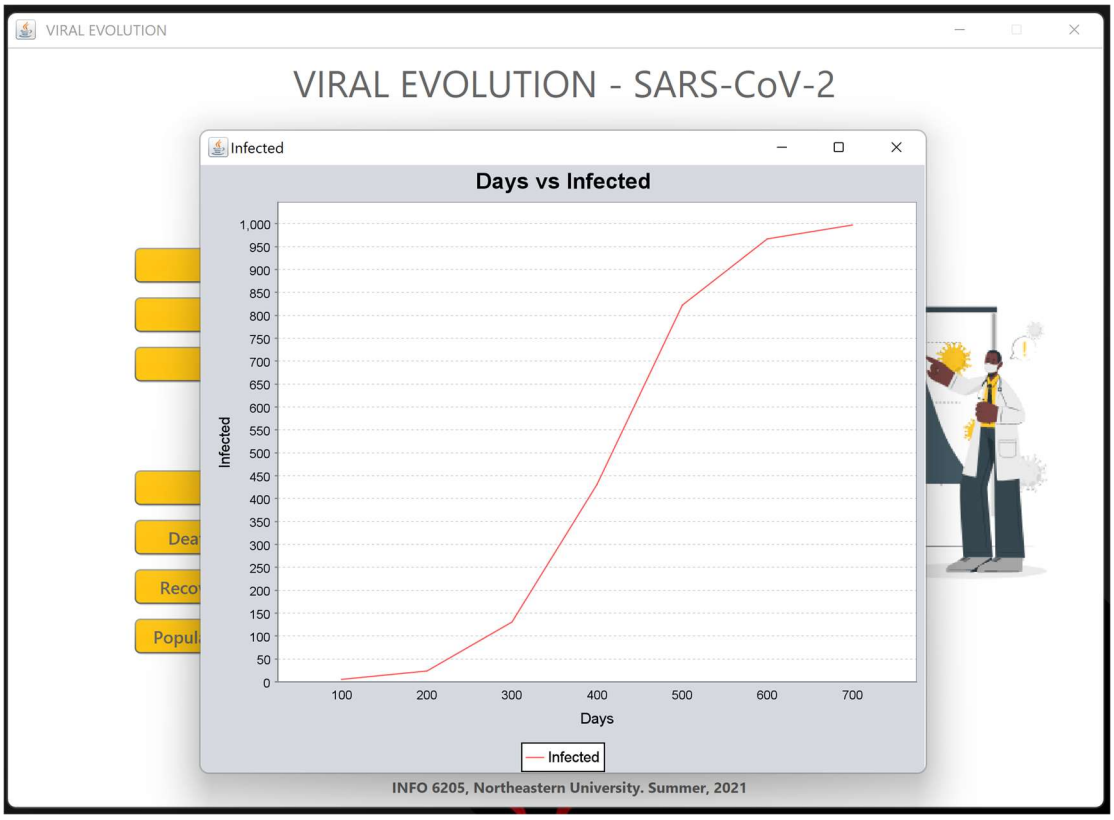
Consists of different statistical Information

Buttons:

- **Days vs infected:** Displays line graph for Number of days passed Vs Number of people Infected
- **Days vs Deaths:** Displays line graph for Number of days passed Vs Number of people dead
- **Days vs Recovery:** Displays line graph for Number of days passed Vs Number of people Recovered
- **Variants vs Fitness:** Displays bar graph for different Variants encountered Vs their respective Fitness
- **Deaths Based on Genotype:** Displays bar graph on Deaths of People based on different Genotypes of people.
- **Recovery Based on Genotype:** Displays bar graph on Recovery of People based on different Genotypes of people.
- **Population Based on Genotype:** Displays bar graph on distribution of People based on different Genotypes of people

