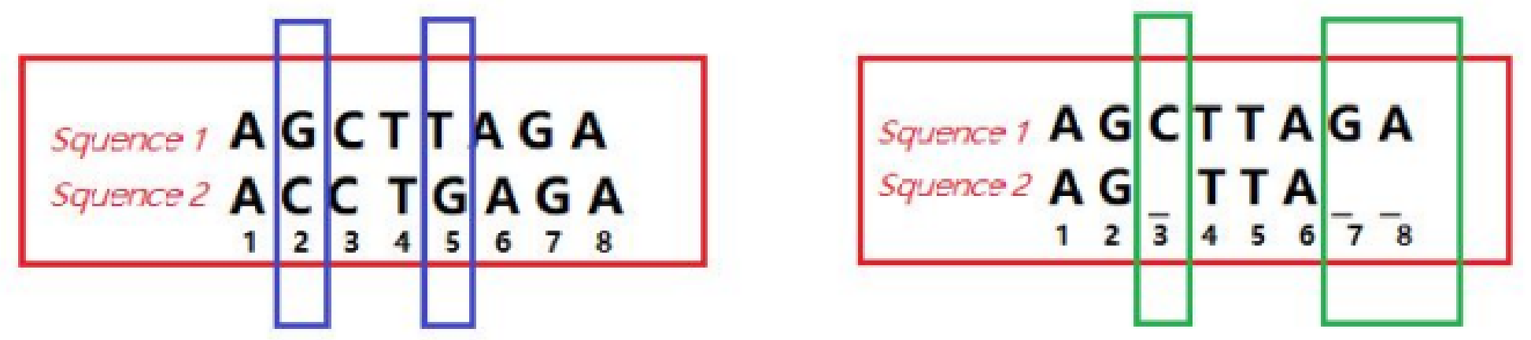
Statistical modeling of SARS-Cov-2 mutation in the U.S.

Yuru Jing, Angela Antonou, Prof.

Dept. of Mathematics and Computer Science, University of St. Francis

Introduction

As a result of the severity of the SARS-Cov-2 outbreak worldwide, researchers have invented vaccines to prevent its spread. However, the higher mutation characteristics of an RNA virus such as SARS-Cov-2 may make it more likely that a group of individuals with a higher average and standard deviation mutation frequency will have a shorter duration of immunity against SARS-Cov-2 after vaccination. To investigate this, we analyzed a random sample of thousands of SARS-Cov-2 RNA sequences from infected individuals within the US and calculated their mutation frequency by aligning each of them against the first outbreak RNA sequence in the database. During the aligned process, there are two primary cases to count the mutants in our model, which are mismatch and gap. They illustrate our general three types of mutation as the Figure 1. The implement of the fast global sequence alignment algorithm with the core math thought, branch-and-bound technique will help us find the final mutation frequency for each case and get the further statistical modeling and analysis based on those mutation frequency data.


Squence 1 A G C T T A G A
Squence 2 A C C T G A G A
1 2 3 4 5 6 7 8
Squence 1 A G C T T A G A
Squence 2 A G _ T T A _ _
1 2 3 4 5 6 7 8

(a) Mismatch in the alignment (b) Gap in the alignment

Figure 1. The aligned process for two RNA sequences: 2 mismatches in (a), 3 gaps in (b)

Fast Optimal Global Sequence Alignment Algorithm (FOGSAA)

In 2013, Chakraborty and Bandyopadhyay proposed the fast optimal global sequence alignment algorithm, which generate the optimal alignment between two sequences by finding the optimal branch in the following branch and bound tree. In our alignment, we assume that the reference sequence cannot contain a gap. Furthermore, we use the scoring scheme described in the Introduction and again in Equation 1.

- Alignment Assumption: FOGSAA assumes that there are two nucleotide sequences S_1 and S_2 with different lengths m and n, respectively, and we assume S_1 to be the reference sequence here. Thus, we can label the sequences:

$$S_1 : (a_1 a_2 ... a_m)$$
$$S_2 : (b_1 b_2 b_3 ... b_n),$$

  where a_i and b_j represent the bases in the ith and jth positions of their respective sequences.Let P_1 and P_2 be pointers to the bases in S_1 and S_2, with initial pointer values set to 0 for both P_1 and P_2.
- Scoring Scheme: In aligning process, three possible outcome are as below: a gap in S_2 (the sample sequence), a match, or a mismatch. To indicate these options, we use the scoring scheme in Equation 1.

$$\begin{cases} -2 & \text{if gap} \\ 1 & \text{if match} \\ -1 & \text{if mismatch.} \end{cases} \tag{1}$$

