Determining the mutation rates of HIV over its viral genome

Noah H. Guale

Amaral Lab Northwestern University

Abstract

Human Immunodeficiency Virus (HIV) is a lentivirus (a subgroup of retrovirus) that is spread through certain bodily fluids and attacks the body's immune system, specifically th*e* *CD4+ cells* (T-lymphocytes). These *CD4+* cells are what help the immune system fight off infections. If left untreated, the virus will reduce the number of *CD4+ cells* to the point the body cannot fight off other diseases. At that time the person will have developed AIDS, the last stage of the HIV infection. And since its discovery, around 30 million people have died from the virus and currently 35 million are diagnosed. No efficacious cure currently exists, as the virus is known for its high rates of mutations, thwarting the development of treatments. However new research development with CRISPR, a form of genome editing, there is hope to completely remove the virus from the system. By conducting a quantitative analysis on the HIV's genome sequence, the mutation rates at each position of the virus's genome sequence can be calculated. This analysis zones in on regions with high rates of mutations and functionality of the virus. With this information, scientists can target those portions that are functional with tools such as CRISPR/Cas9 and render the virus obsolete.

Background Information

The Human Immunodeficiency Virus (HIV) is a lentivirus that can infect the body and with time will develop in Acquired Immunodeficiency Syndrome (AIDS). The primary function of the virus is preventing the body's immune system to function properly by attacking *CD4+ cells* (T-lymphocytes). This increases the probability of the carrier developing other infections and diseases. If the infection is not treated, the person will develop AIDS.

There are two distinct types of HIV – HIV-1 and HIV-2. The most predominant type is HIV-1, which accounts for 95% of infections worldwide. Within HIV-1 there are known to be at least nine genetically distinct subtypes of HIV-1 – A, B, C, D, F, G, H, J, and K. Sadly, even though 50% of people with HIV have subtype C, less research is done for that subtype. This subtype is common among countries in Southern and East Africa.

HIV is spread only in certain body fluids from a person infected with HIV. These fluids are blood, semen, pre-seminal fluids, rectal fluids, vaginal fluids, and breast milk. Once inside the body, the virus targets CD4+ cells. Inside the CD4+ cell, the virus uses reverse transcriptase, an enzyme, to convert its RNA structure into DNA. Once it becomes DNA, it will attach itself to the DNA inside the cell nucleus and will begin to mass replicate.

Because of the high rate of error in the reverse transcriptase process, HIV is more prone to mutations, where a mutation happens approximately once per 2000 incorporated nucleotides. This high mutation rate leads to multiple variations of the virus that can be more resistant to the immune system and to antiviral drugs. However, with the emergence of gene editing tool called CRISPR/Cas9, HIV-1 replication can be stopped along with the removal of any other infected

cells by the process of deleting targeted regions of HIV-1 from the genome rendering the virus useless.

Procedure

Genome Sequencing

With the dataset provided by the Los Alamos National Security, over 3000 variations or genotypes have been sequences from multiple sources. Each genotype is classified by multiple attributes such as year, country, and year sequenced. As HIV is a worldwide infection, each genotype is divided by a subtype which have sequences that more comparable than to sequences in other subtypes. These subtypes represent a multitude of variations such as geographical location and virus lineage. So comparing genotypes from different subtypes is less useful than from those in the same grouping. I decided to only look at genotypes from subtype C and the country Ethiopia for a more focused and concise dataset, which were 26 distinct genotypes.

In order to analyze the dataset, a reference sequence needs to be determined to compare all the sequences. C.ET.1986.ETH2220.U46016 was chosen as the reference, as it was the oldest genotype sequenced. This allows for a better analysis of the mutations of the virus throughout a longer time period.

	subtype	country	year	accession
C.ET.2002.02ET_288.AY713417	C	ET	2002	AY713417
C.ET.2008.ET104.KU319528	C	ET	2008	KU319528
C.ET.2008.ET106.KU319529	C	ET	2008	KU319529
C.ET.2008.ET108.KU319530	C	ET	2008	KU319530
C.ET.2008.ET115.KU319531	C	ET	2008	KU319531
C.ET.2008.ET119.KU319532	C	ET	2008	KU319532
C.ET.2008.ET122.KU319533	C	ET	2008	KU319533
C.ET.2008.ET124.KU319534	C	ET	2008	KU319534
C.ET.2008.ET126.KU319535	C	ET	2008	KU319535
C.ET.2008.ET128.KU319536	C	ET	2008	KU319536
C.ET.2008.ET130.KU319537	C	ET	2008	KU319537
C.ET.2008.ET135.KU319538	C	ET	2008	KU319538
C.ET.2008.ET136.KU319539	C	ET	2008	KU319539
C.ET.2008.ET145.KU319540	C	ET	2008	KU319540
C.ET.2008.ET147.KU319541	C	ET	2008	KU319541
C.ET.2008.ET148.KU319542	C	ET	2008	KU319542
C.ET.2008.ET149.KU319543	C	ET	2008	KU319543

