American Academic Scientific Research Journal for Engineering, Technology, and Sciences

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Prediction of Genomic Signature of Ngs Sequences and Comparative Drug-Likeness

Angela U. Makolo[a], Festus Segun Ajiboye[b*]

[a,b]*University of Ibadan Bioinformatics Group, Department of Computer Science, University of Ibadan, Ibadan 200132, Nigeria*

Department of Computer Science, Federal School of Statistics Ibadan

[a]*Email: aumakolo@gmail.com,* [b]*Email: ajiboyefestussegun@gmail.com*

Abstract

Developing a drug or particular immunotherapy medication for a worldwide epidemic illness caused by viruses (current pandemic) necessitates comprehensive evaluation and annotation of the metagenomic datasets to filter nucleotide sequences quickly and efficiently. Because of the homologs' origin of aligning sequences, space complexity, and time complexity of the analyzing system, traditional sequence alignment procedures are unsuccessful. This necessitates employing an alignment-free sequencing approach in this research that solves the foregoing issue. We suggest a distance function that compresses performance metrics for automatically identifying Short nucleotide sequences used by SARS coronavirus variants to identify critical features in genetic markers and genomic structure. This method provides easy recognition of data compressed by using a set of mathematical and computational tools in the study. We also show that by using our suggested technique to examine extremely short regions of nucleotide sequences, we can differentiate SAR-CoV-2 from SAR-CoV-1 viruses. Later, the Lipinski descriptor (rule of 5) was used to predict the drug-likeness of the target protein in SARS-CoV-2. A regression model using random forest was created to validate the machine learning model for computational analysis. This work was furthered by comparing the regressor model to other machine learning models using lezypredict, allowing scientists to swiftly and accurately identify and describe the SARS coronavirus strains.

Keywords: SAR-CoV-1;SAR-CoV-2; compression complexity; Lempel-Ziv;Lipinski descriptor; Regression model.

1. Introduction

An avalanche of DNA data is being generated by the Human Genome Project, as well as sequences in life sciences and pharmaceuticals. With so much data, software that can extract relevant information from raw DNA sequences is in high demand.

----

* Corresponding author.

It's critical to have tools that can recognize protein-coding regions and predict complete genes. WHO recently proclaimed a viral pandemic known as COV19, which shares the same disease symptoms as past active coronaviruses and is the century's third significant human coronavirus [2] The COVID-19 risk level was established by the World Medical Association evaluation for the geographical and international level to “Extremely high “on February 28, 2020.” The COVID-19 infection is suspected to be a member of the influenza virus family to Beta coronavirus lineage B (Sarbecovirus) based on scientifically confirmed facts and viral protein-based similarities [14]. According to phylogenetic analyses, RNA, and the whole nucleotide of the COVID-19 viral infection and similar coronaviruses, the COVID-19 viral infection is clearly linked to two bat SARS-like coronaviruses reported in China, SARS-COV-1: Urbanni and SAR-CoV-1: BJ01[20]. According to a phylogenetic investigation of whole-genome alignment and similarity plots, the COVID-19 virus shows the greater similarity to SAR-COV-1: Urbanni and SAR-COV-1: BJ01, both of which have been accepted as alignmentsequence phylogenetic distribution [11]. The current pandemic virus nucleotide show almost 99 percent nucleotide alignment and a paucity of variability with the variant, implying a shared homologues and origin, and pointing to the human strain's recent appearance.

It's uncertain if the COVID-19 viral originated independently as a unique variant capable of infecting people or as a result of recombination with previously described bat and unknown coronaviruses. According to the research done so far [14], the COVID-19 virus most likely originated in bats. [2], a Corresponding research, had made a lot of effort into using various machine learning techniques to quickly analyze multiple DNA sequences. The types of genome sequences and the numerous alignment techniques used for genomic analysis were recognized in the early study [5]. Also, they used MLDSP (guided algorithm and digital signal analysis) and a machine learningbased alignment-free algorithm were employed for genomic analysis in their work distance similarity, decision trees, genetic algorithms, clustering, and CNN networks was among the strategies utilized for the classification of new pathogens, according to their research [5]. The paper [8] suggested an Effort-To-Compress (ETC) estimate method to obtain information contained in a sequence data that is predicated on the concept of compression-complexity, as well as a compression-similarity average measure for accurate classification of

SAR coronavirus dwarf varieties in a set of viruses by only using brief gene sequences remnants. Also, Single nucleotide polymorphisms (SNPs), whole-genome sequence phylogeny, protein mutations, and microsatellites was among the genetic markers employed in their investigation [3]. Discovered that the normal hidden Markov model has an inherent protein mutation problem, so the method they advanced was Hidden Markov Model (AHMM) that show to be more prolific in identifying genetic markers than the previous model. Using genomic data, this publication [6] performed a phylogenetic tree-based approach to investigate the homologues connection of SARS-CoV2 with other beta coronaviruses [8]. To generate scores for probable functional sites, a statistical technique called Reference point logistics (RPL) regression was proposed.

2. Research Methodology

The dataset contains four different viruses, with one reference virus having between 27,608 and 29,751 copies. To differentiate between SAR-COV-1 and SAR-COV-2, 300 short reads were picked at random from each virus. The K-means clustering algorithm will be used to map the selected DNA base primary sequence into 0-1 sequences using the three metrics described below [24].

