Predicting Evolution of Resistance Mutations to NRTI Drugs in HIV-1 Using Computational and Statistical Methods

As a result of its high variability and fast evolution, antiretroviral drug treatment for HIV is often unsuccessful. Many modern drug treatments are a regimen of combined and changing antiretroviral drugs to try and circumvent HIV's quick evolution.

This research aims to provide an easier way to predict the fitness (relative ability to survive and reproduce) of different strains of HIV under drug stress. This model can be used to predict further drug resistance mutations as well as predict conserved domains in the amino acid sequence. This will allow for easier treatment of HIV through antiretroviral drugs as well as inform scientists in drug design. In this project, a basic model of the fitness of each strain of HIV was used based on the Ising Spin Model in particle physics (Shekhar 1). This statistical model can predict the relative fitness of each new amino acid sequence of HIV based on its similarity to a consensus sequence, taking into account common double mutations. This model has been shown to also correlate with in vitro fitness of HIV. The model developed in this project is two Ising Spin Models – one using parameters created with drug-naïve sequences and one using parameters with drug stressed sequences. This allows the model to differentiate between resistance mutations and mutations common in a drug-naïve virus.

The result of this research is a model that is able to predict HIV evolution under drug stress, leading to some inferences about new resistance mutations. The model predicts 22 single mutations that confer resistance to NRTIs and 4,291 pairs of mutations that cause resistance. Further research is necessary to test this model in vitro with genetically engineered reverse transcriptase mutants in order to check if the mutations predicted by the model do indeed cause antiretroviral resistance.

Introduction

The treatment of Human Immunodeficiency Virus (HIV) is a major focus of immunologists and virologists. However, the virus tends to evolve very quickly inside a host to avoid immune and drug selective pressure and diversifies with low selection pressure among hosts (Vracken 2). Within hosts, the virus is heavily affected by a variety of immune responses, causing the virus to develop escape mutations quite quickly. Intra-host virus evolution is generally varied between different people and, as a result, most HIV response to immune pressure happens after infection. However, inter-host evolution does not appear to be driven by a similar selective process – strain variety in HIV is most often a result of the "founder effect", in which by chance the first strain in the area is the most prevalent (Lemey 2).

However, one major selective pressure on HIV is drug pressure. There are four main types of antiretroviral drugs that can treat HIV – Nucleoside Reverse Transcriptase Inhibitors (NRTIs), Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), Protease Inhibitors (PIs), and Integrase Inhibitors (INIs). NRTIs and NNRTIs attack the virus' reverse transcriptase protein, which allows it to insert its RNA genome into the genome of the cell host. Nearly twenty-five antiretroviral drugs have been licensed for the treatment of HIV-1, the most common ones being the NRTI and the PI (Shafer 1). HIV-1 develops drug resistance mutations to these variety of drugs very quickly, including more than 50 reverse transcriptase mutations. These mutations are common across different hosts and as a result, these are one example of selective inter-host evolution. A list of these mutations is publicly available in the Stanford HIV Drug Resistance Mutation Database, which hosts not only the list of antiretroviral resistance mutations, but also a scoring system for the effectiveness of each drug under specific mutations, estimates of decreased drug susceptibility, and comments about each resistance mutation in the sequence (Tang 1). However, this database does not contain every studied drug resistance mutation.

A *fitness model* is a computational tool that is able to find viral fitness (ability to survive and reproduce) as a function of amino acid sequences. A fitness model is useful because different mutations have different effects on the virus protein. Some mutations may affect the structure (and function) of the protein more than others. As a result, some viruses (like HIV) develop compensatory mutations – double, triple, or even quadruple mutations that together increase the *net* fitness of the protein, even if one mutation negatively affects the structure of the protein (Ferguson 1).

Fitness landscapes have been created for many types of viruses, notably in depth for influenza A (Łuksza 1) to predict which strains will be most successful in each successive season. This research has even been used to create real-time flu sequence trackers (Neher 1).

