Report - Analysis of SARS-COV-2 Sequences

Mayank Ghogale

September 2021

Institute of Chemical Technology
Guide: Prof Dr.Amiya Bhowmick
Prof Dr.Shamlan Reshamwala

Contents

1 Introduction

Analysis of the nucleotides of any virus can help us find various insights which can be used later on to study the virus more effectively.One such analysis has been done here on the sequences of SARS COV-2 the virus which brought the world to a stand still.The sequences are analyzed using various techniques of data science and the language used was Python.Various plots and figures have been deduced from analyzing the data which can be helpful in understanding the virus.

2 Data Extraction

The data was the sequences of nucleotides of SARS-COV-2 as extracted from the National Center for Biotechnology Information(NCBI)website. From about 3,50,000+ sequences of complete nucleotides present on the NCBI on the date of extraction(1st August,2021) around 15695 were taken for analysis. The sequences taken from analysis constituted sequences from all parts of the world except that from North-America and Europe. These sequences could not be included in the analysis due to lack of compute resources as analysing the sequences require a great deal of compute which was not present with the authors at the time of analysis. These sequences were then downloaded from the NCBI site and preprocessed as described in the following section before they could be used for data analysis.

3 Data Preprocessing

The sequences were first aligned with respect to the Wuhyan sequence, which was the first reported sequence of the SARS-COV-2 virus.This alignment was done using an online tool specifically meant to align the sequences of SARS-COV-2 as they were longer than the usual sequences.The tool used was MAFFT version 7(online version) and after alignment it gave us sequences of length 30216 and we had such 15695 sequences. In this 15695 sequences one sequence was the Wuhyan one and this sequence acted as the reference sequence for the rest of our data analysis. This data was first changed to Comma Separated Value or csv form using BioPython package and this csv file was read into a data frame using Pandas package and the further analysis was done taking this data farme as the basis.

4 Data Analysis

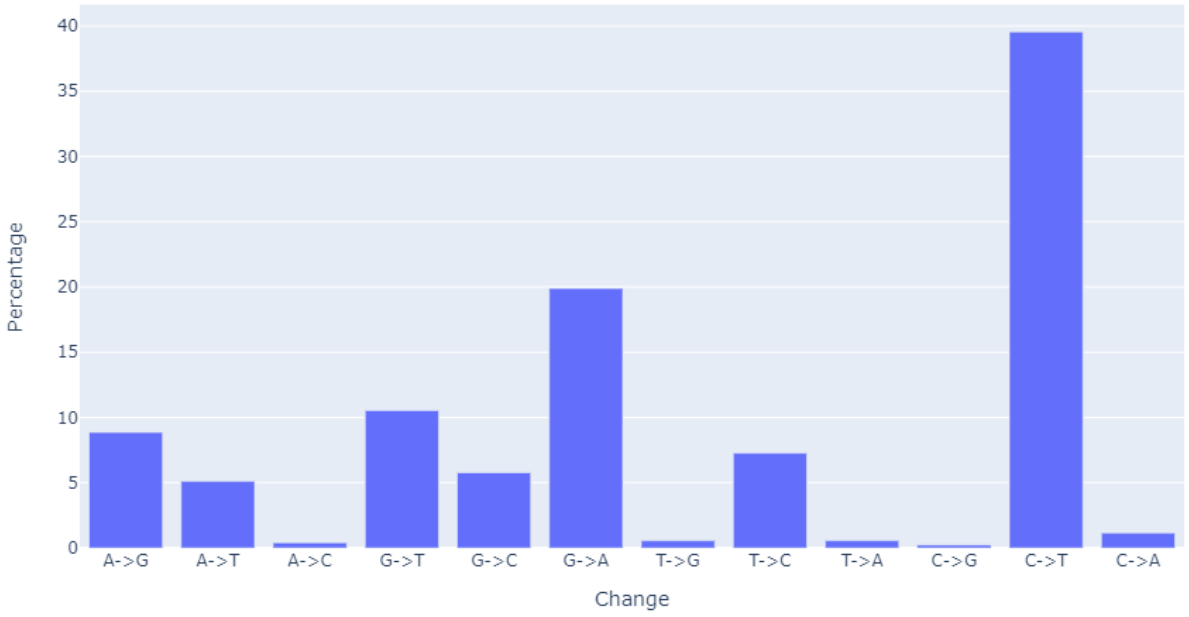
The data frame was analyzed thoroughly via a rudimentary code in Python to look for the changes at various positions of the nucleotides with respect to Wuhyan sequence and this data was then condensed into a data frame. Here we just are considering changes from and to A,G,T and C as those are the ones which are biologically relevant. This data frame had called columns of SeqNo,Position No,Character at Wuhyan and what character it changed to.This is depicted in the figure below.