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

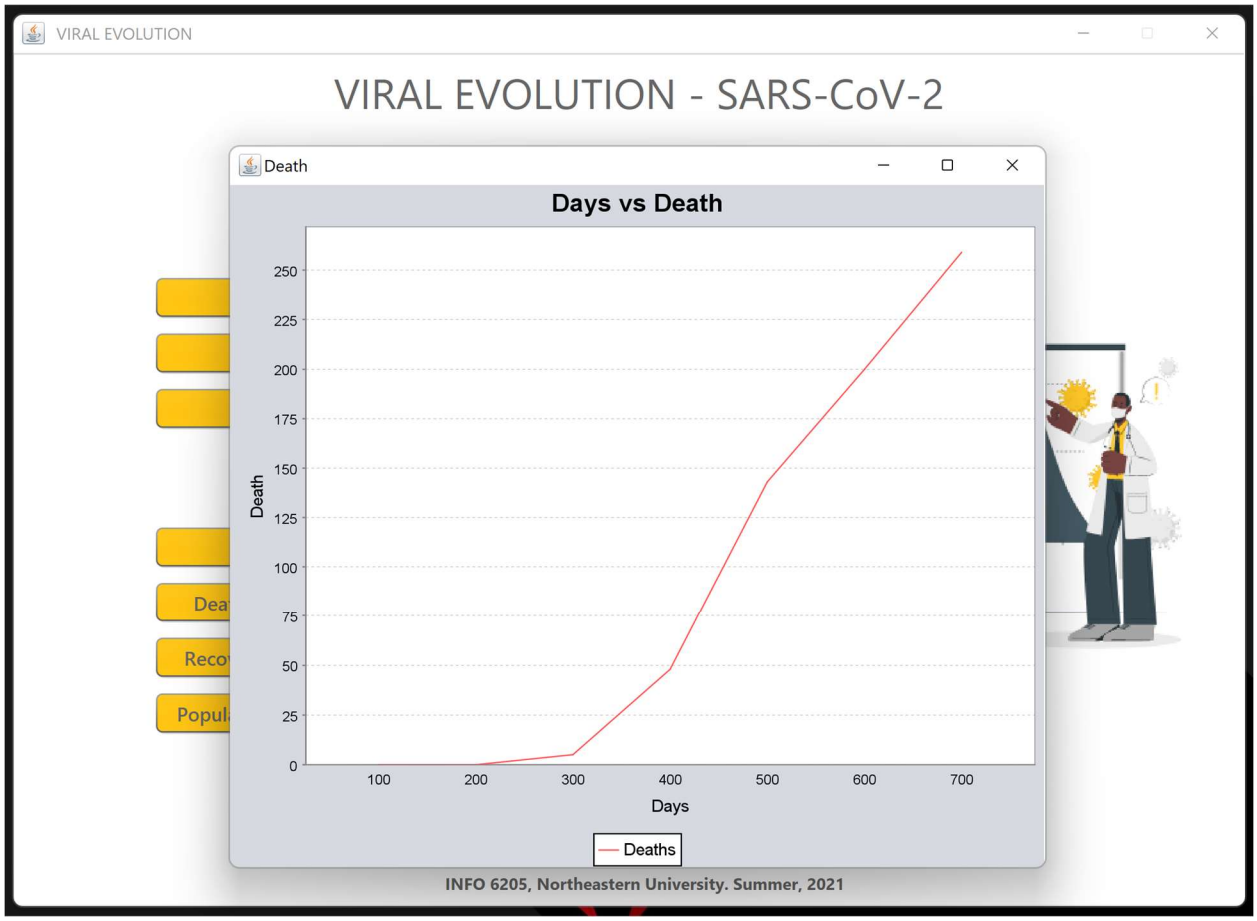
Observations and Graphical Analysis



Analysis:

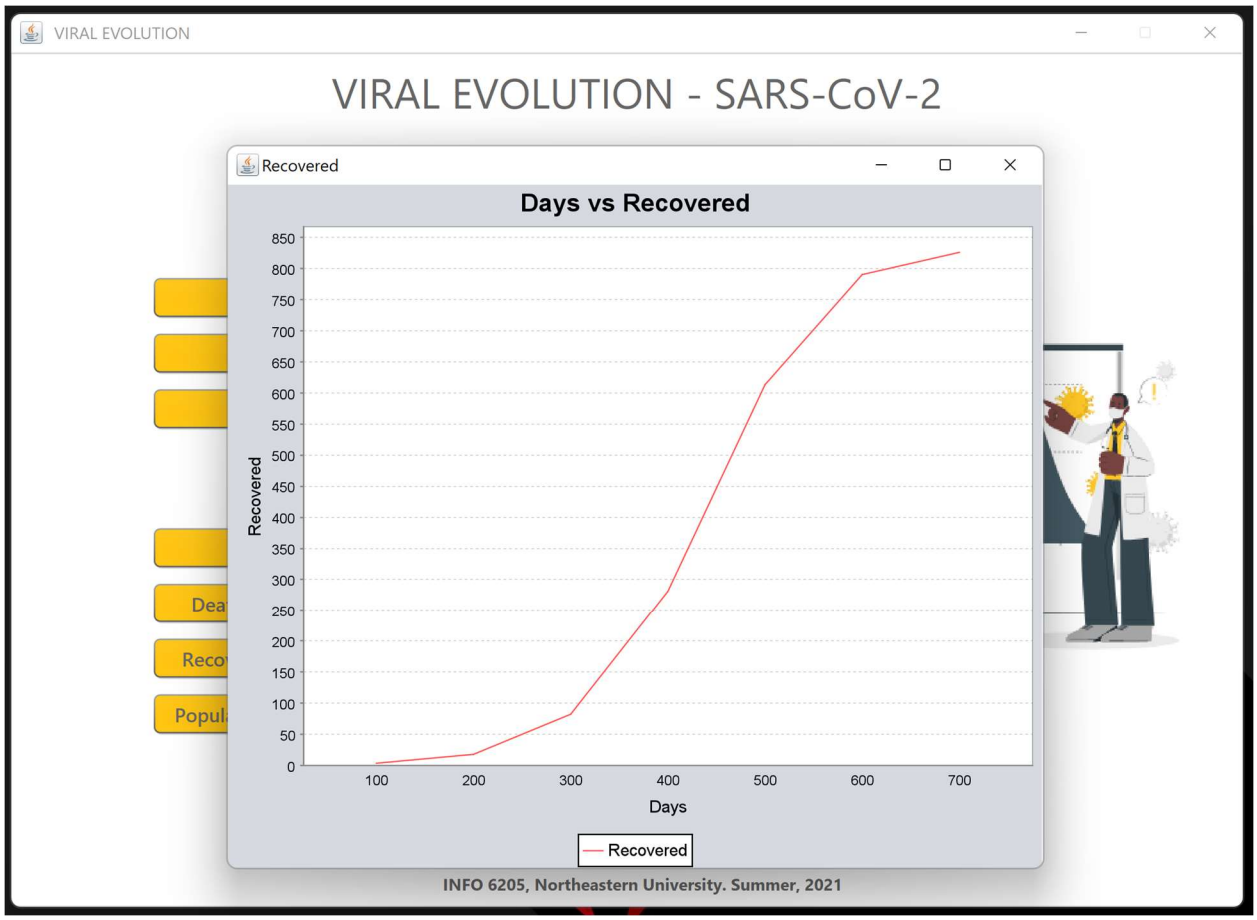
- Initially the count of infected people is less, as the number of days increases the count of people being infected also increased.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



