Finding Positively Selected HIV Mutations in Longitudinal Data

Kevin Hu[1], Tiana Pereria[2] and Michael Lanthier[3]

***Abstract*— Drug resistance is a continuing problem when treating HIV/AIDS. Because of the very high mutation rates of HIV, drug resistance is developed extremely quickly. In order to best treat HIV, being able to identify mutations associated with drug resistance is extremely important. Through the use of a longitudinal data set of 1,700 patients, we performed analysis of HIV reverse transcriptase to identify positively selected positions that may incur drug resistance.**

I. INTRODUCTION

HIV is an infectious virus containing an RNA-based genome. It hijacks and utilizes the host cells enzymes in order to replicate and multiply. There are three key proteins vital to the HIV life cycle: reverse transcriptase (RT), integrase, and protease (PR). Many drugs targeting these key proteins have been developed to treat HIV infection. Integrase inhibitors, non-nucleoside RT inhibitors (NNRTI) and nucleoside RT inhibitors (NRTI) target RT, and protease inhibitors prevent normal HIV enzymatic activity. However, due to high rates of mutation (owing to RTs error rate), the virus acquires drug resistance extremely rapidly. Mutations in RT, integrase, or PR that confer drug resistance is likely to be positively selected. Therefore, identifying positive selection in HIV is crucial for the development of new drugs and prognoses.

Chen et al. has previously described a method using the K_a/K_s ratio to measure positive selection in HIV (1). A K_a/K_s value >1 indicates positive selection, a K_a/K_s value $=1$ indicates neutral selection, and a K_a/K_s value <1 indicates negative selection. In their study, 40,000 HIV sequences were used to identify positively selected amino acid positions in PR and RT. With their method, they were able to identify 19/23 and 20/34 known drug resistant positions in PR and RT, respectively (1). Chen et al. also discovered novel positions that are candidates for drug resistance or fitness.

However, the K_a/K_s method requires an extraordinarily large data set that may not be easily obtained. Instead, we applied K_a/K_s using a longitudinal set of HIV sequences (roughly 1750 patients) to identify positive selection. We hoped to still find positively selected positions using this data because longitudinal data should reduce noise from evolutionary drift. Even with less sequences, the higher data quality should allow us to match the positions found in Chen et al. We also hope to show our method is more efficient in identifying positively selected positions and even novel ones.

[1]K. Hu is an undergraduate within the Physical and Biological Sciences department at the University of California, Santa Cruz.

[2]T. Pereria is an undergraduate within the Bio-molecular Engineering Department in the Jack Baskin School of Engineering at the University of California, Santa Cruz.

[3]M. Lanthier is an undergraduate within the Computer Science Engineering Department in the Jack Baskin School of Engineering at the University of California, Santa Cruz.

II. MATERIALS

A. Sequences

Testing was performed on a longitudinal data set of RT sequences taken from 1,715 patients with HIV. Every patient has a corresponding FASTA file which contains at least two sequences, an initial sequence (t_0) and a final sequence (t_f). In between these two time points, all patients had received new drug treatments. While not all, many of the patients had received drug treatment prior to t_0 as well.

All the sequences were aligned prior to receiving the data. This was further verified by running a multiple sequence alignment on all the t_0 sequences and verifying they remained unchanged.

In order to provide a reference for comparison, a reference sequence of HIV was used. A full HIV sequence was used (HIV-1 vector pNL4-3). In order to rerun the tests using the reference sequence, every patient FASTA file was realigned to the reference using the MUSCLES multiple sequence alignment tool. After alignment both the reference sequence and patient sequences were trimmed to align with the original patient sequences in order to make codon comparisons.

III. METHODS

A. Calculations of K_a/K_s Ratios

Our K_a/K_s ratios are calculated as defined by Chen et. al. The main difference is rather than calculate ratios for mutations to specific amino acids, our method calculated mutations for any amino acid.