Table 2.1: various methods for categorizing DNA sequences into various sets of 0-1 sequences.

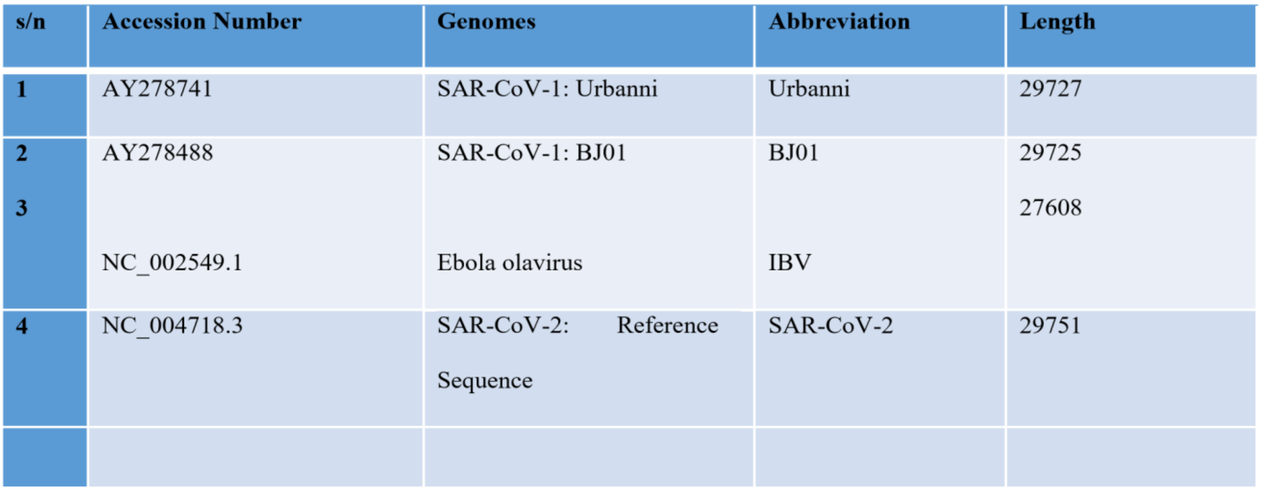
Seq No	Steps based on	sequence arranged to 0 sets mapped to 1 sets
1	Molecular weight(MW)	{A,G}set to 0} Pyrimidine {{C, T}set to 1}
2	Carboxylic acid	Amino{{A,C}set 0} Keto-{{G,T}set to1}
3	Strength of hydrogen electron	{A,T set 0} Strong h-b electron {G,C set 1}

The research method is divided into seven (7) stages which are:

2.1 Analyzed the Primary

We first analyzed the primary DNA of the proposed sequence which are the under listed virus are: SAR coronavirus Urbanni (AY278741.1), SAR coronavirus BJ01 (AY278488.2) and Ebola olavirus (NC_002549.1) SAR-CoV-2: Reference genome (NC_004718.3). The initial two viruses belong to the SAR-CoV-1 variant. Then, nucleotide data was collected from Genbank database,

Table 2.2: DNA Sequences and Length.



s/n	Accession Number	Genomes	Abbreviation	Length
1	AY278741	SAR-CoV-1: Urbanni	Urbanni	29727
2	AY278488	SAR-CoV-1: BJ01	BJ01	29725
3	NC_002549.1	Ebola olavirus	IBV	27608
4	NC_004718.3	SAR-CoV-2: Reference Sequence	SAR-CoV-2	29751

Before we can evaluate complexity levels, we must first convert primary sequences to 0-1 in our study.

We looked at three distinct ways for categorizing genome sequences [21] and mapping to the numbers 0 and 1 depending on their: • Molecular mass (MW)

- Carboxylic acid
- Hydrogen bond electron

2.2 *Complexity Measures for Lempel-Ziv (LZ)*

We employ Lempel-Ziv [1] compression measures to assess the distribution of short-length regions of nucleotide map sequences. Lempel–Ziv compression (LZ) is a common and universally used metric for determining how many steps the the input pattern must be compressed to an uniform length in the Non-Sequential Recursive Pair Substitution machine leaning (or a DNA of zero entropy). The compressibility of an input sequence is estimated using the complexity measure.

2.3 Measure and Identification of Distance

The concatenations approach was then used to determine the compressed sequences identification distance using a computed compression complexity measure (LZ). We had three virus genome sequences (W1, W2, and W3). Concatenation is used to create new sequences W1W2 and W2W1. The complexity measurements LZ(W1), LZ(W2), LZ(W1W2), and LZ(W1W2) are then computed (W2W1). In accordance with [7]'s methodology. The mean of the similar difference among the compression values of the two concatenated sequences W1W2 and W2W1 is proposed as a distance measure.