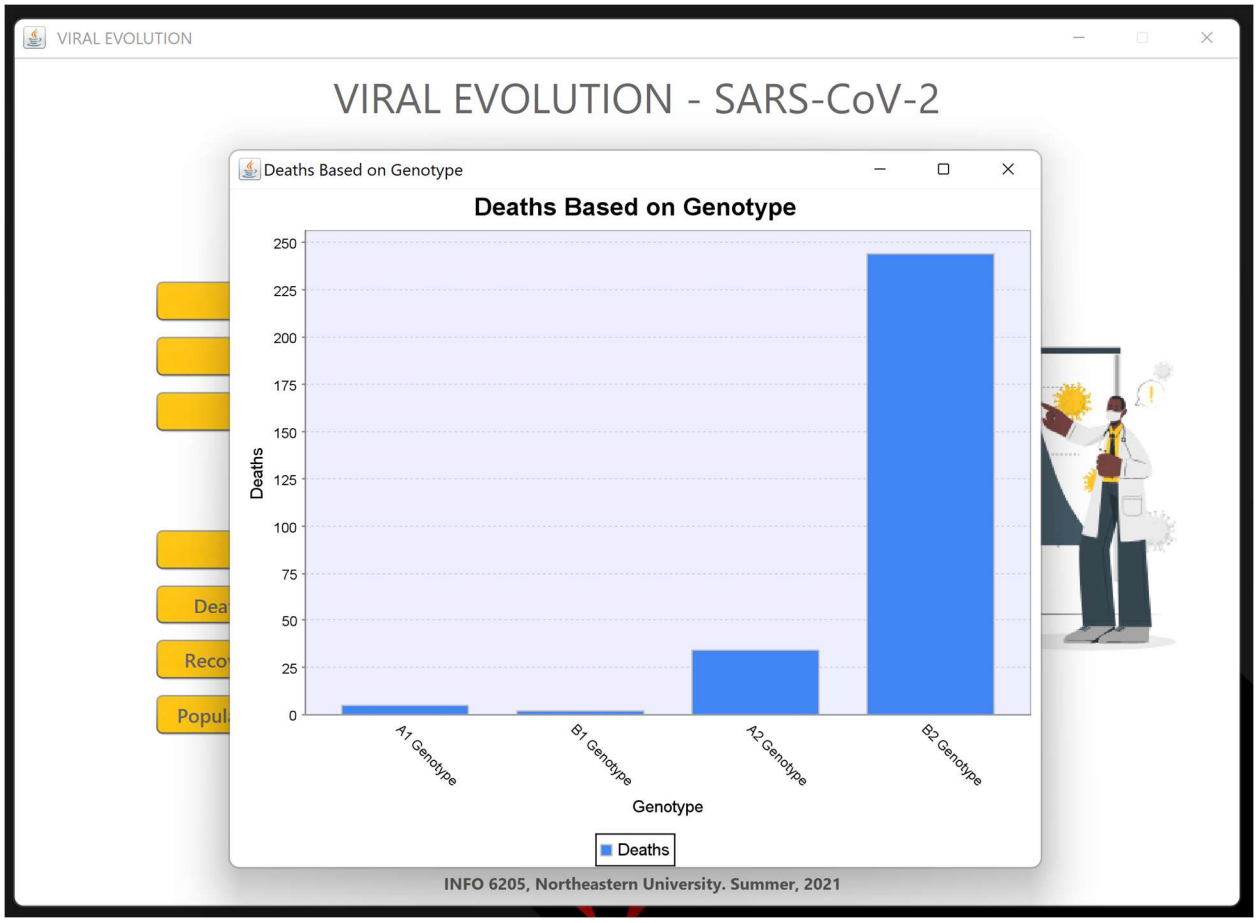
- As the number of infections in people increases based on the fitness of human, their Genome type and as the days pass with the increase in fitness of virus due to mutation and recombination deaths of virus increase.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