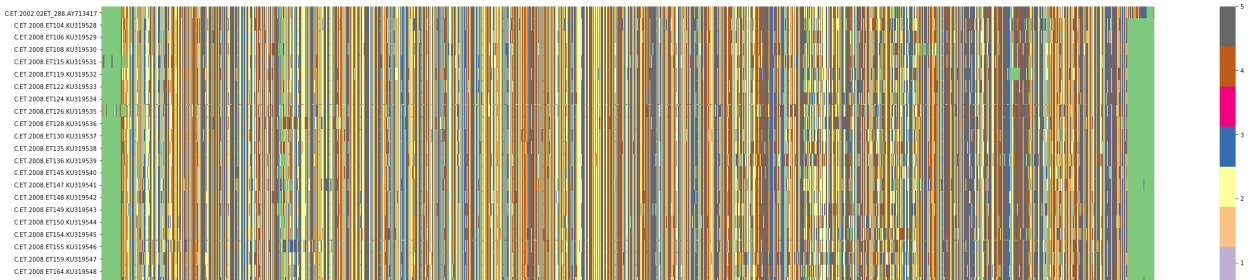
Sequence Alignment.

With the dataset including 26 distinct genotypes of HIV, the process to compare them is called sequence alignment. This procedure is done to compare two more biological sequences to identify regions of similarity. As the HIV virus is heavily mutable, insertions, deletions, and/or point mutations can occur at given sequence positions, so sequence alignment is necessary.

I developed my own sequence alignment tool in Python to align any target sequence to the reference sequence, C.ET.1986.ETH2220.U46016. The resulting sequence would have the same length of nucleotides to the reference. The alignment works by mapping out the corresponding nucleotides to the positions of the reference sequence and adding an (d) for a deletion, and an (i) for an insertion to the new aligned target sequence. The alignment would loop around for all 26 sequences and produce a list of new aligned sequences.

As all the new sequences had the same length, they could be put in a table by position and represented by a qualitative color map.

	C.ET.2002.02ET_288.AY713417	C.ET.2008.ET104.KU319528	C.ET.2008.ET106.KU319529
1	d	d	d
2	d	d	d
3	d	d	d
4	d	d	d
5	d	d	d
6	d	d	d
7	d	d	d
8	d	d	d
9	d	d	d
10	d	d	d
11	d	d	d
12	d	d	d
13	d	d	d
14	d	d	d
15	d	d	d



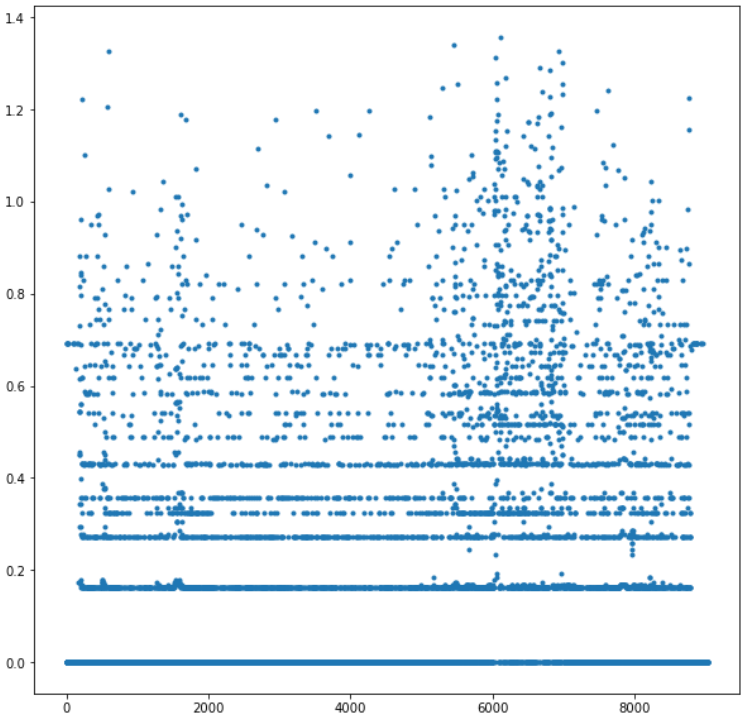
Analysis

Calculating Entropy

Shannon's entropy is a general idea that quantifies the uncertainty of probability distributions.