They are mathematically described as:

$$dLZ(W1; W2) = \frac{(LZ(W1W2)\ LZ(W1)) + (LZ(W2W1)\ \ LZ(W2))}{2} \qquad (2.1)$$

$$dLZ(W1:W3) = \frac{(LZ(W1W3) - LZ(W1)) + (LZ(W3W1)-LZ(W3).}{2} \qquad (2.2)$$

$$dLZ(W2:W3) = \frac{(LZ(W2W3) - LZ(W2)) + (LZ(W2W3)-LZ(W3).}{2} \qquad (2.3)$$

The minimum distance of the set d(W1; W2); d(W1; W3); d(W2; W3) is then determined. The distances in the mathematical equations 2.1, 2.2, and 2.3 will always be non-negative and symmetric, yet genomic complexity is not uniform. The time complexity of regions with more regular genes is lower than that of regions without a gene. For each of the three sets, the LZ distances are determined independently. The distance between the two sequences is calculated using the average values. The two SAR coronaviruses (Urbanni and BJ01) ought to have the smallest mean average (in complexities). In order to uniquely differentiate the SAR virus, the distance between Ebola and any of the SARS coronaviruses is compared to the distance between Ebola and any of the SARS coronaviruses (Urbanni and BJ01). If there are more than three sequences, this steps of display can readily be expanded. Perform Gen translation

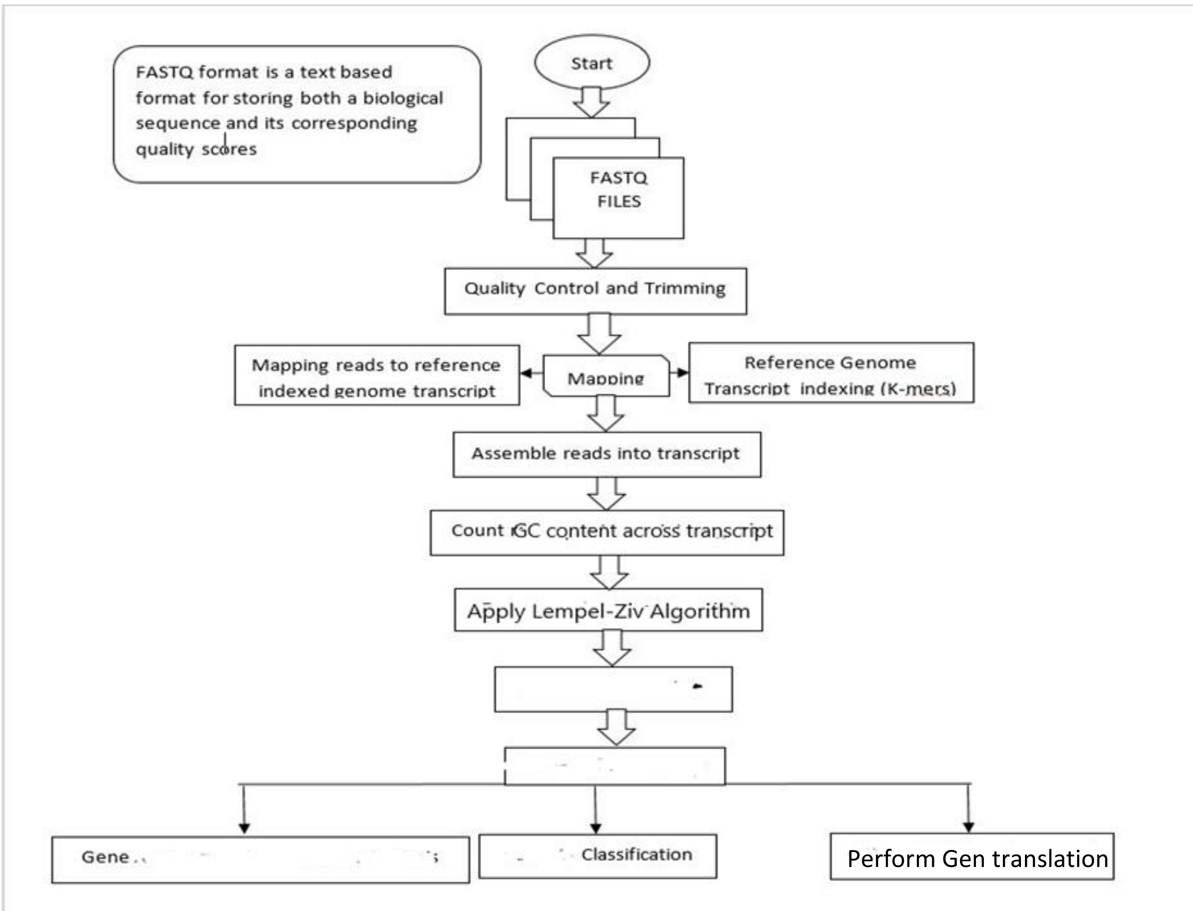

FASTQ format is a text based format for storing both a biological sequence and its corresponding quality scores
Start
FASTQ FILES
Quality Control and Trimming
Mapping reads to reference indexed genome transcript
Mapping
Reference Genome Transcript indexing (K-mers)
Assemble reads into transcript
Count GC content across transcript
Apply Lempel-Ziv Algorithm
Gene
Classification
Perform Gen translation

Figure 2.1: Comparative Analysis Procedure.

2.4 Analyze Short SARS-COV-1 VS. SARS-COV-2

We proceed with the research to analyze SAR-CoV-2 virus from SAR-CoV-1 virus after demonstrating the performance of LZ-based measures in effectively differentiating virus by examining extremely small portions of nucleotide sequences (the result presented in result and analysis).

The sequences below are used to accomplish this.

Table 2.3: Genbank accession numbers, names, abbreviations and lengths of coronaviruses.