- Branching Criteria: Each node in a FOGSAA tree also stores two additional components: the Present Score(PrS) and the Fitness Score(F), which are used to determine the criteria for branching as well as the bound needed to terminate the algorithm.The Present Score (PrS) represents the sum of scores from the root node to the present node.

$$F_{min} = \begin{cases} x_2 \cdot (-1) + (-2) \cdot (x_1 - x_2), & x_2 < x_1 \\ x_1 \cdot (-1) + (-2) \cdot (x_2 - x_1), & \text{otherwise} \end{cases}$$

$$F_{max} = \begin{cases} x_2 \cdot (1) + (-2) \cdot (x_1 - x_2), & x_2 < x_1 \\ x_1 \cdot (1) + (-2) \cdot (x_2 - x_1), & \text{otherwise.} \end{cases}$$

Supposing that $(P_1, P_2) = (i_k, j_k)$. Then the present score for this node will be:

$$PrS = \sum_{\forall i_p j_p, 1 \le p \le k} SC_{i_p j_p}, \tag{2}$$

where

$$SC_{i_p j_p} = \begin{cases} -2, & \text{if } b_{j_p} = gap \\ 1, & \text{if } a_{i_p} = b_{j_p} \\ -1, & \text{if } a_{i_p} \neq b_{j_p} \text{ and } b_{j_p} \neq gap. \end{cases}$$

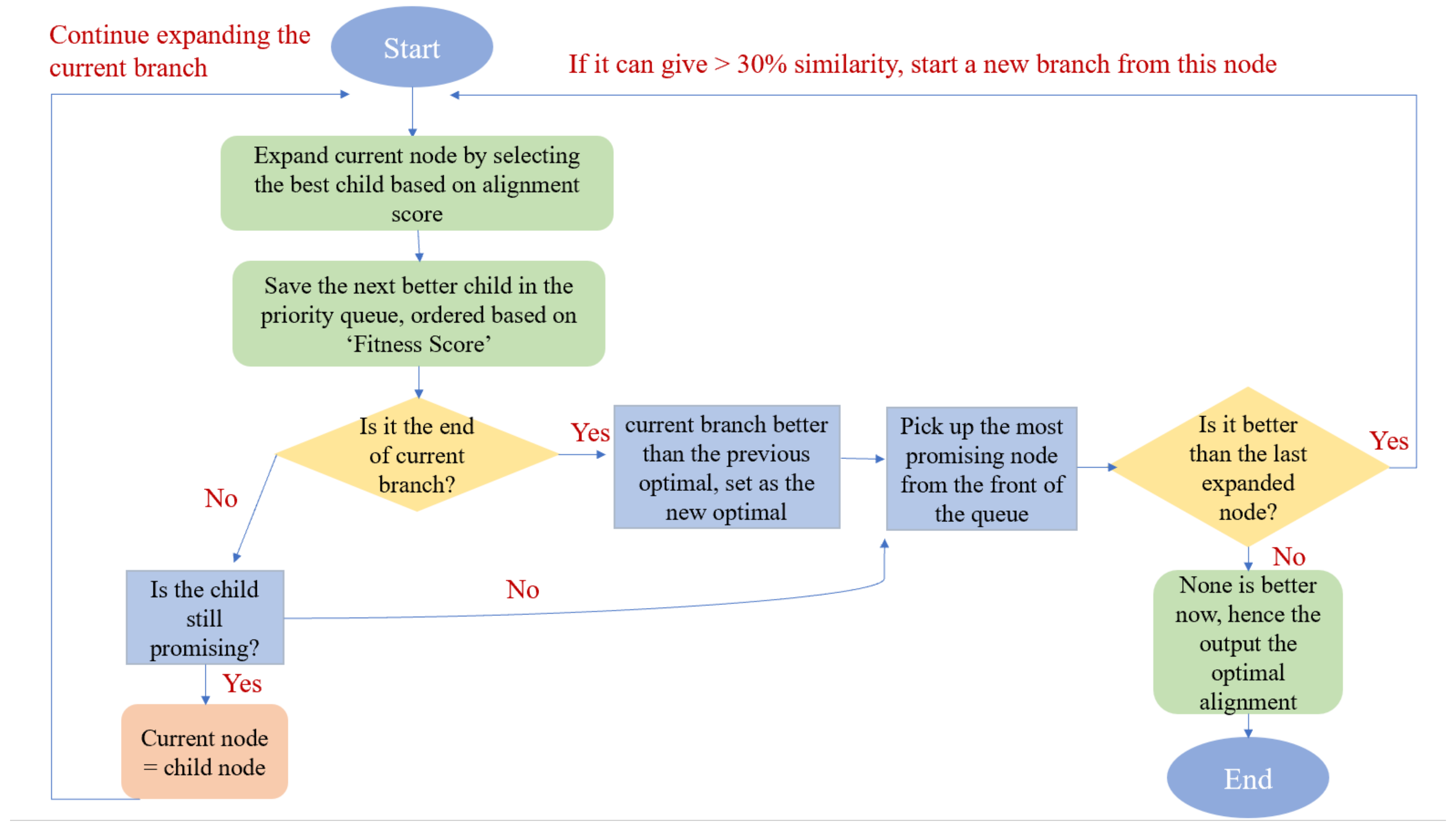
The Fitness Score is a combination of the Present Score and another measure, the Future Score. The Future Score contains two values F_{min} and F_{max} and indicates the highest and lowest possible scores in aligning the remaining portions of the sequences. Taking x_1 to be $m - P_1$ and x_2 to be $n - P_2$, the Future Score can be determined by the following formulas (recalling the scoring scheme in Equation 1):