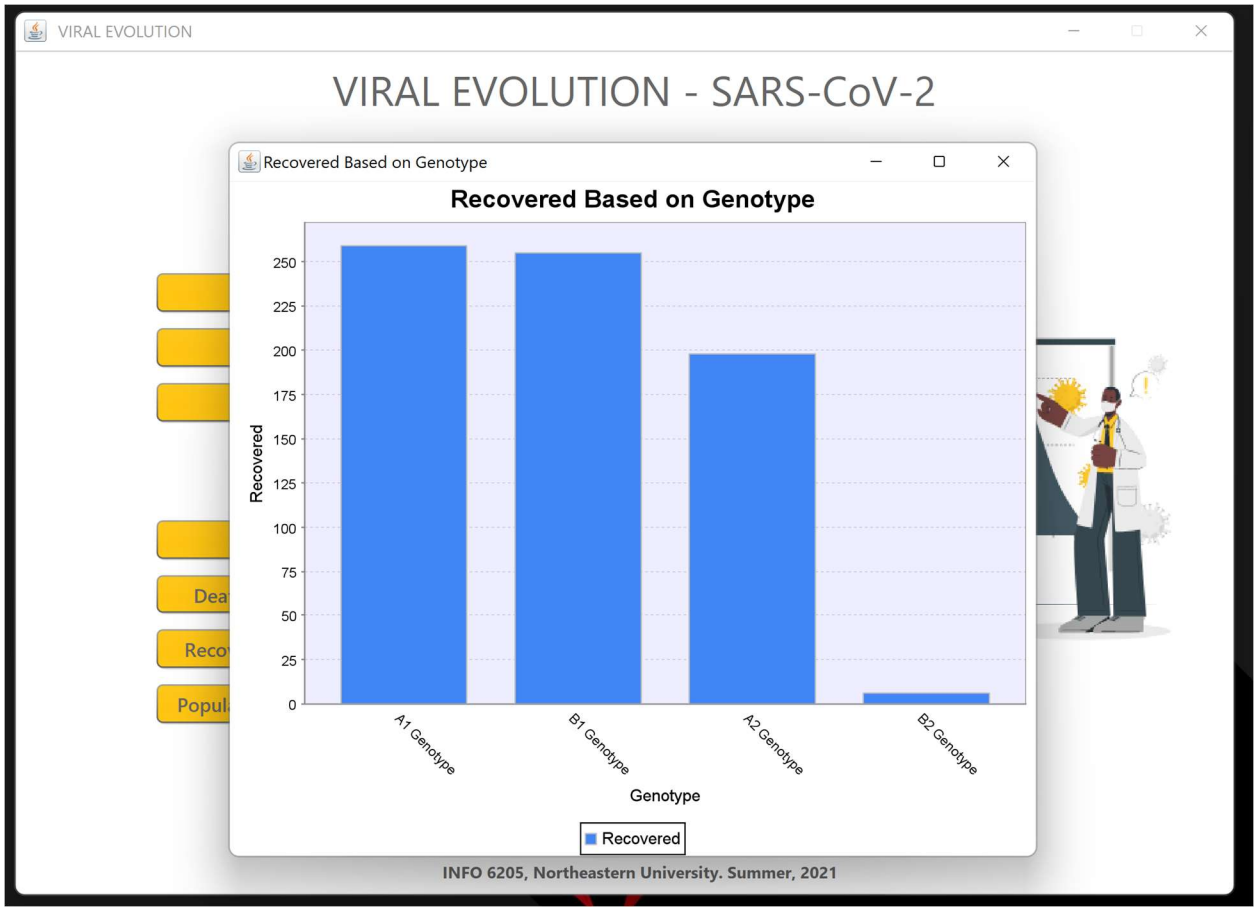
- As the number of days increases people start developing antibodies and after the vaccination the rate of recovery increases.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