S.NO	Accessions Number	Genome	Abbreviations	Lengths
1	AY278741	SAR-CoV-1: Urbanni	Urbanni	29728
2	AY278488	SAR-CoV-1: BJ01	BJ01	29726
3	NC_004718.3	SAR-CoV-2: Reference Sequence	SAR-CoV-2	29754

2.5 Machine Learning Implementation

The target amino acid is extracted from the detected SARS-COV-2, and we use Lipinski's Rule for analyzing the drug-likeness of the nucleotide to predict the drug-likeness of the extracted protein. According to Christopher Lipinski, such drug-likeness is predicated on Absorption, Distribution, Metabolism, and Excretion (ADME). He also looked at all orally active FDA-approved medications in order to come up

The following is rule-of-five:

Molecular Structure < 500 Dalton

- Hydrogen Donor < 5
- Hydrogen acceptor < 10
- Octanol-water partition (LogP) < 5

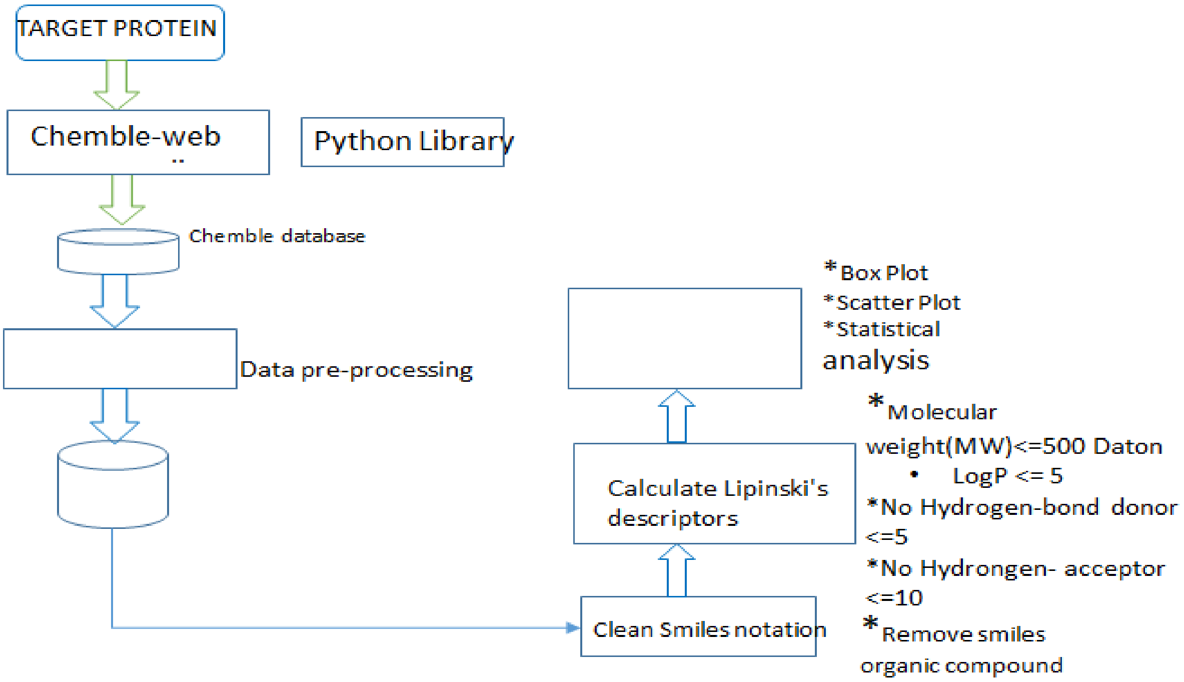

TARGET PROTEIN
Chemble-web
Python Library
Chemble database
Data pre-processing
*Box Plot
*Scatter Plot
*Statistical analysis
*Molecular weight(MW)<=500 Daton
• LogP <= 5
*No Hydrogen-bond donor <=5
*No Hydrongen- acceptor <=10
*Remove smiles organic compound
Calculate Lipinski's descriptors
Clean Smiles notation

Figure 2.2: Machine Learning Implementation Chart.