	SeqNo	Position	Wuhyan	ChangedTo
0	2	305	c	t
0	2	740	c	t
0	2	3101	c	t
0	2	6386	a	g
0	2	6601	t	g
...	...	...	...	...
0	15690	8071	a	c
0	15690	9604	c	t
0	15693	8852	c	t
0	15693	28242	t	c
0	15693	29206	c	t

159487 rows × 4 columns

4.1 Analysis of total mutations

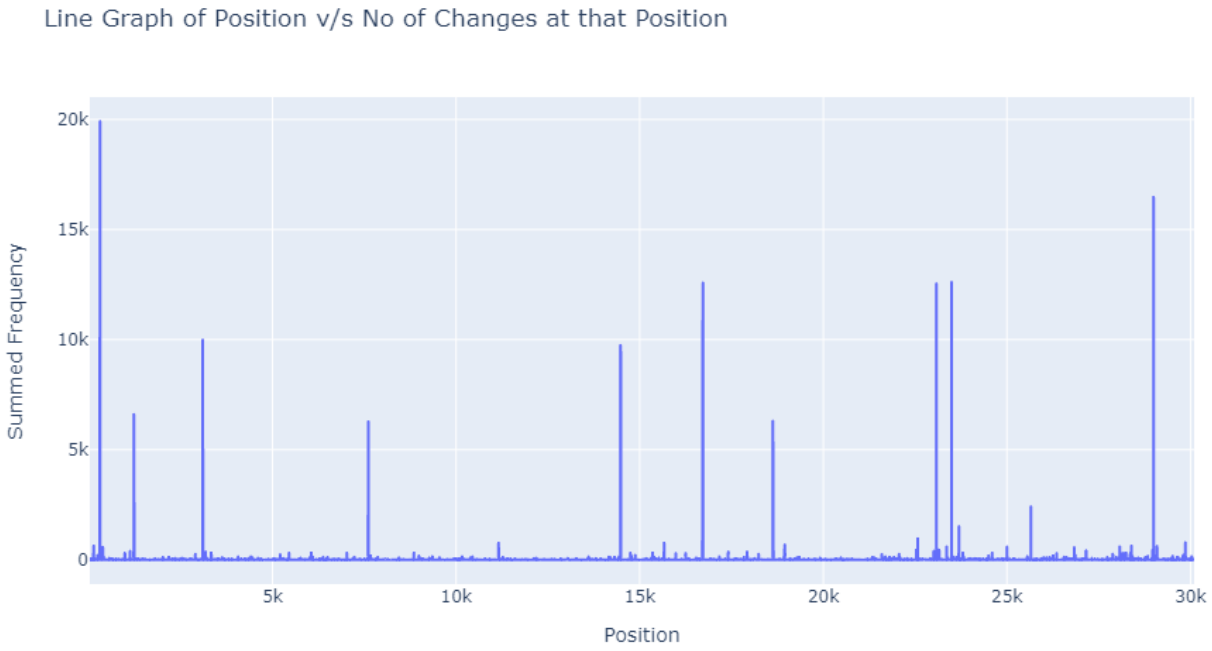
The data frame was then analyzed to calculate the frequencies of various transitions(A to G, A to T and so on) and the percentage of these frequencies were plotted with the help of Plotly Package in the form of a Bar Plot.The bar graph depicting the percent changes is given below.