Just like Chen et. al. we calculated K_a/K_s ratios at every codon and normalized them by the random mutational model as defined by them. In order to provide an accurate model of random mutations, we had to take into account the high transition/ transversion ratio that HIV has. In order to do so, we went through all of the t_0 and t_f sequences and counted the number of transitions and transversions we encountered. Because of the varying sequence lengths the formula used in previous methods was not helpful. We iterated through all the patients counting the number of transitions (N_t) and transversions (N_v) as well as the total number of known bases encountered (N_b). The transition and transversion frequencies were then calculated using the following formulas as the transition and transversion frequencies, respectively: $f_t = N_t/N_b$ and $f_v = N_v/N_b$. The approach above ignored ambiguous bases, only counting A, T, G and Cs. From there we are able to calculate the full K_a/K_s ratio as defined by Chen et al.

$$\frac{K_a}{K_s} = \frac{\frac{N_y}{N_s}}{\frac{n_{y,t}f_t + n_{y,v}f_v}{n_{s,t}f_t + n_{s,v}f_v}} \quad (1)$$

N_y/N_s is the true K_a/K_s ratio. This is the number of non-synonymous mutations observed at that specific codon(N_y) divided by the number of synonymous mutations observed (N_s). This ratio is then normalized by the denominator, the random mutational model. $n_{y,t}$ is the number of possible transition mutations in the codon while $n_{s,t}$ is the number of possible transition mutations at the codon that are synonymous. $n_{y,v}$ and $n_{s,v}$ are the same as above but for transversions.

This produces the equation defined above. When the numerator is greater than the random mutational model we get a K_a/K_s ratio greater than one. That indicates that there is positive selection pressure at that codon.

B. Significance Testing

1) Average K_a/K_s Ratios: In order to derive statistical significance from our relatively small dataset we are using bootstrap replication. Every replication selects twenty percent of the data and calculates K_a/K_s ratios based on them. After running 10,000 bootstrap replications, the K_a/K_s ratios are averaged at every codon.

2) P-Value Testing: In order to run a P-value test, we define a null hypothesis as whenever the K_a/K_s ratio is less than or equal to one. This is an indicator of neutral or negative selection. While bootstrapping, count the number of times the null hypothesis is rejected at every codon. Then calculate the P-values at each codon as defined below.

$$P_{value} = 1 - \frac{\#\ of\ rejections\ of\ the\ null\ hypothesis}{Number\ of\ bootstrap\ replicates} \quad (2)$$

C. Code Structure

The program was written using Python 3.7 and intended to be run using a linux based operating system. It takes in all the patient FASTA files and reads them into a list of all patients. The mutations that are observed between t_0 and t_f are all identified as well and stored for each patient as well. In addition to that, the transition and transversion frequencies are calculated. From there bootstrapping occurs. Twenty percent of the patients are selected and K_a/K_s ratios are calculated for each codon. 10,000 different iterations are then run generating many lists of K_a/K_s ratios. From there the positively selected ratios are counted and all ratios are averaged at each codon.

IV. RESULTS

We found 40 codons in RT that had a K_a/K_s value greater than one, indicating positive selection. However, when testing the p-value for each of these codons, only 7 passed our statistical significance testing. All 7 codons were additionally found in the previous literature, as well as 4 of the 7 matching known drug resistance mutations in the Stanford HIV mutation database. These include codons 35, 41, 103, 135, 184, 200, and 215 (shown in figure 2). All of these mutated codons have K_a/K_s values less than 10, with the exception of codon 184 with a K_a/K_s value of about 39. Out of the 33 codons that failed the significance test, 15 of which were found in previous literature while 9 matched other known mutations for drug resistance. The p-values for every positively selected codon indicated, including those that passed and failed the p-test, are included in figure 3.