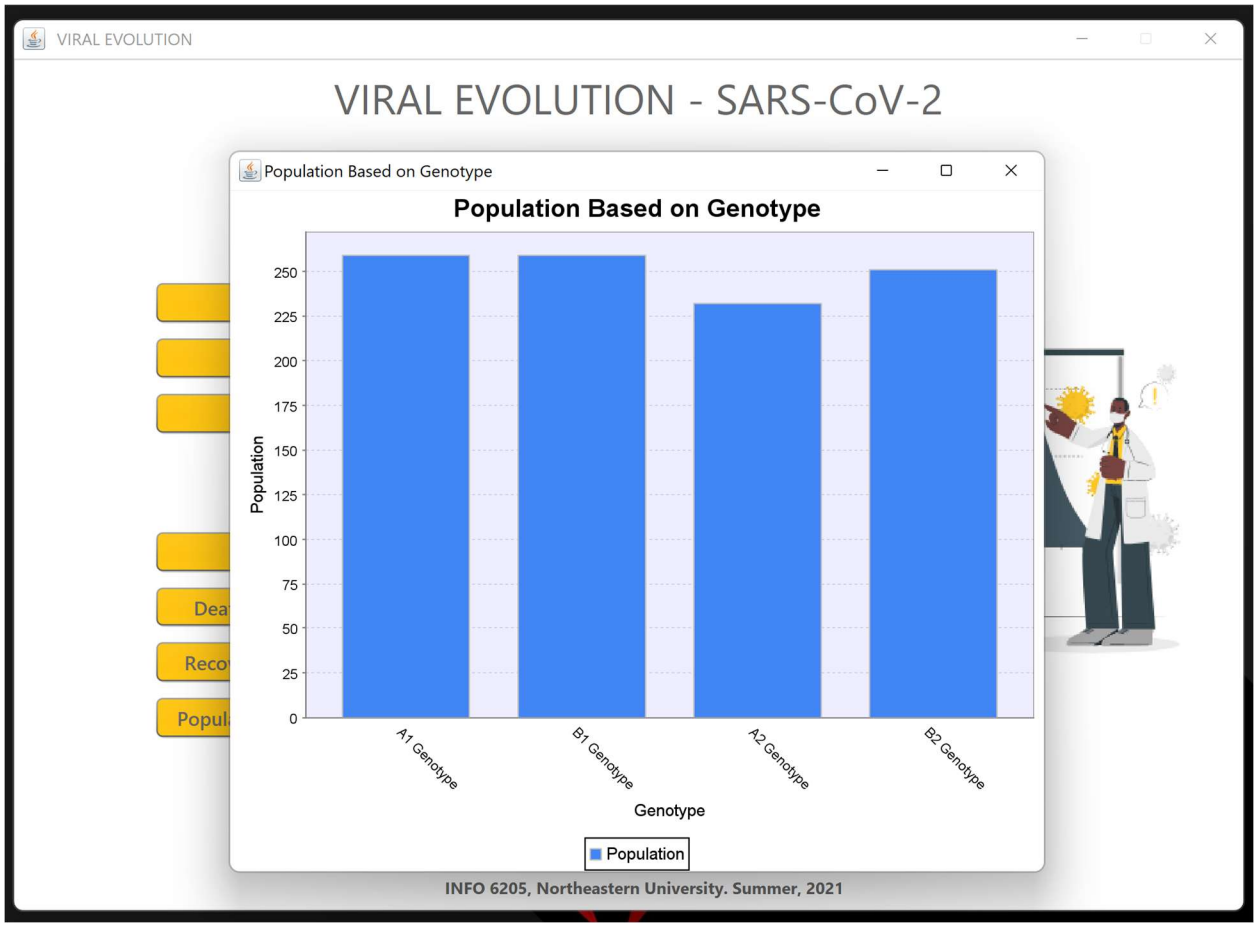
- Since B2 genotype has high probability of being effected by virus the deaths of B2 genotype is more compared to deaths of other genotype people.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



- Since A1 genotypes has low probability of being affected by virus the recovery of A1 is high compared to the recovery of other genotypes.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



- Division of different Genotypes of people based on random selection.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

RESULTS AND MATHEMATICAL ANALYSIS

Mathematical Analysis

Probability of Virus to mutate correctly && calculating the fitness of Virus:

SARS-cov-2 consists of 30000 base pairs made of 10 individual Genes (**ORF1a, ORF1b, S, ORF3a, E, M, ORF6, ORF7a, ORF8, ORF10**). The proportion in which these individual genes divide 30000 base pairs are as follows.

48%, 28%, 14%, 3%, 1%, 2%, 1%, 2%, 2%, 4%

So, to calculate the fitness of we have taken Virus RNA to be an array of 100 elements made from the above individual Genes. Based on the positioning of these individual genes we calculated the fitness, i.e., for Example ORF1a occupies 48% of the original SARS-cov-2 virus RNA. Therefore, when ORF1a gene is placed under 48 indices then the fitness count is incremented.