However, HIV is a more complex virus than influenza. Prediction of its sequence is a lot more demanding on researchers. A basic model of the fitness of each strain of HIV has been created based on the Ising Spin Model in particle physics (Ferguson 1). This statistical model can predict the relative fitness of each new sequence of HIV based on its similarity to a consensus sequence. This model is useful because it only calculates the fitness of a protein based on its single and double mutations, and has been shown to predict triple and quadruple mutations (Ferguson 3). This model has been used by some researchers, notably in predicting HIV's evolution as a result of immune pressure from the human body's series of weak immune attacks on the virus. This model has been worked to fit simulations of HIV among many different hosts with different immune responses (Shekhar 1). The model has also been advanced to predict escape mutations to immune response by using a more complex technique – power law scaling (Barton 1), which is a way of evaluating the commonality of sequences of high fitness and in low fitness. This model of HIV fitness using immune pressure is one step on the long path to developing a vaccine to the virus.

Before a vaccine can be developed, killing the virus is the way to treat those who get it. However, as a result of its high variability and fast evolution, antiretroviral drug treatment for HIV is often unsuccessful. Many modern drug treatments are a regimen of combined and changing antiretroviral drugs to try and circumvent HIV's quick evolution. Using an approach to develop a fitness model for drug pressures similar to that used by Ferguson et. al. for immune pressures may allow researchers to develop drug regimens for patients more easily and develop new drugs that avoid possible resistance mutations. In the short term, this model can predict further resistance mutations for study that are not contained in the HIV Drug Resistance Mutations database. Therefore, this research study was conducted to create a model to predict the drug resistance of a strain of HIV-1 and to generate resistance mutations that are not contained in the HIV Drug Resistance database.

Materials and Methods

All computation in this research was performed on an 11-inch mid-2013 MacBook Air, at maximum using only a few hours. Two databases were used – the HIV Drug Resistance Database (hivdb.stanford.edu) and the HIV Sequence Database (hiv.lanl.gov). To create multiple sequence alignments (MSAs) and consensus sequences, software tools were used – Unipro UGENE as a GUI and consensus finder and MUSCLE for aligning sequences. The actual mathematical model was programmed entirely in-house in Java.

The model itself is based upon a mathematical function that takes as input a binary vector representing the amino acid sequence and outputs the computed probability of finding that particular sequence in a person. The sequence translated into the binary vector $\vec{z}$ in which $z_i = 0$ if the sequence has the same amino acid as the wild-type protein and $z_i = 1$ if the sequence and WT differ. Note that this “binary approximation” of the protein loses some resolution (as some pairs of amino acids are more similar than others), but it simplifies computation immensely. Using a regularized Potts model would remove this problem (Mann 3).

The basic model, the Ising Model, is calculated in equation 1:

$$H(\vec{z}) = \sum_{i} h_i z_i + \sum_{i<j} J_{ij} z_i z_j \qquad (1)$$

This uses vector constants $\vec{h}$ and $\vec{J}$, both of which are calculated based on initial input values.

Using H($\vec{z}$) (also known as the Hamiltonian), one can then calculate the probability of finding a strain $\vec{z}$ using equation 2:

$$P(\vec{z}) = \frac{1}{Z} e^{-H(\vec{z})} \qquad (2)$$

Where Z is the partition function, the sum of all possible binary sequences, as described by equation 3. This is calculated through a Monte Carlo sampling.

$$Z = \sum_{\{z_i\}=\{0,1\}} e^{-H(\vec{z})} \qquad (3)$$

To calculate the vector $\vec{h}$, a simple form of supervised learning was used – inputting a MSA of specific HIV amino acid sequences and implementing a form of gradient descent known as Boltzmann learning (Roudi 2), which aims to have the average probability of a single mutation z_i in the training data equal that amino acid's average probability in the Ising distribution (equation 4). η is the constant learning rate.

$$\delta h_i = \eta (< z_i >_{data} - < z_i >_{Ising}) \quad (4)$$

"Average probability over the Ising distribution" is calculated using equation 5:

$$< z_i >_{Ising} = \frac{\sum_{\forall \vec{z}: z_i = 1} \frac{1}{Z} e^{-H(\vec{z})}}{\sum_{\forall \vec{z}: z_i = 1} 1} \qquad (5)$$

This averages all possible sequence's probabilities that contain a mutation in position i. Since there are a huge number of possible sequences, Metropolis Monte Carlo Sampling (MMC) is used. MMC is used instead of regular Monte Carlo sampling because it is much more time efficient, only requiring calculation of the Hamiltonian once and augmenting its value each iteration with O(1) time complexity (Beichl 2). The MMC algorithm for the Ising model is well-documented and is not described here, but can be found in “The Metropolis Algorithm”, by Beichl and Sullivan.