It can be seen from the above bar graph that the most common type of change is from C to T and the next common being from G to A. These two changes as seen from graph constitute almost sixty percent of the total mutations and this information can be well used to obtain certain biological conclusions.

4.2 Analysis of mutations at a position

From the data frame, the number of total mutations which occurred through out various nucleotides at a particular position were calculated and this was plotted in form of a line plot which is shown below.



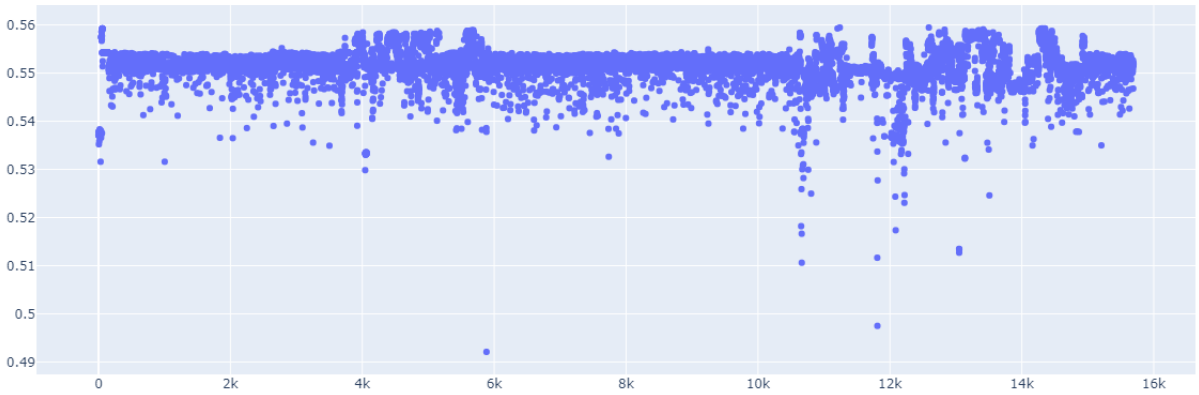
We can see from the plot that a position in the first few hundred positions and a position near the last position has seen a maximum number of mutations and we can easily from the data frame shown in the figure below find out things like what was the Wuhyan character at that position and what did it change to for that position. For clearly visibility the output does not show all the rows but any row can be easily accessed by just a single line command. We can safely ignore the first column of this data frame as it just has the default serial number given by Pandas package and it is of no use in our analysis.

	Position	Wuhyan	ChangedTo	Freq
0	25	a	t	8
1	25	a	g	8
2	25	a	c	8
3	26	t	a	8
4	26	t	c	8
...	...	...	...	...
6971	30095	a	c	2
6972	30096	a	g	2
6973	30097	a	c	1
6974	30098	a	t	1
6975	30099	a	t	1

6976 rows × 4 columns

4.3 Mutual Information between Sequences with respect to Wuhyan Sequence

The sequences were analyzed and the mutual information between them was found using functions from the Scikit-Learn package. For this the data frame used just had sequences in its rawest form and those were analyzed(including things like -,N,etc) To find the mutual information using the package stated above it was necessary to change the sequence into an array of zeros and ones which could be used for the calculation of the mutual information. To do this a small python script was written and the mutual information was then calculated with the help of the output of the python script using the package of stated above. The values of the mutual information have been summarized in a scatter plot as shown below.



It can be easily inferred from the plot that the mutual information with respect to Wuhyan is more or less the same and is around fifty six percent stating that the there is almost a fifty six percent correlation between any sequence which is not the reference Wuhyan sequence and the reference Wuhyan sequence.

5 Conclusion

Various nucleotide sequences of SARS-COV-2 Virus were analyzed using techniques of Data Analysis in Python and the results were displayed in form of tables and plots. This analysis can prove useful to various biologists who are working on similar topics as they can extract something meaningful from these charts and tables.