Exploring Viral Genomes

Tshepo Yane

Project Motivation

Virus genome analysis is a critical tool in the fight against viral diseases. By studying the structure, function, and evolution of viral genomes, we can develop better diagnostic tests, treatments, and vaccines. In addition, virus genome analysis can help us identify new viruses, and track the spread and evolution of existing viruses over time. This knowledge is essential for protecting public health and saving lives.

However, there are currently limited tools available to bioinformaticists who need to analyze genome sequences. Researchers and developers need tools to parse biological data formats, such as FASTA and GenBank files, as well as modules for common tasks such as clustering and sequence alignment. The best available tool in python is the Biopython module, but it is a large and complex library that can be difficult to learn and use. The Biopython project was formed in August 1999 as a collaboration to collect and produce open-source bioinformatics tools written in Python. Additionally, because Biopython is developed by volunteers, its quality and documentation can vary. Furthermore, because it is a third-party library, it may not always be up-to-date with the latest developments in the field of bioinformatics.

We need better tools to help us understand and combat viral diseases. By investing in the development of new, user-friendly bioinformatics tools, we can empower researchers and developers to make breakthroughs in the fight against viruses. Join us in supporting this important work and help protect public health around the world

Solution

Our project provides a central source for high-quality bioinformatics tools that researchers can use. We have created a custom Python class specifically designed for virus genome analysis, which gives us complete control over its design and implementation. This allows us to optimize the class for our specific needs and ensure that it is the best possible tool for the job. Unlike third-party libraries like Biopython, our class is tailored to the unique challenges of virus genome analysis. It can be used for genome analysis on FASTA files, a common file format for storing nucleotide or protein sequences.

Creating our own class from scratch can be time-consuming and may require significant investment in development and testing, but the benefits are clear. It is important to carefully weigh the costs and benefits before deciding whether to create your own class or use an

existing library like Biopython. By investing in the development of custom bioinformatics tools, we can empower researchers to make breakthroughs in the fight against viruses.

Method

In this project, we used the `SeqIO.read()`function from the Biopython library to load the viral genomic sequence data. The data was then instantiated as an instance of the custom-built `Genome` class. This class was designed to process and analyze a string sequence of nucleotide bases.

The `Genome` class has several methods for analyzing the nucleotide base composition of a genome. The base composition is calculated as the proportion of each nucleotide base (adenine, guanine, cytosine, and thymine) in the sequence. This information is important because it can provide insight into the length and potential function of the sequence. The class also has a method that visualizes the base composition of the genome on a bar plot.

The `Genome` class also has a method for calculating the guanine-cytosine (GC) content of a sequence. This is calculated by dividing the number of G and C nucleotides in the sequence by the total number of nucleotides and multiplying by 100 to express the result as a percentage. GC content is an important metric because it has been shown to be correlated with a virus' ability to evade the host immune system and its resistance to chemical and physical stresses.

Finally, the `Genome` class has methods for modeling transcription and translation. To model transcription, the coding strand of the genome is converted to RNA by replacing the thymine (T) nucleotides with uracil (U). The resulting mRNA transcript is then translated into a sequence of amino acids to form a polypeptide chain. The translation is performed using a dictionary of codons that map 3-nucleotide sequences to specific amino acids. The resulting polypeptide chains and nucleotide sequences can be visualized using color-coded strings, where each nucleotide or amino acid is represented by a different color. This representation makes it easier to compare and analyze the sequences

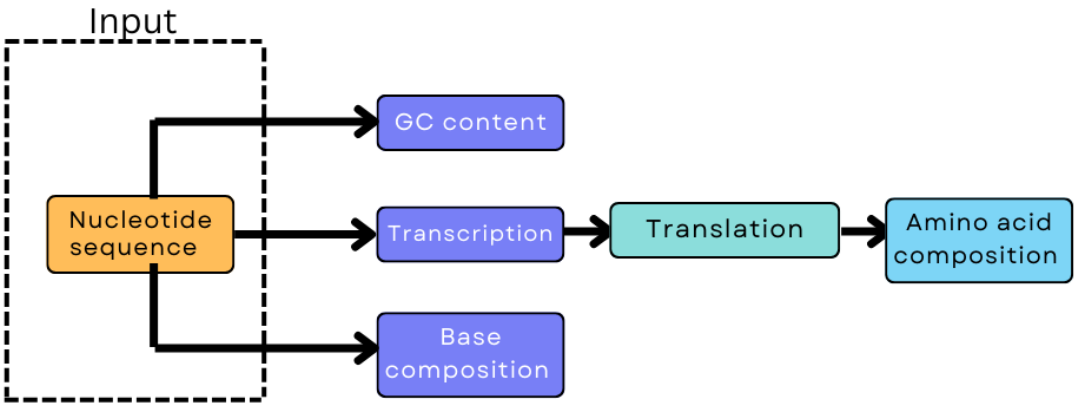

Figure 1: Shows a diagram of process of operations of the program

=

Results and Discussion

Through the use of our custom-built `Genome` class, we sought to investigate the similarities and differences between the genomes of SARS, MERS, Covid-19, and Ebola. Our analysis revealed that SARS, MERS, and Covid-19 possess a similar base composition, characterized by high levels of thymine (T) and low levels of cytosine (C). In contrast, the most prevalent base in the genome of Ebola was adenine (A), with guanine having the lowest proportion. This finding is consistent with the known phylogenetic relationships among these viruses, as SARS, MERS, and Covid-19 are all members of the coronavirus family, while Ebola belongs to a different viral family.

An unexpected result of our analysis was the observation that SARS, MERS, and Ebola had similar GC contents, ranging from 40.7% to 41.2%, while Covid-19 had the lowest GC content at 38%. This discrepancy may be correlated with the lower mortality rate of Covid-19 compared to the other viruses.

Upon transcription and translation, we also noted that Covid-19, SARS, and MERS have similar amino acid compositions, which further supports the notion that these viruses are closely related. In addition, we identified a conserved string of lysine amino acids at the end of the coronavirus genome, which is absent in the genome of Ebola.

Overall, our study has shed light on the genetic similarities and differences among these viral pathogens, and has provided insight into their evolutionary relationships. Furthermore, we were able to conduct this analysis using relatively low computational resources

amino	Covid-19	SARS	MERS	Ebola
L	8.890227	13.795885	17.045228	8.954041
S	8.127634	7.432432	8.487747	8.890650
T	6.813165	6.948366	5.618649	7.242472
C	6.371664	3.156515	5.040845	2.805071
F	5.950231	4.407019	4.512851	4.469097
R	5.599037	4.205325	5.180315	6.592710
V	5.498696	6.887858	6.485356	4.421553
Y	5.067229	3.438887	3.466826	3.248811
N	4.736103	4.215409	2.839211	5.007924
I	4.374875	5.707947	5.628611	7.321712
K	4.134056	5.062525	3.426977	5.483360
G	3.953442	4.417104	2.918908	4.041204
A	3.762793	5.798709	4.164176	3.787639
H	3.331327	3.005244	3.367205	2.218700
Q	3.261088	3.912868	3.357242	4.580032
P	2.929962	3.166599	3.825463	5.213946
D	2.909894	3.227108	1.205419	3.185420
E	2.709211	3.680920	1.564057	3.343899
W	2.638973	1.109318	2.022315	1.362916
M	1.173992	3.680920	2.978681	1.854200

Figure 2: Amino acid composition of the 4 viruses