Continuing with calculating the constant vectors in the Hamiltonian, the independent-pair approximation is employed (Roudi 3) to calculate $\vec{J}$, shown to correlate with Boltzmann learning (Roudi 5). The independent-pair is used to reduce runtime; this particular approximation was used due to its ease of implementation. The independent approximation is calculated in equation 6:

$$J_{ij} = \frac{1}{4} ln(1 + \frac{< z_i z_j >_{data}}{(1+ < z_i >_{data})(1+ < z_j >_{data})}) \quad (6)$$

where $<z_i z_j>_{data}$ is the probability of a double mutation at positions i and j in the training data.

Using the definition of the Ising model in equations 1-3 and the description of calculating constants in 4-6, the model for the actual model of HIV drug resistance can be described. The model involves two Ising spin models (equation 1) that are created using different vector constants $\vec{h}$ and $\vec{J}$ due to different training Multiple Sequence Alignments (MSAs).

Both MSAs were taken from the HIV Drug Resistance Database – one for isolates untreated by NRTI drugs and one for treated NRTI drugs. These were all taken from HIV-1, subtype B, using the amino acid sequence for the reverse transcriptase protein. The untreated data included 35 isolates (the maximum number of drug-naïve isolates in the HIV Drug Resistance Database); clones were analyzed by consensus sequence. The treated data included 1346 isolates that had a treatment shift from 0-8 NRTI drugs to 1-8 different NRTI drugs. Using MUSCLE, these sequences were aligned and cut to contain 410 amino acids each and a consensus sequence was found using Unipro UGENE (the consensus sequence used for *both* models was created from the untreated data). The untreated data creates a Hamiltonian for an environment without drugs and the treated data creates a Hamiltonian for an environment where NRTI drugs are present.

These two different MSAs were then fed into the Boltzmann learning algorithm described by Equation 4; the $\vec{h}$ vector was calculated after 15 steps of Boltzmann learning and a learning rate η of 2. The value of the partition function was calculated with 100000 Monte Carlo samples and the Metropolis Monte Carlo for averaging over the Ising Distribution used 10000 walking steps with one acceptance for every 500 changes (a rejection rate of 500). These numbers could certainly be made larger for a more accurate result using more computation.

Then, the two Ising models were used to predict new resistance mutations by assigning each possible single mutation (out of 410 possible amino acids) and each possible double mutation (out of 83,845) a score, calculated in equations 7 and 8:

$$i_{score} =< z_i >_{treated} - < z_i >_{untreated} \quad (7)$$

$$\{ij\}_{score} =< z_i z_j >_{treated} - < z_i z_j >_{untreated} \quad (8)$$

After scoring every possible single mutation and double mutation, the average score of known resistance mutations in the HIV Database was calculated. All mutations not contained in the Database with a score above this average were considered candidate resistance mutations.

The model operates under several assumptions:

1. The level of accuracy calculated in this particular experiment was sufficient to make some rough predictions.
2. The mutation scores calculated in Equations 7 and 8 will be significantly different between “known” resistance mutations and other mutations.
3. For a non-resistance mutation i, $<z_i>_{untreated}$ and $<z_i>_{treated}$ are essentially equivalent.
4. For a resistance mutation j, $<z_j>_{untreated}$ and $<z_j>_{treated}$ are significantly different, with $<z_j>_{treated}$ being higher than$<z_j>_{untreated}$ because the mutation should have a higher probability when in an environment where NRTI drugs are present.

These assumptions will be further addressed in the “Results” section.

Results

Figure 1 shows the output of the probabilities of each possible single mutation in the reverse transcriptase protein. High probabilities are more likely to be found in a colony with the specified environment.

Figure 2 is a graph of the scores of each possible single mutation. Higher numbers are more likely to be resistance mutations than lower numbers. Numbers over the cutoff are predicted resistance mutations. The cutoff is the average score of a resistance mutation.