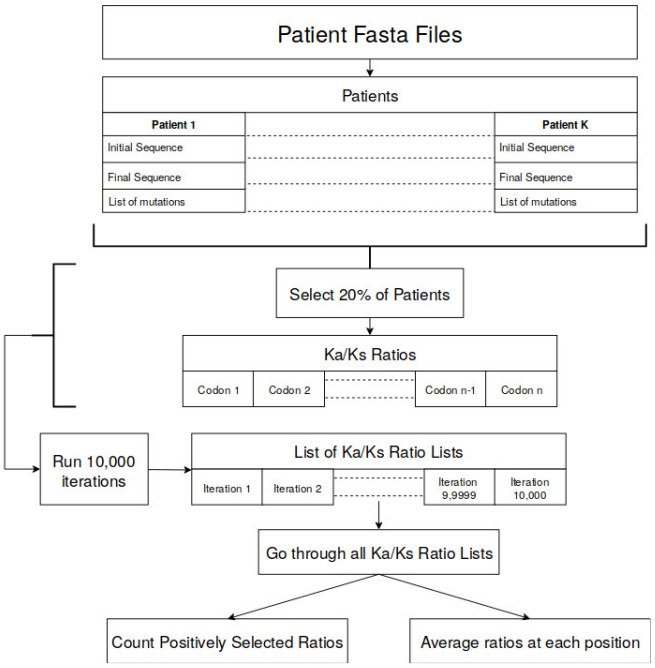

Fig. 1. K_a/K_s Flow of data within the program

In addition to testing the longitudinal data, we used the same methods but compared each t_f sequence to an reference genome. In this experiment, we found 95 codons that indicate positive selection, with 59 codons passing the p-test. Three of these codons matched known drug resistance. Of the 36 codons that did not show statistical significance, three were found in previous literature and one matched known drug resistance mutations in the Stanford HIV mutation database(2,3).

V. DISCUSSION

Using the longitudinal data set, we were able to identify 40 positively selected positions in RT. However, only 7 positions are of statistical significance. In parallel, comparing HIV sequences from patients at t_f to a reference sequence performed better. Using this comparison we identified a total of 95 positively selected codons, 59 of those statistically significant. We reasoned that the reference-based comparison performed better because positive selection is more clearly defined and identifiable over a longer evolutionary distance.

The longitudinal data set allowed us to identify 13 drug resistant mutations (4 being significant) compared to 4 in the reference-based comparison (3 being significant) based on previously published literature (1,2,3). Since the longitudinal data compares HIV sequences during patient drug treatment,

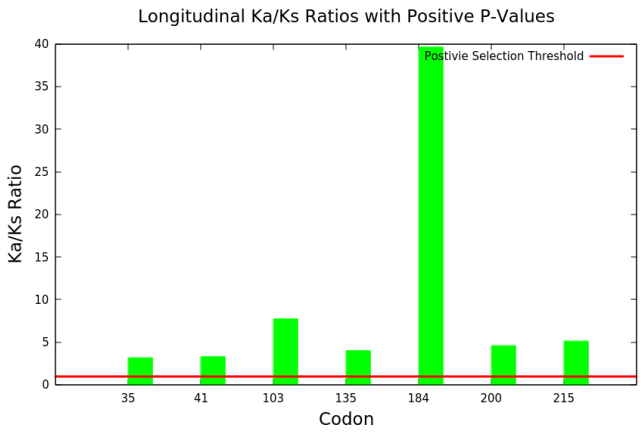

Fig. 2. K_a/K_s ratios calculated using the longitudinal data. All ratios have passed the P-value test

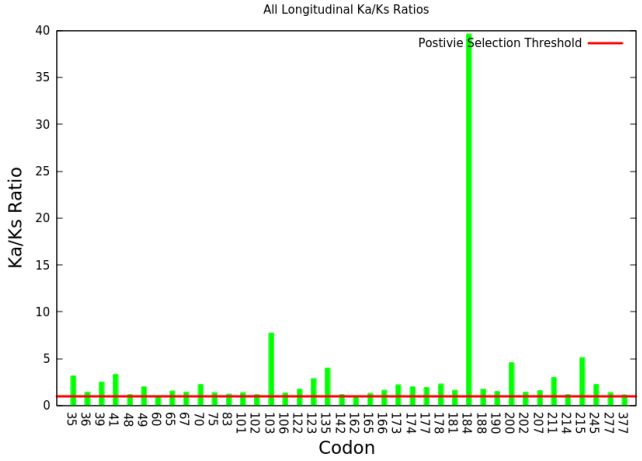

Fig. 3. K_a/K_s ratios calculated using the longitudinal data regardless of P-values.

it is not surprising the data set allowed us to identify three-fold more resistance mutations.

Chen et al. used the Ka/Ks ratio to identify positively selected codons, but also at the level of individual amino acid changes. In fact when they looked at protease, some individual amino acid mutations displayed positive selection even though the codon position itself was negatively selected. There may be a similar case in RT where this is possible. Thus, we may have not utilized the utmost power of our method. It is quite possible that our results may yield more promising results if we also analyzed such mutations.

Another consideration is altering our method of accounting for statistical significance. The sample size of the longitudinal data was 1750 sequences compared to 40,000 used in the Chen groups study. This may be a big factor resulting in the lack of statistical significance of our results. Since our sample size is quite small, we decided to sample 20% of our data set over 10,000 bootstrap replicates to calculate p-values for each codon. This may have even further diluted the already small sample size.

