Testing the Validity of the Use of Genealogical Trees to Predict Virus Evolution

Team Members:
Dominique Hughes
Ling Yang
Nur Izzati Johari

Project Description

- **Purpose**: Our project is to validate an existing study from Neher et al. about using genealogical trees to predict virus evolution [1].
- **Motivation**: Neher et al. uses their own secondary model to validate their primary model; we wanted to test their primary model on real influenza data.
- **Literature**: Neher et al. provides the mathematical model we are validating, termed the Inferred Fitness Model (IFM, Equation (1)) [1].
  - The IFM determines the fitness of a virus strain.
  - The IFM uses the shape of the genealogical tree of a virus to determine it's fitness, without understanding the genetic components [1,2].

$$P(x|T) = \frac{p_0(x_0)}{Z(T)} \prod_{i=0}^{n_{int}} g(x_{i1}, t_{i1} | x_i, t_i)\, g(x_{i2}, t_{i2} | x_i, t_i) \quad (1)$$

Equation (1): The IFM equation. The x-variables give the fitness of each node. Higher fitness is related to the branching patterns in the genealogical tree, which is related to how many offspring (or in this case, future strains) a given strain should be able to produce.

- **The need for a model**: The flu vaccine is a prediction of the virus, but because of the random virus mutate, prediction of the strains can be difficult and the vaccine may not be an accurate predictor of the most prominent flu viruses [3].
  - Genealogical trees can provide us with substantial knowledge about the evolution of a virus without requiring an in depth knowledge of its DNA [2].
- **Project Goal**: We want to determine the efficacy of the IFM by testing it on several decades' influenza data and recording its accuracy [1].

Scientific Challenges

- While this method does not require an understanding of how genetic mutations affect fitness, genetic data for the virus is still required to create a genealogical tree.

Project Challenges

- Neher et al. states in their paper that they slightly alter a portion of the code to account for small branching [1], but do not specify how it was modified so we could not recreate it.
- We had to create our own method for validating the IFM's prediction of the most fit virus strain.

Potential Applications

- We are currently facing a outbreak of a new virus, COVID-19, which is an evolved form of the SARS virus [4].
  - This method of virus evolution prediction could assist in vaccine making for these types of situations, or any situation where we do not have an in depth understanding of the genetic code.
- Also, this model can help the improvement of the vaccine every year.

Methodology

1. We used the program provided by Neher et al. to rank the viruses in terms of fitness [1].
   a. We ranked the HA genetic portion of influenza AH3N2 virus from the years 1968 to 2014 of 40 states.
   b. The highest ranked virus is considered the progenitor virus for the next year.
2. We created a phylogenetic tree [5] to determine if the top ranked virus strain was a likely progenitor for the virus strains present in 2015 (Figure 1, 2).
   a. The phylogenetic tree was created using the HA genetic portion of influenza AH3N2 virus strains from the years 1968 to 2015 of the same 40 states.
3. Each state's phylogenetic tree was observed and the genealogical distance from the top ranked strain and the 2015 strains was recorded (Figure 1, 2)

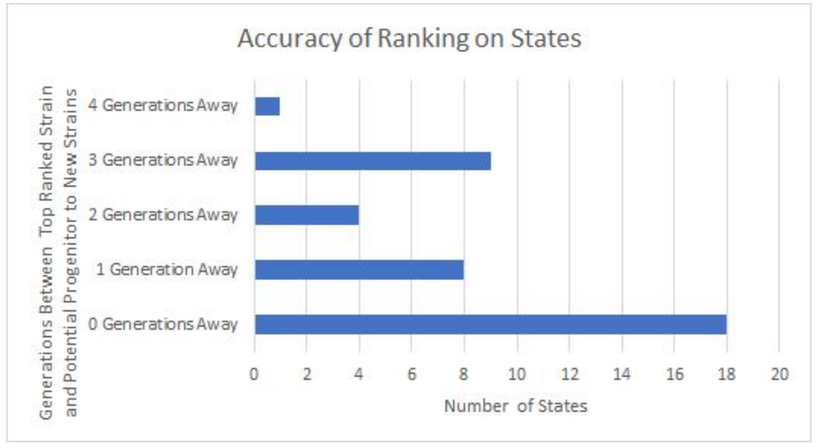

Figure 3: Results of the comparison between top ranked virus strain and it's place in the phylogenetic tree. Strains that are zero generations indicate a high accuracy prediction, whereas strains that four generations away indicate a low accuracy prediction.

Results

1. The code provided by Neher et al. made a high accuracy prediction (zero or one generation away) for 65% of the samples (Figure 3).
2. The code provided by Neher et al. made a medium accuracy prediction (two generations away) for 10% of the samples (Figure 3).
3. The code provided by Neher et al. made a low accuracy prediction (three or four generations away) for 25% of the samples (Figure 3).

Glossary of Technical Terms

Rank: the program provided by Neher et al. ranks the strands of the virus according to the shape of a genealogical tree, and ranks the strands based on fitness

Fitness: the fitness of the virus is the likelihood that it will survive and reproduce itself

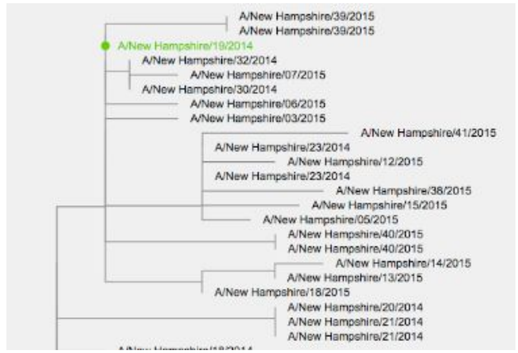

Figure 1: The top ranked virus strain for New Hampshire is highlighted in green. This strain is zero generations from the 2015 mutations, giving this prediction a high accuracy.

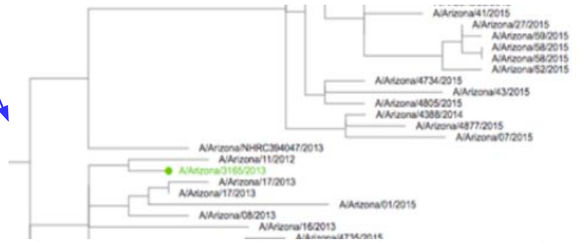

Figure 2: The top ranked virus strain for Arizona is highlighted in green. This strain is three generations from a connection to the 2015 mutations, giving this prediction a low accuracy.

References

1. Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *ELife, 3.* doi: 10.7554/elife.03568
2. Gregory, T. R. (2008). Understanding Evolutionary Trees. *Evolution: Education and Outreach, 1*(2), 121–137. doi: 10.1007/s12052-008-0035-x
3. Branswell, H. (2019, September 30). Flu Vaccine Selections Suggest This Year's Shot May Be Off the Mark. Retrieved from https://www.scientificamerican.com/article/flu-vaccine-selections-suggest-this-years-shot-may-be-off-the-mark/
4. Maragakis, L. L. (n.d.). Coronavirus Disease 2019 vs. the Flu. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-disease-2019-vs-the-flu
5. Influenza Research Database - Influenza genome database with visualization and analysis tools. (2020, March 22). Retrieved from https://www.fludb.org/brc/home.spg?decorator=influenza

Acknowledgments

This project was mentored by Nicole Fider, whose help is acknowledged with great appreciation.