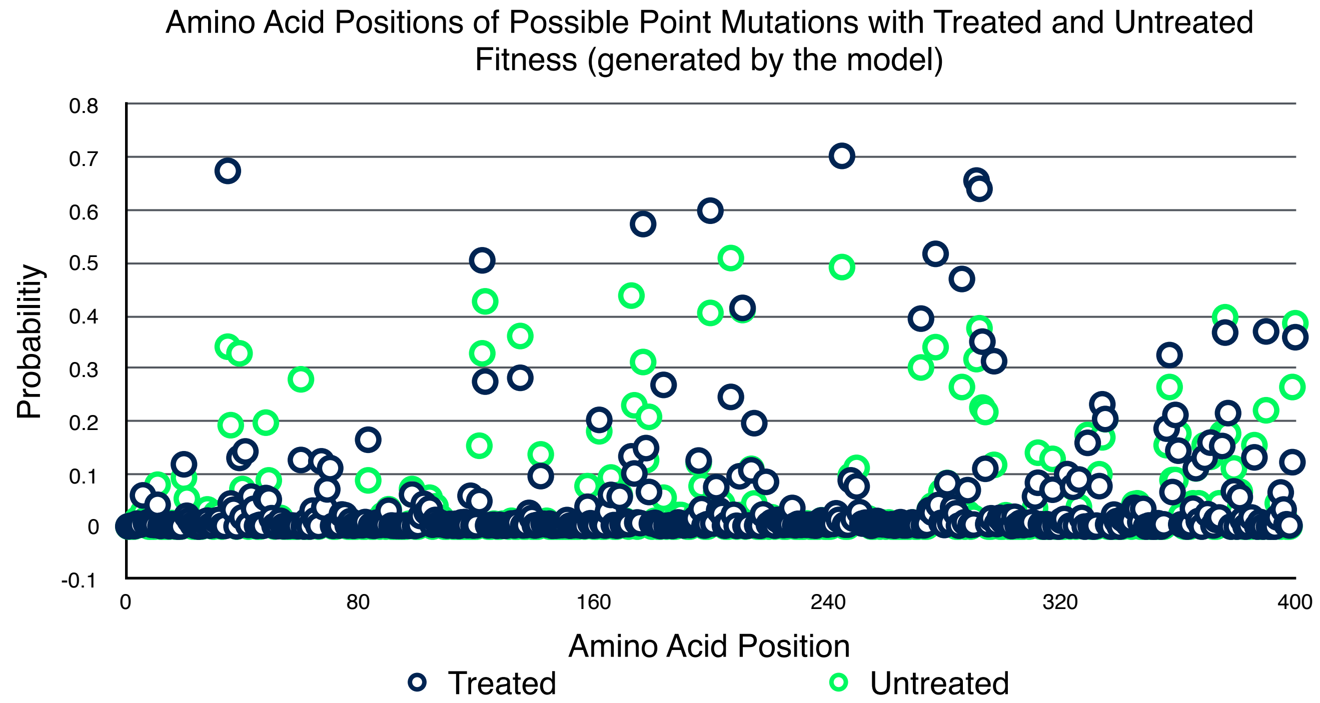

Figure 1. Graph of amino acid positions of possible point mutations with treated and untreated fitness (generated by the model).