2.6 Step 6: Model Performance

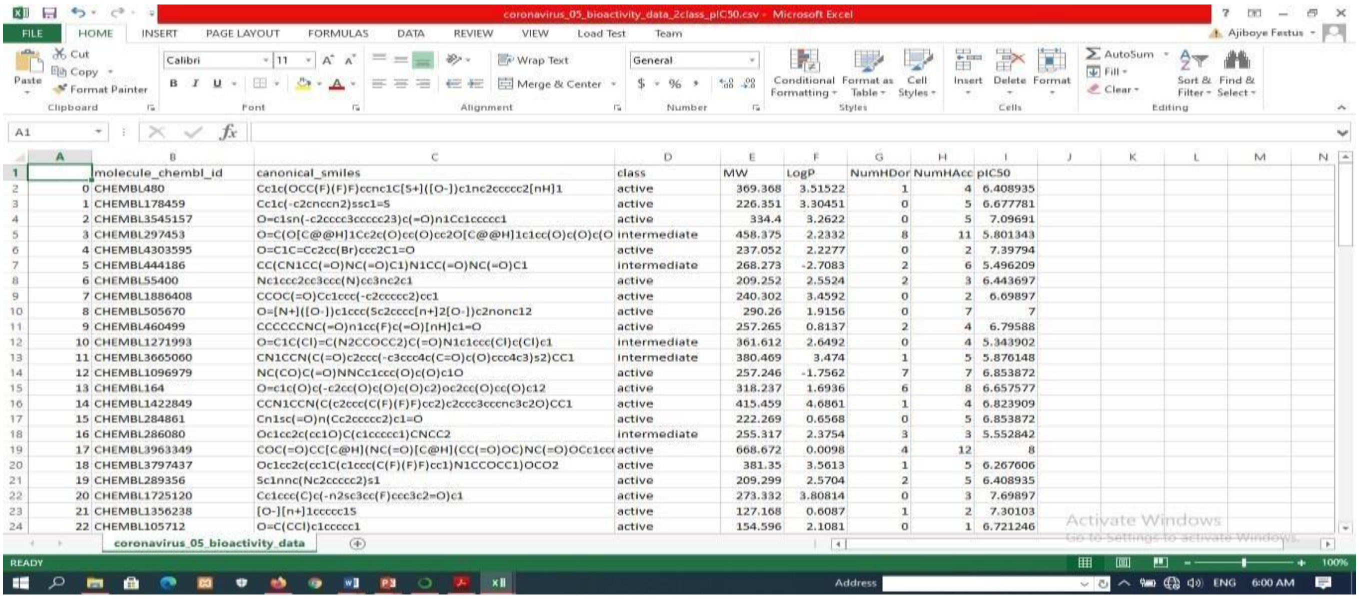
This section describes how to classify new observations using the data set provided. The procedures in developing the model for the section, on the other hand, are divided into four (4) phases: i. Data gathering/preprocessing ii. Model training iii Model construction iv Model comparison

2.2.1 PRE-PROCESSING OF DATA

Following the collection of data, the dataset is preprocessed to match the desired purpose, as stated in the previous heading.

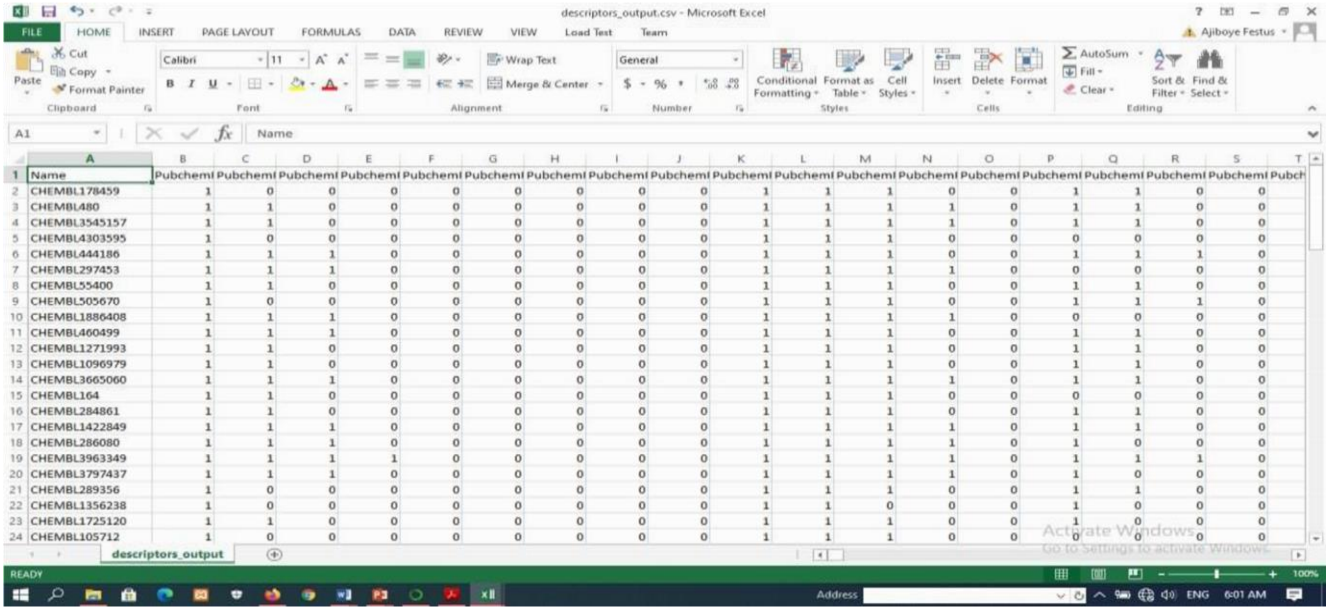
The proposed model was created with the help of the padel descriptors, which contain approximately 25400 characteristics.