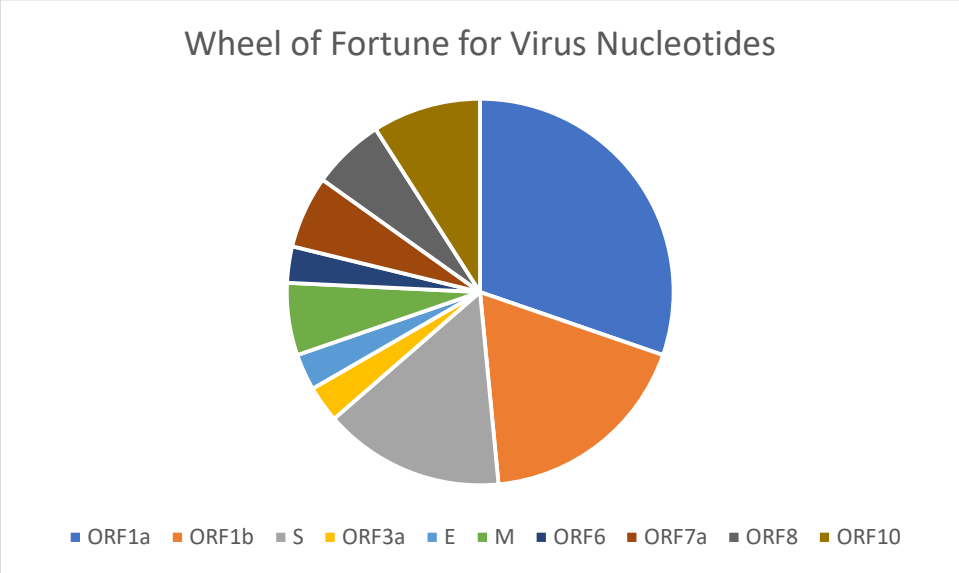
```
public void setFitness() {
    int count = 0;
    this.fitness = 0;
    for (String s : this.genome) {
        if (s.equals("ORF1a") && count < 48)
            this.fitness++;
        else if (s.equals("ORF1b") && (count >= 48 && count < 76))
            this.fitness++;
        else if (s.equals("S") && (count >= 76 && count < 90))
            this.fitness++;
        else if (s.equals("ORF3a") && (count >= 90 && count < 93))
            this.fitness++;
        else if (s.equals("E") && count == 93)
            this.fitness++;
        else if (s.equals("M") && (count >= 94 && count < 96))
            this.fitness++;
        else if (s.equals("ORF6") && count == 96)
            this.fitness++;
        else if (s.equals("ORF7a") && (count >= 97 && count < 99))
            this.fitness++;
        else if (s.equals("ORF8") && (count >= 99 && count < 101))
            this.fitness++;
        else if (s.equals("ORF10") && (count >= 101 && count < 105))
            this.fitness++;
        count++;

    }
}
```

On the other hand, our pool or Data structure from which genes are selected for making or mutating virus consist of following Genes.

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

```
String[] nucleotides = {
        "ORF1a","ORF1a","ORF1a","ORF1a","ORF1a","ORF1a","ORF1a","ORF1a","ORF1a","ORF1a",
        "ORF1b","ORF1b","ORF1b","ORF1b","ORF1b","ORF1b",
        "S","S","S","S","S",
        "ORF3a",
        "E",
        "M","M",
        "ORF6",
        "ORF7a","ORF7a",
        "ORF8","ORF8",
        "ORF10","ORF10","ORF10"
};
```



The probability of selecting ORF1a for its correct position is:

(10/33) *(48/105) = 0.13. Which implies that out of 10 successful mutations only one of them has chance of correct mutation which increases the fitness

The amount of human fitness reduced when fighting with a virus.

From this resource we have observed that the chances of different age groups being hospitalized (virus cells attacking the host cells) is as follows:

	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Cases[2]	<1x	1x	Reference group	1x	1x	1x	1x	1x	1x
Hospitalization[3]	<1x	<1x	Reference group	2x	2x	4x	6x	9x	15x

So, on average we have considered the above values for our different genotypes as follows:

A1:2x, A2: 1x, B1:5x, B2:12x

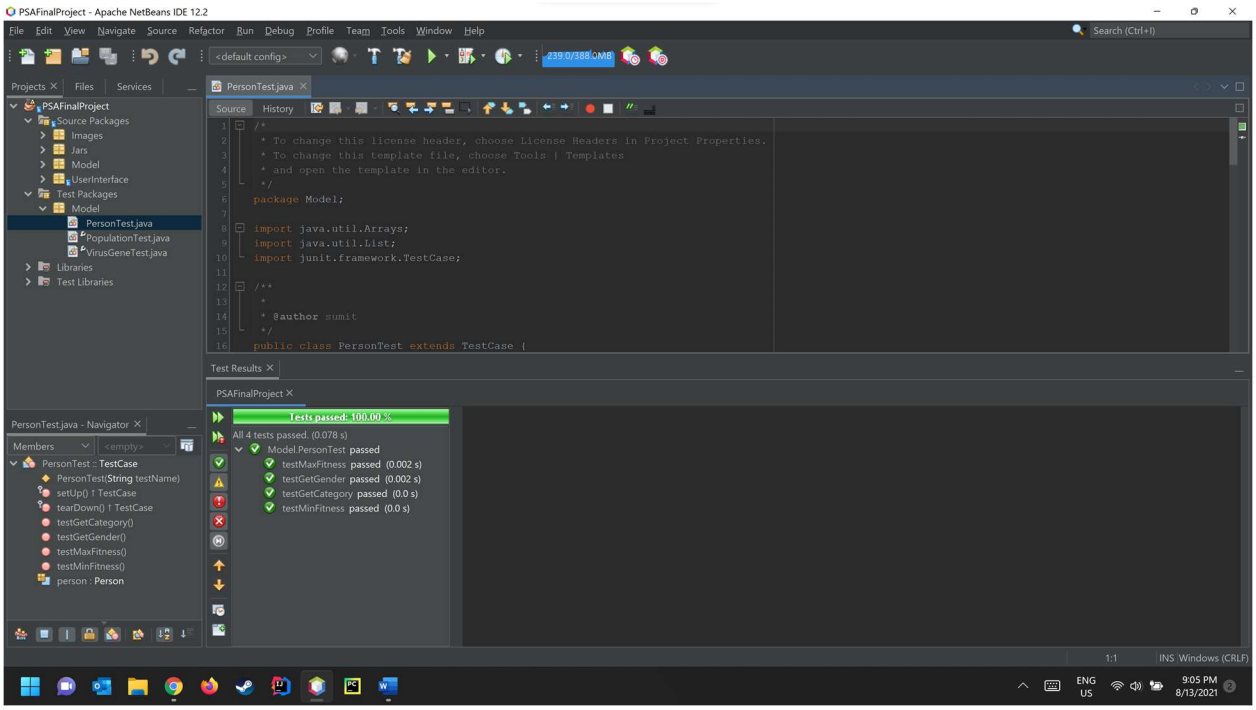
Chances or probability of A1 being hospitalized: 2/20
Chances or probability of A2 being hospitalized: 1/20
Chances or probability of B1 being hospitalized: 5/20
Chances or probability of B2 being hospitalized: 12/20

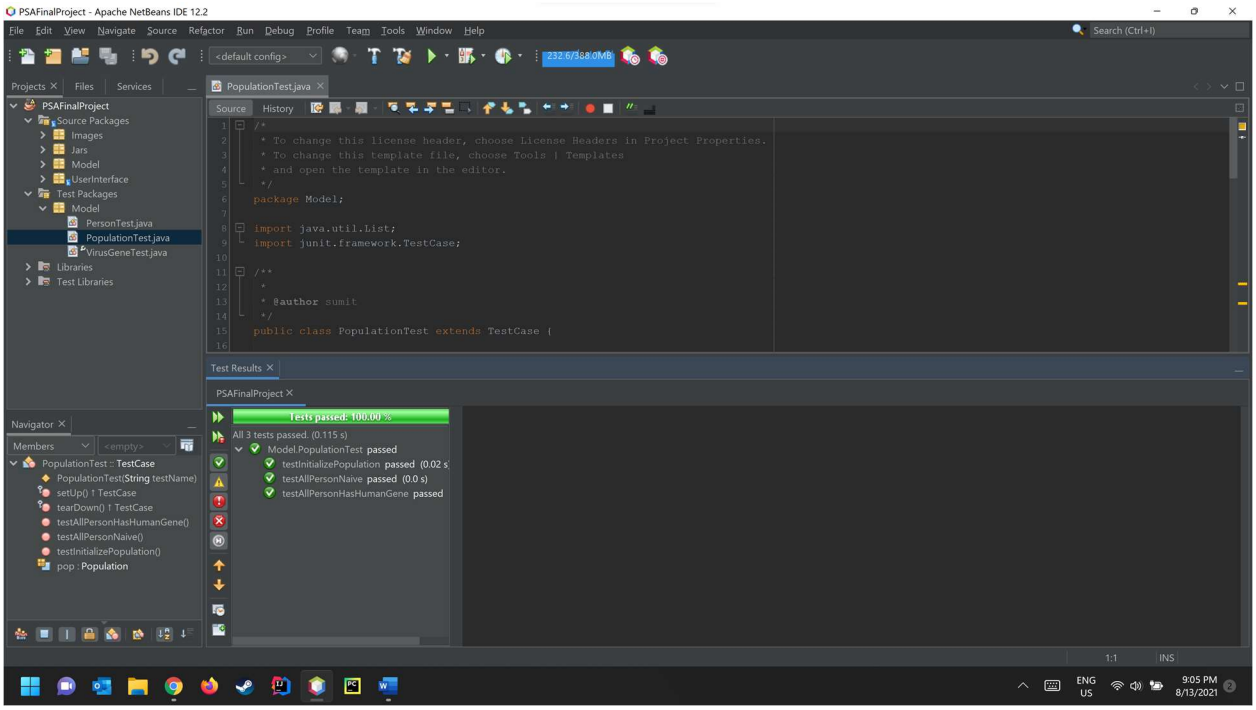
Therefore, when a human of different Genotype is fighting, its respective probability of being hospitalized is multiplied to the fitness of virus.

```
double virusNewFitness = 0;
if (this.category.equals("A1")) {
    virusNewFitness = this.virusGene.getFitness() * RuntimeAttributes.fitnessTable.get(this.virusGene.getName()).get(0);
} else if (this.category.equals("B1")) {
    virusNewFitness = this.virusGene.getFitness() * RuntimeAttributes.fitnessTable.get(this.virusGene.getName()).get(1);
} else if (this.category.equals("A2")) {
    virusNewFitness = this.virusGene.getFitness() * RuntimeAttributes.fitnessTable.get(this.virusGene.getName()).get(2);
} else if (this.category.equals("B2")) {
    virusNewFitness = this.virusGene.getFitness() * RuntimeAttributes.fitnessTable.get(this.virusGene.getName()).get(3);
}
double newFitness = this.getFitness() - virusNewFitness;
this.setFitness(newFitness);
if (this.getFitness() < RuntimeAttributes.naiveHumanThreshold) {
    setAlive(false);
    RuntimeAttributes.deathCount++;
    RuntimeAttributes.naiveDeathCount++;
}
```