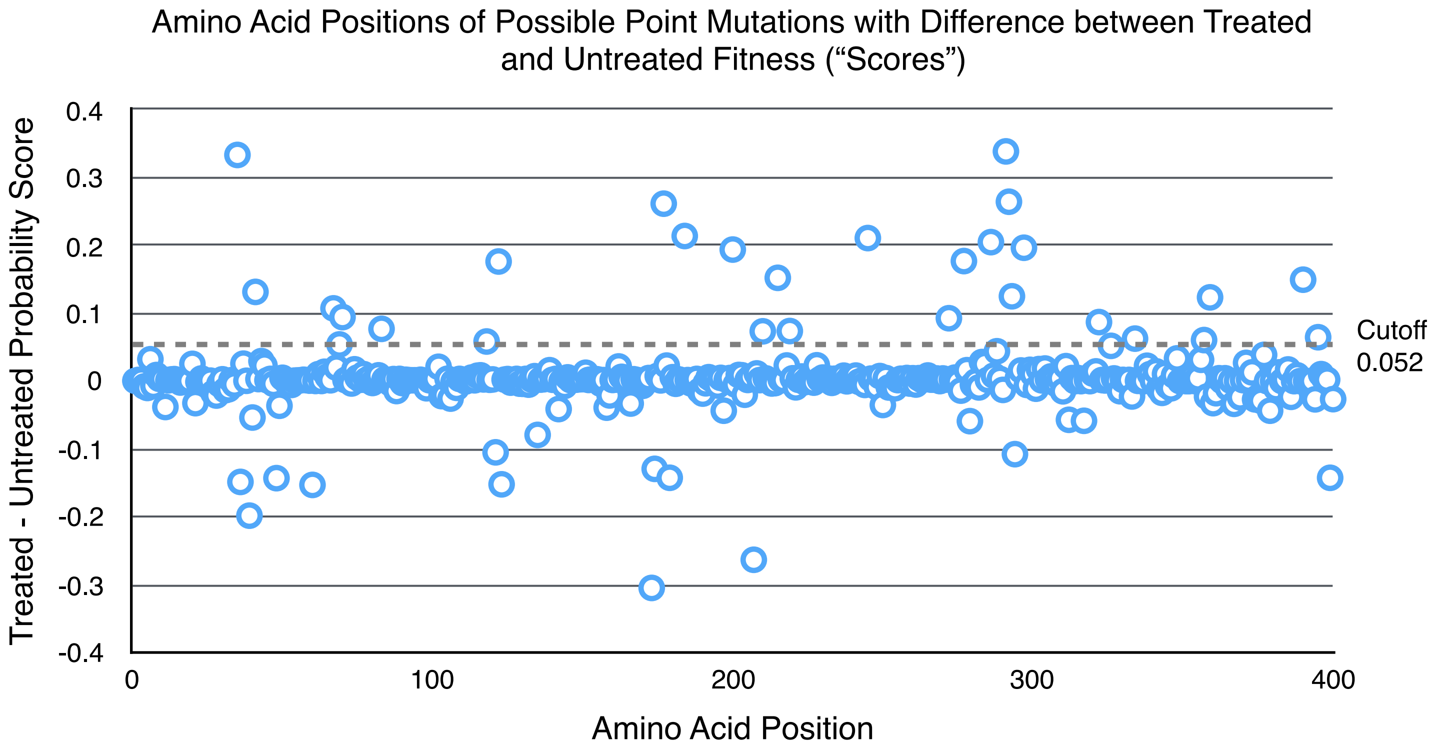

Figure 2. Graph of amino acid positions of possible point mutations with difference between treated and untreated fitness ("scores").

Table 1. New Predicted Resistance Single Mutations (Amino Acid Positions).

Existing Resistance Mutations (20)	**New Resistance Mutations (22)**
40, 41, 44, 62, 65, 67, 69, 70, 74, 75, 77, 115, 116, 118, 151, 184, 210, 215, 219, 348	35, 83, 122, 177, 201, 211, 216, 220, 246, 273, 278, 287, 292, 293, 294, 298, 323, 335, 358, 360, 391, 396

Table 1 displays the results of the model – new resistance single mutations

Double mutations are not graphed due to the huge amount of data involved, and the predicted 4,291 double mutations are not pasted here. The predicted double mutations can be found on this project's GitHub at https://github.com/gautam-prab/hiv-prediction/blob/master/FinalData/CandidateDoubleMutations.txt.

Three different t tests were performed in order to validate the model for single mutations. The result of the t tests is that the model accurately differentiates resistance and non-resistance mutations for both single and double mutations.

Summary –

1. Paired t-test of resistance treated vs. untreated: p=0.0015, statistically significant.
2. Paired t-test of all treated vs. untreated: p=0.0448, not as statistically significant.
3. Welch's t-test of resistance vs. non-resistance: p=0.0024, statistically significant.

The first paired t-test is between resistance mutations using treated constants and untreated constants. This tests assumption 4 of the model – these two should have a statistically significant difference in means, indicating that resistance mutations have, on average, larger probabilities when using treated constants rather than untreated constants.

Second, a paired t-test between all treated and untreated mutation probabilities. This should be not statistically significant or have a p value that is greater than the above p value, as that would indicate that untreated and treated have essentially equivalent means unless one is looking at resistance mutations. This tests assumption 3 of the model.