$$F_{min} = \begin{cases} x_2 \cdot (-1) + (-2) \cdot (x_1 - x_2), & x_2 < x_1 \\ x_1 \cdot (-1) + (-2) \cdot (x_2 - x_1), & \text{otherwise} \end{cases}$$

$$F_{max} = \begin{cases} x_2 \cdot (1) + (-2) \cdot (x_1 - x_2), & x_2 < x_1 \\ x_1 \cdot (1) + (-2) \cdot (x_2 - x_1), & \text{otherwise.} \end{cases}$$

T is the fitness score, which can help us select the best child to determine the optimal sequence.

$$T_{min} = PrS + F_{min}$$
$$T_{max} = PrS + F_{max}.$$


Continue expanding the current branch

Start

If it can give > 30% similarity, start a new branch from this node

Expand current node by selecting the best child based on alignment score

Save the next better child in the priority queue, ordered based on 'Fitness Score'

Is it the end of current branch?

Yes

current branch better than the previous optimal, set as the new optimal

Pick up the most promising node from the front of the queue

Is it better than the last expanded node?

Yes

No

None is better now, hence the output the optimal alignment

End

No

Is the child still promising?

No

Yes

Current node = child node

Figure 2. The workflow of the algorithm

Statistical Model Analysis and Results

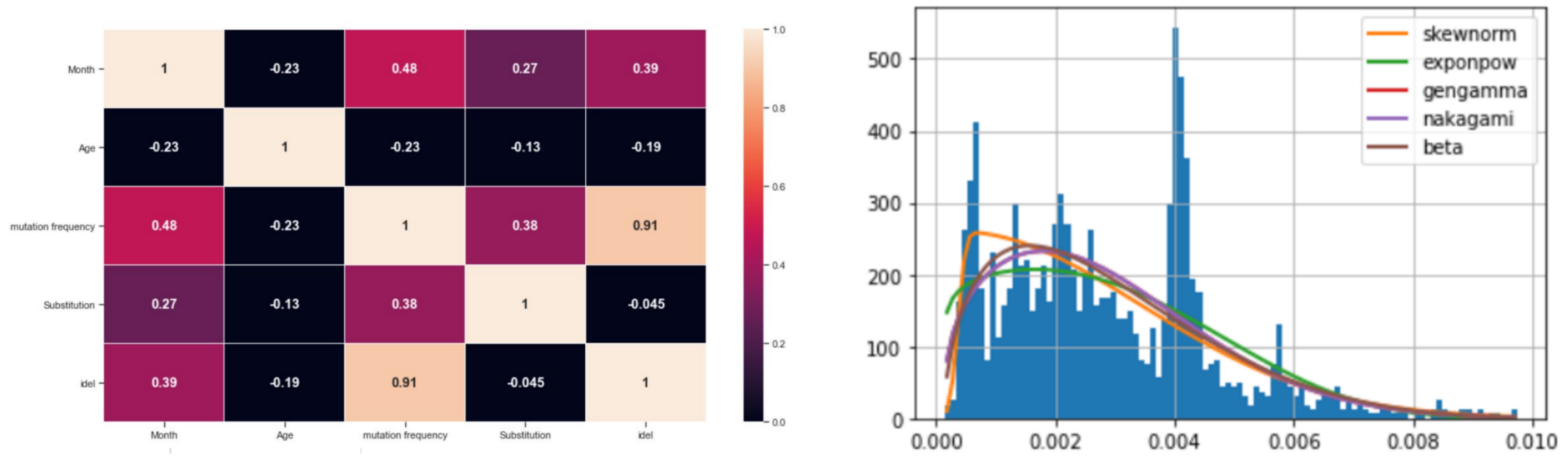


Figure 3. The left graph is the correlation for the entire features without outliers; The right graph is the mutation distribution with the fitting test(The skew-normal distribution with the smallest AIC -791.59 and BIC 9802.77)