However, to obtain an ideal model, a bash script was used to generate the descriptor features, which were then normalized.



	molecule_chembl_id	canonical_smiles	class	MW	LogP	NumHDor	NumHAcc	pIC50
0	CHEMBL480	Cc1c(OCC(F)(F)F)ccnc1C[S+]([O-])c1nc2ccccc2[nH]1	active	369.368	3.51522	1	4	6.408935
1	CHEMBL178459	Cc1c(-c2cnccn2)ssc1=S	active	226.351	3.30451	0	5	6.677781
2	CHEMBL3545157	O=c1sn(-c2cccc3ccccc23)c(=O)n1Cc1ccccc1	active	334.4	3.2622	0	5	7.09691
3	CHEMBL297453	O=C(O[C@@H]1Cc2c(O)cc(O)cc2O[C@@H]1c1cc(O)c(O)c(O	intermediate	458.375	2.2332	8	11	5.801343
4	CHEMBL4303595	O=C1C=Cc2cc(Br)ccc2C1=O	active	237.052	2.2277	0	2	7.39794
5	CHEMBL444186	CC(CN1CC(=O)NC(=O)C1)N1CC(=O)NC(=O)C1	intermediate	268.273	-2.7083	2	6	5.496209
6	CHEMBL55400	Nc1ccc2cc3ccc(N)cc3nc2c1	active	209.252	2.5524	2	3	6.443697
7	CHEMBL1886408	CCOC(=O)Cc1ccc(-c2ccccc2)cc1	active	240.302	3.4592	0	2	6.69897
8	CHEMBL505670	O=[N+]([O-])c1ccc(Sc2cccc[n+]2[O-])c2nonc12	active	290.26	1.9156	0	7	7
9	CHEMBL460499	CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	active	257.265	0.8137	2	4	6.79588
10	CHEMBL1271993	O=C1C(Cl)=C(N2CCOCC2)C(=O)N1c1ccc(Cl)c(Cl)c1	intermediate	361.612	2.6492	0	4	5.343902
11	CHEMBL3665060	CN1CCN(C(=O)c2ccc(-c3ccc4c(C=O)c(O)ccc4c3)s2)CC1	intermediate	380.469	3.474	1	5	5.876148
12	CHEMBL1096979	NC(CO)C(=O)NNCc1ccc(O)c(O)c1O	active	257.246	-1.7562	7	7	6.853872
13	CHEMBL164	O=c1c(O)c(-c2cc(O)c(O)c(O)c2)oc2cc(O)cc(O)c12	active	318.237	1.6936	6	8	6.657577
14	CHEMBL1422849	CCN1CCN(C(c2ccc(C(F)(F)F)cc2)c2ccc3cccnc3c2O)CC1	active	415.459	4.6861	1	4	6.823909
15	CHEMBL284861	Cn1sc(=O)n(Cc2ccccc2)c1=O	active	222.269	0.6568	0	5	6.853872
16	CHEMBL286080	Oc1cc2c(cc1O)C(c1ccccc1)CNCC2	intermediate	255.317	2.3754	3	3	5.552842
17	CHEMBL3963349	COC(=O)CC[C@H](NC(=O)[C@H](CC(=O)OC)NC(=O)OCc1ccc	active	668.672	0.0098	4	12	8
18	CHEMBL3797437	Oc1cc2c(cc1C(c1ccc(C(F)(F)F)cc1)N1CCOCC1)OCO2	active	381.35	3.5613	1	5	6.267606
19	CHEMBL289356	Sc1nnc(Nc2ccccc2)s1	active	209.299	2.5704	2	5	6.408935
20	CHEMBL1725120	Cc1ccc(C)c(-n2sc3cc(F)ccc3c2=O)c1	active	273.332	3.80814	0	3	7.69897
21	CHEMBL1356238	[O-][n+]1ccccc15	active	127.168	0.6087	1	2	7.30103
22	CHEMBL105712	O=C(CCl)c1ccccc1	active	154.596	2.1081	0	1	6.721246

Figure 2.3: Dataset before fingerprint descriptor.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Name	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	PubchemI	Pubch
2	CHEMBL178459	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
3	CHEMBL480	1	1	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	
4	CHEMBL3545157	1	1	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	
5	CHEMBL4303595	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	
6	CHEMBL444186	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	
7	CHEMBL297453	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	
8	CHEMBL55400	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
9	CHEMBL505670	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	
10	CHEMBL1886408	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	
11	CHEMBL460499	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
12	CHEMBL1271993	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
13	CHEMBL1096979	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
14	CHEMBL3665060	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	
15	CHEMBL164	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	
16	CHEMBL284861	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
17	CHEMBL1422849	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	
18	CHEMBL286080	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	0	0	0	
19	CHEMBL3963349	1	1	1	1	0	0	0	0	0	1	1	1	1	0	1	1	1	0	
20	CHEMBL3797437	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	0	0	0	
21	CHEMBL289356	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	
22	CHEMBL1356238	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	
23	CHEMBL1725120	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	
24	CHEMBL105712	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	