Finally, a Welch's t test between scores of resistance vs. non-resistance mutations. This tests assumption 2 of the model. If the assumption is valid, the difference between the two probabilities should be larger for resistance mutations than for non-resistance mutations. This is a Welch's t test because Welch's t-test is more accurate with two groups with radically different sample sizes.

The same process applies to double resistance mutations:

Summary –

1. Paired t-test of resistance treated vs. untreated: p=0.0000825, statistically significant.
2. Paired t-test of all treated vs. untreated: p=6.8*10^-255, very statistically significant.
3. Welch's t-test of resistance vs. non-resistance: p=0.0001, statistically significant.

Analysis and Conclusions

The single-mutation t tests went exactly as expected for the model, indicating that it is correctly predictive for single mutations.

The double-mutation t tests did not go as expected – the second t test had a much lower p value than expected. This may be due to a few reasons – most likely due to having too little data in the untreated section, causing the probability under untreated for many uncommon mutations to be ~0. Additionally, there may have been too little data in the HIV Drug Resistance Database, which only contained 17 pairs of mutations that were considered drug resistance double mutations. Doing this again with more data would yield more accurate results; however, this sort of data is very difficult to find. The model is still likely results in an accurate model because it still passes the other two t tests; the middle test is less important than passing the other two. However, attempting this again with better data, if it could be found, would perhaps improve the results.

The bottom line – the model works based on its t-tests; it generated 22 candidate single resistance mutations and 4,291 candidate double resistance mutations. Each possible resistance mutations need to be tested for prevalence in HIV in the wild, and perhaps tested in a laboratory setting.

Results were limited by computing power and time, running some of these algorithms again with more computational power and more data would greatly improve the accuracy of the model.

In order to use this model to assist in drug treatment of HIV further, it can easily be extended to involve all four major types of antiretroviral drug, allowing doctors to make more informed choices when choosing antiretroviral drug treatments for HIV patients.

Works Cited

Agyingi, Lucy, et al. "The Evolution of HIV-1 Group M Genetic Variability in Southern Cameroon is Characterized by Several Emerging Recombinant Forms of CRF02_AG and Viruses With Drug Resistance Mutations." *Journal of Medical Virology* 86.3 (2014): 385-93. *PubMed*. Web. 6 Oct. 2015.

Barton, John P., Mehran Kardar, and Arup K. Chakraborty. "Scaling laws describe memories of host–pathogen riposte in the HIV population." *Proceedings of the National Academy of Sciences* 112.7 (2015): n. pag. *Proceedings of the National Academy of Sciences*. Web. 15 Sept. 2015.

Beichl, Isabel, and Francis Sullivan. “The Metropolis Algorithm.” *Computing in Science and Engineering* 2.1 (2000): 65-69. *Hua Zhou | Stat 758*. Web. 3 Dec. 2015.

Ferguson, Andrew L., et al. “Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design.” *Immunity* 38.3 (2013): 606-17. *Cell*. Web. 26 Oct. 2015.

“HIV Sequence Database.” *HIV Databases*. Los Alamos National Security, 15 Mar. 2016. Web. 30 Mar. 2016. <http://www.hiv.lanl.gov/>.

Lemey, Philippe, Andrew Rambaut, and Oliver G. Pybus. “HIV Evolutionary Dynamics Within and Among Hosts.” *AIDS* 8.3 (2006): 125-40. *Department of Zoology, Oxford*. Web. 25 Sept. 2015.

Luksza, Marta, and Michael Lässig. “A Predictive Fitness Model For Influenza.” *Nature* 507.7490 (2014): 57. *Advanced Placement Source*. Web. 19 Aug. 2015.

Mann, Jaclyn K., et al. “The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing.” *PLOS Computational Biology* 10.8 (2014): n. pag. *PLOS Computational Biology*. Web. 26 Oct. 2015.

Neher, Richard A., and Trevor Bedford. “nextflu: real-time tracking of seasonal influenza virus evolution in humans.” *Bioinformatics* (2015): 1-3. Web. 6 Oct. 2015.

Njai, Harr F., et al. “The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness.” *Retrovirology* 3.40 (2006): n. pag. *Retrovirology*. Web. 6 Oct. 2015.