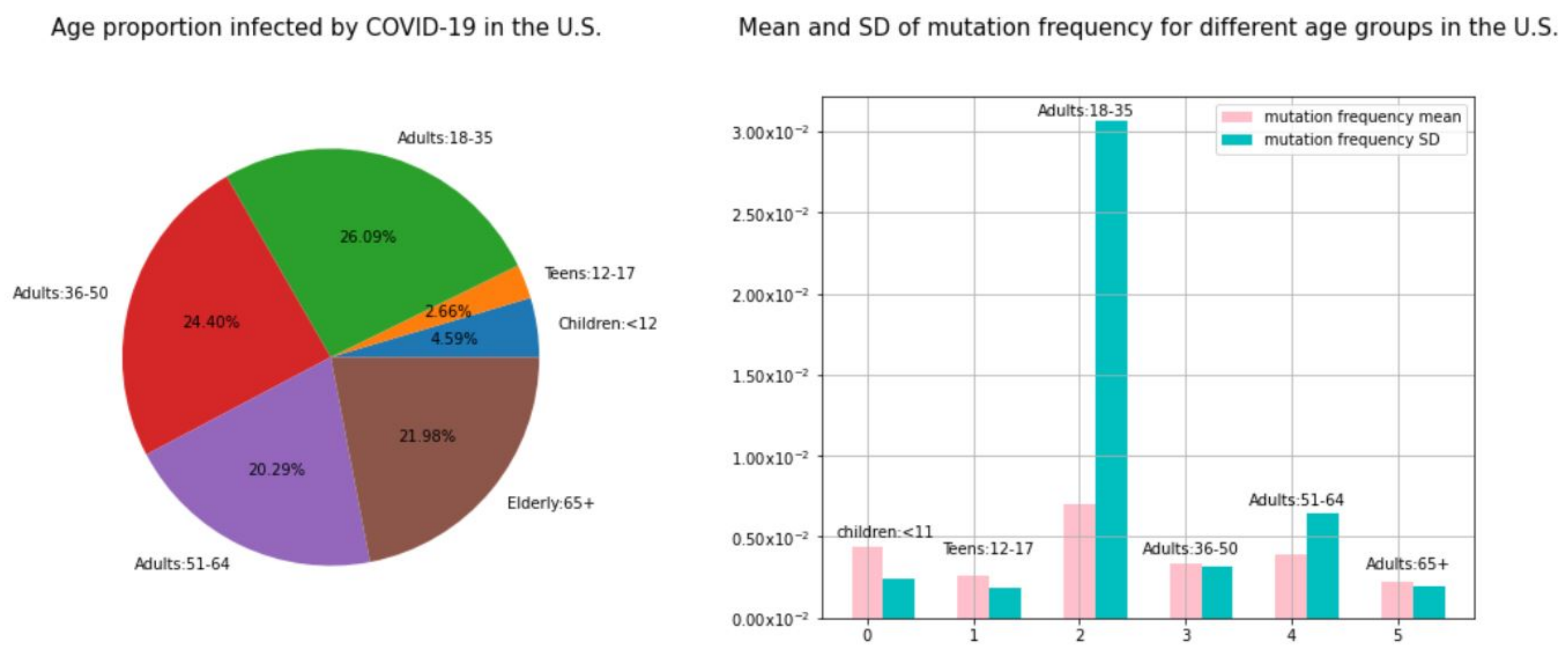


Figure 4. Proportion of age groups represented in the sample and a bar chart with the mean and SD mutation frequency within these age groups

Table 1. The cumulative density function table of the skew normal distribution

Probability portion(cdf)	0.2	0.4	0.6	0.8
mutation frequency	$1.15 * 10^{-3}$	$1.97 * 10^{-3}$	$2.95 * 10^{-3}$	$4.3 * 10^{-3}$

Using the Stepwise Selection to find the best mutation frequency forecast model, where x_1 is day, x_2 is age:

$$-9.188 * 10^{-7} x_2^3 - 1.412 * 10^{-6} x_1^3 + 2.67 * 10^{-4} x_2^2 + 0.67 \tag{3}$$

Conclusion

- There is linear positive relationship between all features and mutation frequency except the age feature.
- The shorter immunization group after vaccination is:(1) younger age group, especially for the age range from 18 to 35. (2) The Midwest and western regions of the U.S. (3) The lineage B and the clade with letter "G" type.
- The mutation distribution is skew-normal and the forecast mutation model is cubic.

Presentation for the Symposium on 2021 BEER

Contact Info: Aantonou@stfrancis.edu

YuruJing@stfrancis.edu