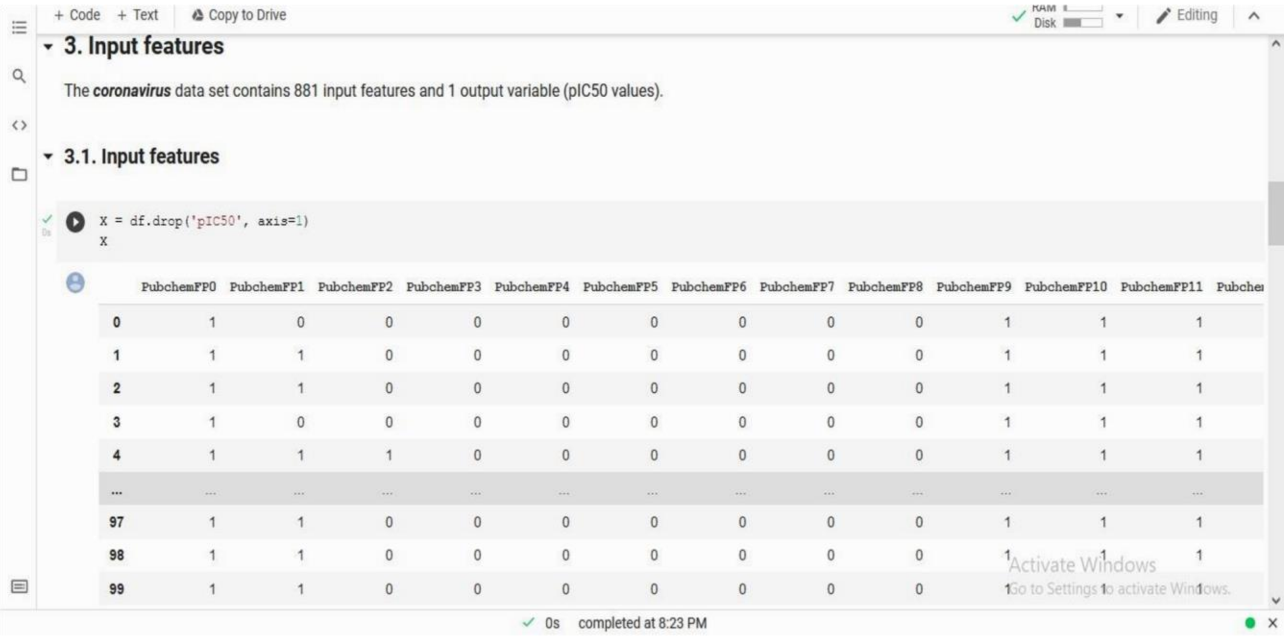
Figure 2.4: Dataset after fingerprint descriptor.

3 .Computational Drug-Likeness Regression Models with Random Forest

Using the fingerprint descriptor bioactivity data, we created an algorithm model. The random forest approach were adopted to create a regression model of SAR-COV-2 inhibitors.

3.1 Input features

The ***SAR-Cov-2*** data set contains 881 input features and 1 output variable (pIC50) values



3. Input features

The ***coronavirus*** data set contains 881 input features and 1 output variable (pIC50 values).

3.1. Input features

```
X = df.drop('pIC50', axis=1)
X
```

	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9	PubchemFP10	PubchemFP11	Pubcher
0	1	0	0	0	0	0	0	0	0	1	1	1	
1	1	1	0	0	0	0	0	0	0	1	1	1	
2	1	1	0	0	0	0	0	0	0	1	1	1	
3	1	0	0	0	0	0	0	0	0	1	1	1	
4	1	1	1	0	0	0	0	0	0	1	1	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
97	1	1	0	0	0	0	0	0	0	1	1	1	
98	1	1	0	0	0	0	0	0	0	1	1	1	
99	1	1	0	0	0	0	0	0	0	1	1	1	

✓ 0s completed at 8:23 PM

Figure 3.1: Input features.

4. Results and Analysis

4.1 SARS-COV-1 Identification

For LZ measurements, the average difference between the two SAR-CoV-1 virus, BJ01 and Urbanni is the greatest. At p 0:05, BJ01 and Ubanni signal statistical significance. The complexity of a genome is not uniform. The time complexity of regions with more regular genes is higher than that of regions without a gene.

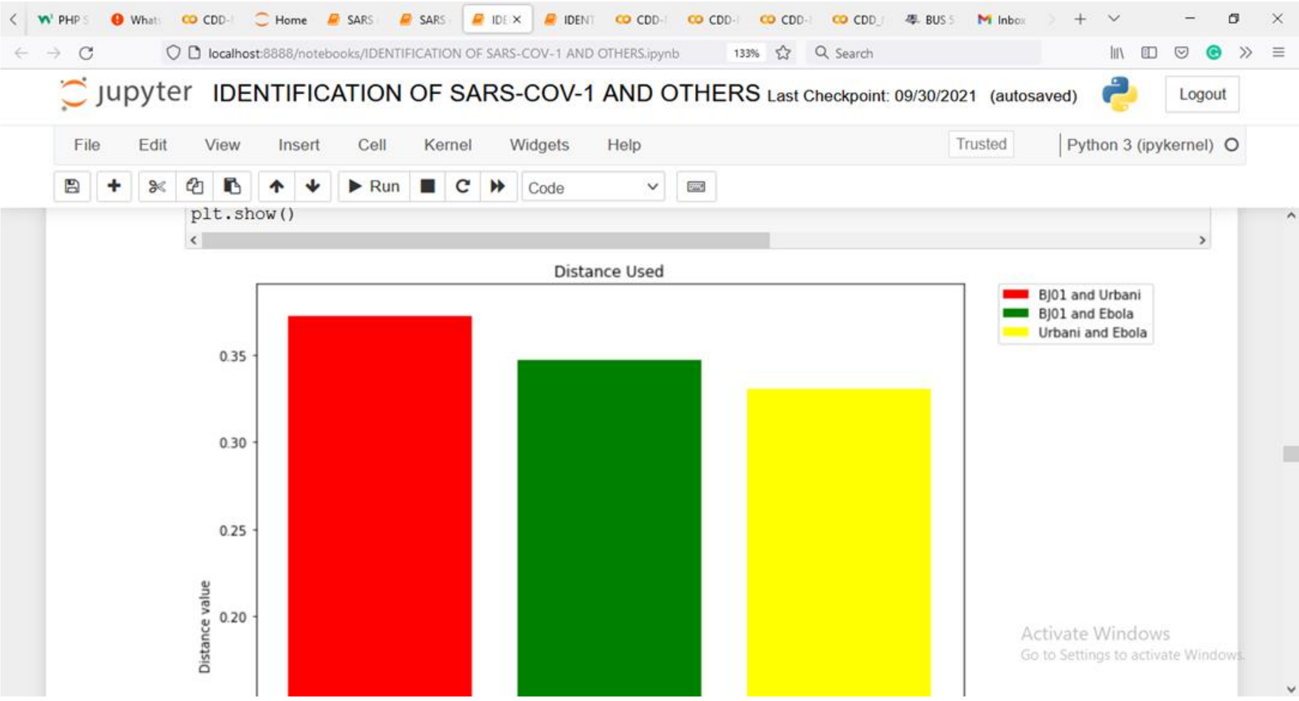
Table 4.1: Pairwise average measure for the selected viruses – Ebola, BJ01 and Urbanni. We discover mean difference between segments of lengths 300 DNA bases each. Those 300 was selected at random point of the whole sequences. The average difference among the two SAR-CoV-1 virus BJ01 and Urbanni is shown to be higher for LZ. [3]