Northeastern University- INFO 6205- Final Project
Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796

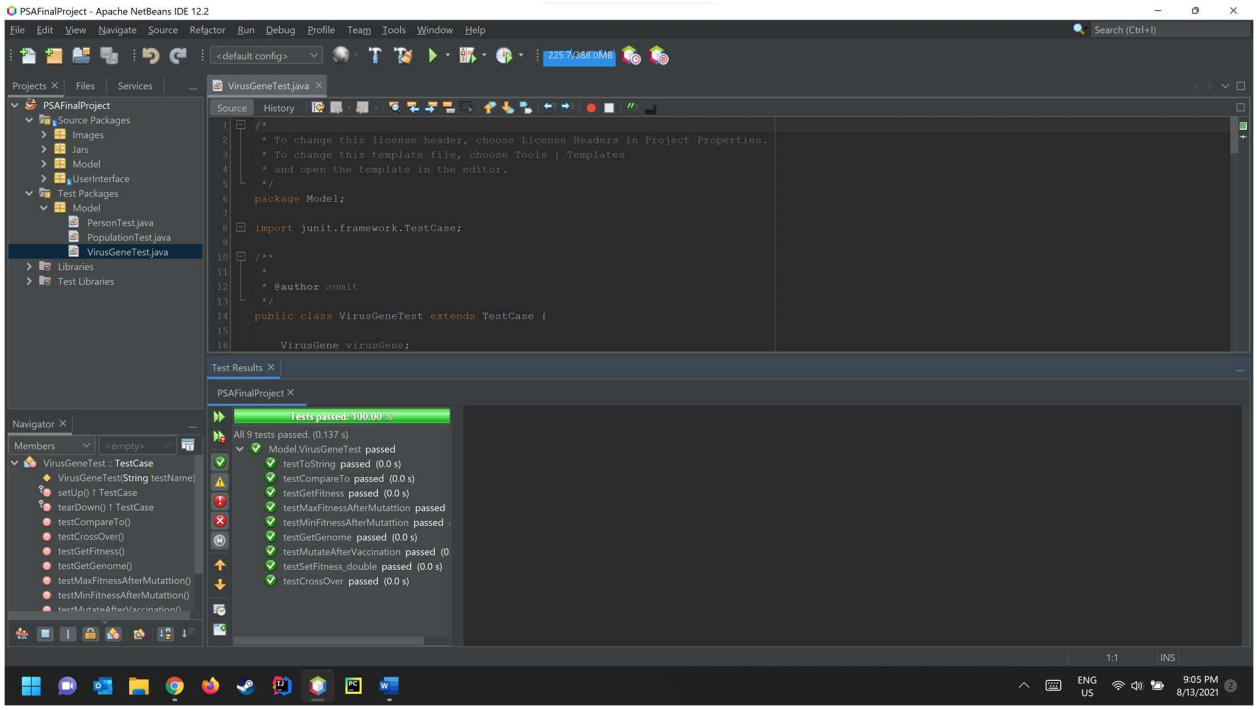
TESTCASES





Northeastern University- INFO 6205- Final Project

Sumit Subbanna Madgi- 001581012
Aditya Ganesh Satish- 001042389
Chinmay Ganesh Chavan- 001568796



CONCLUSION

By the end of the simulation these are the following conclusions:

- The probability of virus mutating successfully is very low.
- As in our simulation since B2 genotypes have high probability of being affected by virus the number of deaths in B2 genotypes are more.
- Similarly, since A1 genotypes have low probability of being affected by virus the number of deaths in A1 genotypes is less.
- As vaccination in people increases the fitness the number of people who recovered from virus increases.
- The mutation rate in virus for naive people is less than the mutation rate in vaccinated people.

REFRENCES

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2

https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome_coronavirus_2#/media/File:SARS-CoV-2_genome.svg

https://towardsdatascience.com/how-to-define-a-fitness-function-in-a-genetic-algorithm-be572b9ea3b4