PROJECT PROPOSAL

1. **The people on your team.**
   Emily Brunelli, Gunnar Johnson, Jonathan Mah

2. **A project title.**
   A Discrete-time Markov Process for Predicting Influenza Evolution.

3. **A clear statement of the goal of the project, and a clear explanation of the phenomenon to be modeled.**
   Influenza has 8 major protein coding regions, two of which are surface proteins suspected to be responsible for the functional mechanism by which Influenza escapes the human immune response. These two surface proteins are Hemagglutinin and Neuraminidase, and are commonly used to identify specific strains of Influenza (e.g., H1N1 or H3N2). By analyzing several years worth of sequenced DNA, we are able to approximate the annual rate of amino-acid substitution for each site in a protein. **Our goal is to investigate the use of these rate values to construct a Markov Chain in order to model the long-run behaviour of amino-acid substitution in Hemagglutinin.** This model should provide insight towards the evolutionary behavior Hemagglutinin, and by extension, insight towards possible targets for vaccine design.

4. **Importance of the problem. Who will care about your solution? How did you come up with the project idea?**
   Certain viral diseases, like Influenza and HIV, currently lack an effective treatment due to the extreme rate and diversity of viral evolution. In hopes of one day curing these diseases, it is imperative to understand the underlying biological mechanisms which allow these diseases to escape the human immune response. One of our members works in a lab which investigates this type of problem, but uses experimental (wet-lab) data and maximum likelihood instead of discrete mathematics. We briefly searched through relevant literature and found that Markov chains can be used to simulate the traversal of so-called “evolutionary space”.

5. **An idea of the methods you plan to use in modeling your problem.**
   Given input data consisting of strain-specific FASTA sequences, we will iterate through a site-specific discrete-time Markov chain in an attempt to predict the rate and diversity of the indicated virus’s evolution. Additionally, we may use random walk or Monte Carlo sampling to detect individual sites under positive selection, meaning sites which evolve “faster” than expected.

6. **An idea of how you will gather data for your project. You must use real data whenever possible.**
   We will be downloading FASTA alignments from the NCBI Influenza Virus Database.

7. **Provide a small example of your main problem. Translate it into a mathematical problem which you can solve.**
A basic form of this problem involves 4-dimensional FASTA sequences. After being chronologically sorted, a "sequence matrix" is constructed where each row represents one sequence. Difference vectors are calculated by column in order to calculate an index-specific probability matrix, denoted as A_i.

$$A_i = \begin{bmatrix} \pi_{AA} & \pi_{AC} & \pi_{AG} & \pi_{AT} \\ \pi_{CA} & \pi_{CC} & \pi_{CG} & \pi_{CT} \\ \pi_{GA} & \pi_{GC} & \pi_{GG} & \pi_{GT} \\ \pi_{TA} & \pi_{TC} & \pi_{TG} & \pi_{TT} \end{bmatrix}$$

where π_{jk} represents the probability of a site i evolving from j to k. We plan to expand upon this problem by translating DNA into amino-acids, which results in FASTA sequences with 20 dimensions instead of 4, and consequently, a 20×20 transition matrix. Additionally, we were interested in possibly computing the variation of a sites evolution, and mapping out specific sites which evolve faster than expected.

8. **A reference related to your topic (research articles, news articles, textbooks, previous work on the subject).**
This article from the Bedford Lab:
http://bedford.io/pdfs/papers/morris-predictive-modeling.pdf
titled "Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology" is a good review of current efforts to predict Influenza evolution for informed vaccine design.

9. **A reference related to your proposed method of solution.**
This article by Sergei Pond:
https://academic.oup.com/mbe/article/22/5/1208/1066893
titled "Not So Different After All: A Comparison of Methods for Detecting amino-acid Sites Under Selection" gives a great review of different methods for tackling this type of problem. Most methods start from the basic 4×4 site-specific transition matrix and then expand upon it. For example, some papers introduce external weighting factors like protein structure, and other papers pre-define a site's expected rate of evolution in order to detect abnormalities.