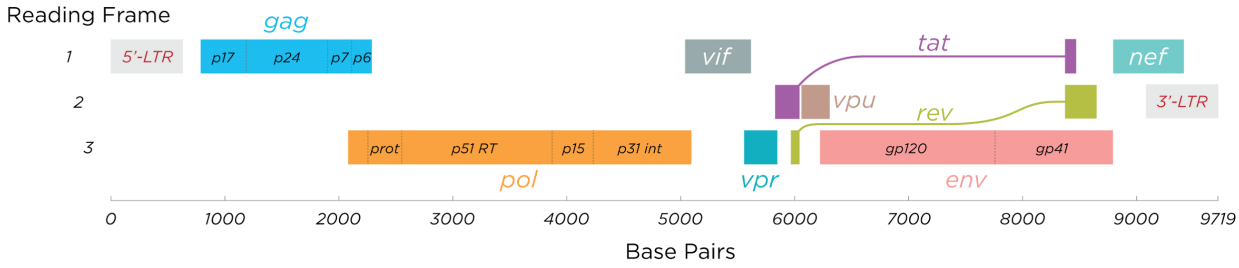
$$\lim_{p \to 0+} p \log(p) = 0.$$

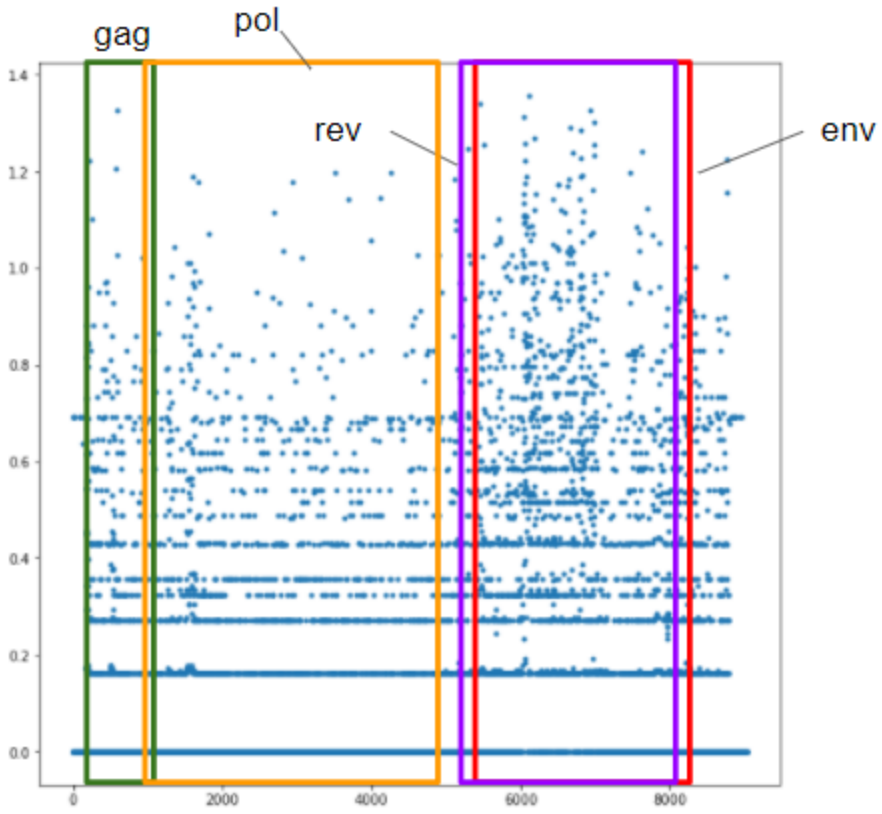
Specifically, this can applied to the distribution of nucleotides in sequence alignments. When entropy is calculated at every position, a graph can be created to measure the variability through each genotype. The values at each position is the measure of the lack of predictability. This is a quantifiable representation of the mutation rates throughout the virus' genome.