Palm, Angelica A., et al. “Faster progression to AIDS and AIDS-related death among seroincident individuals infected with recombinant HIV-1 A3/CRF02_AG compared to sub subtype A3.” *Journal of Infectious Diseases* 209 (2014): 721-28. *Journal of Infectious Diseases*. Web. 6 Oct. 2015.

Rambaut, Andrew, et al. “The Causes and Consequences of HIV Evolution.” *Nature Reviews Genetics* 5 (2004): 52-61. *Nature Reviews Genetics*. Web. 6 Oct. 2015.

Rhee, Soon-Yon, et al. “Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database.” *Nucleic Acids Research* 31.1 (2003): 298-303. Print.

Roudi, Yasser, Joanna Tyrcha, and John Hertz. “Ising model for neural data: Model quality and approximate methods for extracting functional connectivity.” *Physical Review E* 79.5 (2009): n. pag. *SNN Adaptive Intelligence*. Web. 3 Nov. 2015.

Shafer, Robert W., and Jonathan M. Schapiro. “HIV-1 Drug Resistance Mutations: An Updated Framework for the Second Decade of HAART.” *AIDS reviews* 10.2 (2008): 67–84. Print.

Shekhar, Karthik, et al. “Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes.” *Physical Review E* 88.6 (2013): n. pag. *American Physical Society*. Web. 15 Sept. 2015.

Tang, Michele W., Tommy F. Liu, and Robert W. Shafer. “The HIVdb System for HIV-1 Genotypic Resistance Interpretation.” *Intervirology* 55.2 (2012): 98-101. *Karger Publishers*. Web. 6 Jan. 2016.

Vrancken, Bram, et al. “The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates.” *PLoS Computational Biology* 10.4 (2014): n. pag. Web. 6 Oct. 2015.

Whitcomb, Jeannette M., et al. “Broad Nucleoside Reverse-Transcriptase Inhibitor Cross-Resistance in Human Immunodeficiency Virus Type 1 Clinical Isolates.” *Journal of Infectious Diseases* 188.7 (2003): 992-1000. *Oxford Journals*. Web. 12 Jan. 2016.

Reflection

I did this project because I am very interested in the application of biology and computer science to solve real-world problems, and this seemed like a great option to actually help people with the kind of knowledge that I have. I was curious about this topic because I think that virology and pathology have an intrinsic interest – especially because they are incredibly complex problems that can take very interesting solutions.

I expected to find out a lot about the reproduction of HIV and the kind of algorithms used to predict the fitness of viruses, and I of course learned a lot about those things. However, I also learned a lot about gathering and manipulating amino acid data and general algorithms for sampling the average value of a function. Even more importantly, I learned a lot about how difficult it is to read technical papers, and I think I am a lot more able in that field after doing the project .Exploring this subject did really help me to see scientific studies in a new light – I now understand how much of a struggle it is just to read scientific papers, let alone do a project and write a paper myself. A lot of this research project was struggling to finish on time and trying to make sure that my results were as accurate as possible. I didn't just code the entire thing once and be finished.

I do not think there is a topic that I would have liked to learn more about relating to this project; I read plenty of papers on most things that related to the project. This project was full of frustrations – inaccuracies in the results of my program or large bugs in the program that I had to slowly dissect – both on the surface, programming level and on the deeper, algorithmic level – in order to ensure a correct answer. I overcame these frustrations by slowly going through my process and ensuring that each portion was producing correct results. I had to run my program many times to ensure that it was still working; unfortunately the program took hours to run so I often had to run my program overnight and fix it in the morning.

If I were to redo the experiment, I would have been much more careful about the data I was using to generate my constants; I had to regenerate constants several times just because my constants were incorrect. I would have been a lot more careful about ensuring that each model had the same number of isolates (and made sure I could retrieve these isolates from the correct databases). For students who have to complete their own research studies, planning and time management are key. Expect that things will go wrong, so plan for extra time to fix any problems.

If I were to do another research study of this magnitude I have no idea what topic I would explore. I do this huge research project because I get to work for a long time learning about a topic that interests me and perhaps contributing to human knowledge about that topic. I learned many skills in the process – writing papers, reading articles, presenting the project, and many much more technical things. All of these things will apply to any future in STEM.