REFERENCES

[1] Chen, L., Perlina, A., & Lee, C. J. (2004). Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. Journal of virology, 78(7), 3722-3732.

[2] Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J. Betts, Jaideep Ravela, and Robert W. Shafer (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Research, 31(1), 298-303

[3] Shafer RW(2006). Rationale and Uses of a Public HIV Drug-Resistance Database. Journal of Infectious Diseases 194 Suppl 1:S51-8

[4] Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97

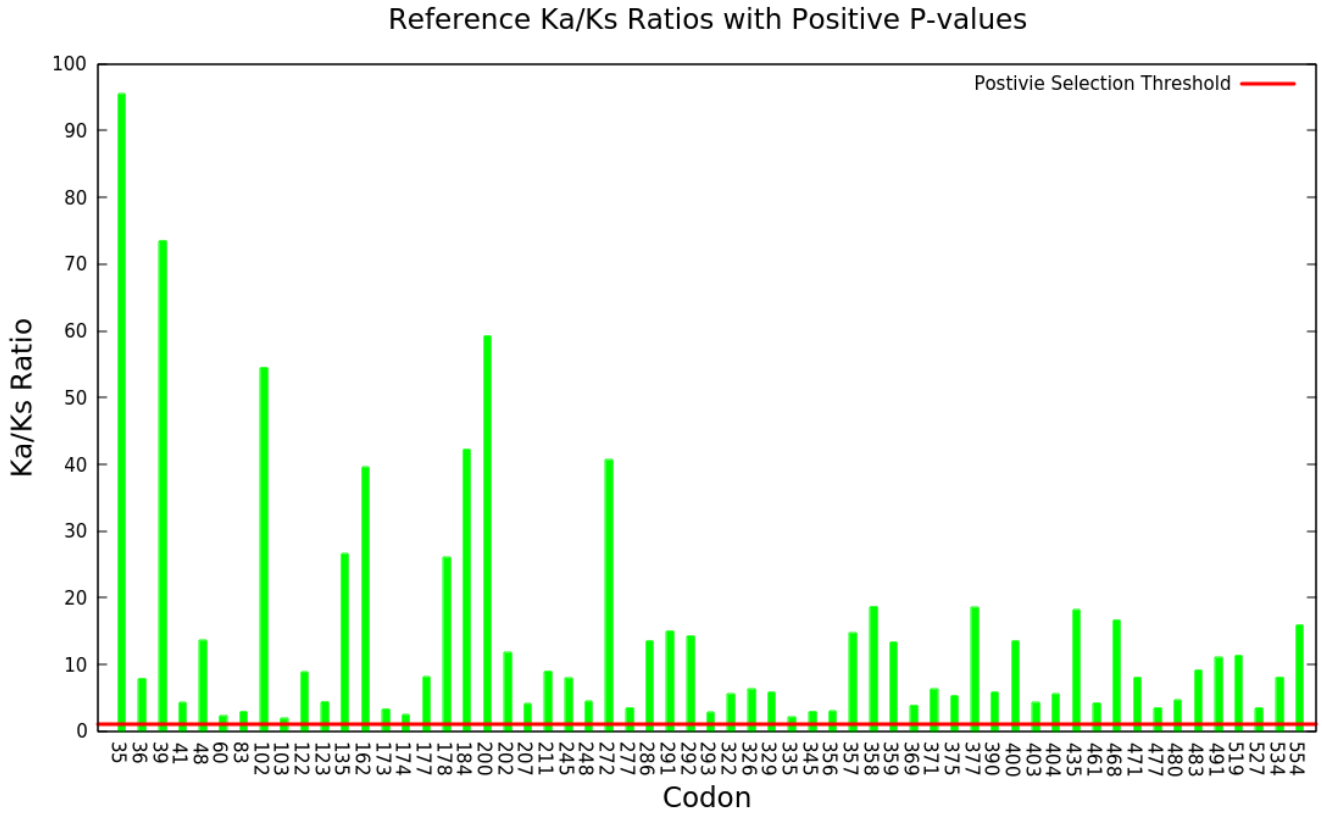

Fig. 4. K_a/K_s ratios calculated using the reference genome. All ratios have passed the P-value test