Interpreting Entropy

This entropy plot can be used then to compare the variability of the sequence positions to important regions of the virus. The entropy of each position can be compared to some biological function that can be represented throughout the genome sequence. So when regions show high rates of entropy or mutation those biological functions also mutate at high rates. Entropy can be then set onto the gene-map which consists of the proteins coded by the genome of HIV.





These four proteins that show high rates of mutation are also integral to the virus' functionality. The gag protein is associated with the viral assembly, the pol protein is associated with reverse transcriptase, and the env protein is outer envelope of the virus. And because the regions that code for these proteins are at higher rates of mutations these proteins will too, allowing the virus to change its function rapidly, making it hard to target.

Conclusion

According to the World Health Organization, over 35 million people are reported to be diagnosed with HIV with no reliable cure for HIV to be seen and since the start of its epidemic, more than 35 million have died.

New hope has been discovered recently with the use of CRISPR/Cas9 can target regions of HIV-1 from its genome rendering the virus obsolete. And based on this study, which has calculated the regions of most importance through being attributed to high rates of mutations, scientist can use these findings to target these regions with tools like CRISPR/Cas9. These approaches can a real cure for millions of people affected by HIV worldwide.

Acknowledgements

I would like to thank Dr. Luis Amaral for allowing me to join his lab over the summer. Also I would like to give thanks to Thomas Stoeger and Sophia Liu, who were both my mentors for this project and couldn't have done this without them. I'd also like to finally, most importantly, to my physics teacher, Dr. Mark Vondracek for allowing my love of research blossom to reality and giving me support at all times.

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

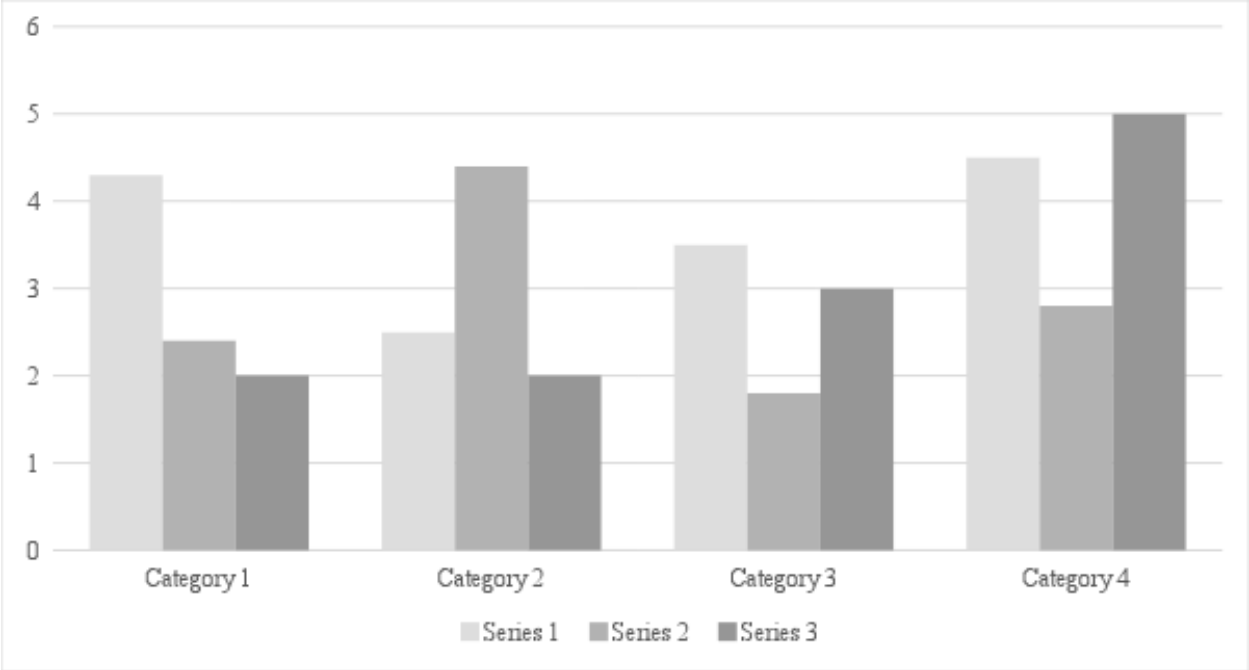


Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.