Table 4.1: pairwise mean average for the three viruses – Ebola, BJ01 and Urbanni.

Pair of Virus	Distances: dLZ (μ ± O)
Ebola and BJ01	0.34685 ± 0.0386
Ebola and Urbanni	0.3302 ± 0.0351
BJ01 and Urbanni	0.37185 ± 0.0445*

***shows significance value at p < 0:05**

Figure 4.1 shows that outcome were statistically confirmed using 95 percent confidence interval plots. Based on the sample data, we can conclude that LZ can distinguish the SAR coronaviruses among the specified collection of viruses using just small contiguous segments comprises of 300 DNA bases picked at random among the whole sequence, with a total error rate of 0.05



```
plt.show()
```

localhost:8888/notebooks/IDENTIFICATION OF SARS-COV-1 AND OTHERS.ipynb
jupyter IDENTIFICATION OF SARS-COV-1 AND OTHERS Last Checkpoint: 09/30/2021 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
Distance Used
BJ01 and Urbani
BJ01 and Ebola
Urbani and Ebola
0.35
0.30
0.25
0.20
Distance value
Activate Windows
Go to Settings to activate Windows.

Figure 4.1: 95 percent confidence intervals for mean dLZ distance estimates for Ebola, Fig 4.1: 95% confidence interval for mean dLZ mean distance for the selected virus (pairwise) – Ebola, BJ01 together with Urbanni. The mean position among the two SAR-CoV-1 virus is proven to be higher for LZ.

4.2. SARS-CoV-2 Identification

We expand the research to distinguish SAR-CoV-2 virus from SAR-CoV-1 virus, having established the effectiveness of LZ-mean difference successfully differentiating viruses by evaluating very small regions of DNA sequences. For this experiment, we'll use the sequences (Table 3.1). Table 4.2 shows the results. For the selected viruses – SAR-CoV-2, BJ01, and Urbanni – pairwise mean distances were calculated. We calculated an average based on 300 short, contiguous segments. These 300 bases was picked at random sites across the whole sequences. For LZ measurements, the average position among the selected SAR-CoV-1 virus BJ01 and Urbanni is the greatest. This is a statistically significant outcome.

Table 4.2: pairwise average for LZ measures between the three viruses – SAR-CoV-1: BJ01, SAR-CoV-1: Urbanni and SAR-CoV-2.

Pair of Viruses	**Distances: dLZ** (μ ± Ʊ)
SAR-CoV-2 and BJ01	0.4031 ± 0.0524
SAR-CoV-2 and Urbanni	0.38375 ± 0.0474
BJ01 and Urbanni	0.37185 ± 0.0446*

***shows significance value at p < 0:05**

Table 4.2 shows pairwise distances for LZ [3] measures for the selected viruses – SAR-CoV-1: BJ01, SAR-CoV-1: Urbanni, also SAR-CoV-2 – calculation of the selected DNA sequences, randomly selected from the whole genome sequence.

Table 4.2 shows the mean pairwise distances (and variance).The sequence SAR-CoV-1 viral match (BJ01 and Urbanni), the LZ-pair mean position provided the least value (statistically significant). The pairwise conclusion of the outcome is deduced at p-value 0.05 interval graph as in Figure 4.2. with reference to the sample dataset, with a total error rate of 0.05, it can be deduced that LZ can discriminate between SAR-CoV-1 and SAR-CoV-2 virus adopting just small segment section comprising of 300 segments) of the whole sequence. The graph of mean average measure for LZ for the selected viruses for a one selectively segment of minimal length three hundred nucleotide shown at Figure 4.2 showing the outcome of variant length differentiating SAR-CoV-1 viruses from SARCoV-2 viruses, the graph of mean average measure for LZ of the selected virus of a pair selected lengths 300 sequences of the table

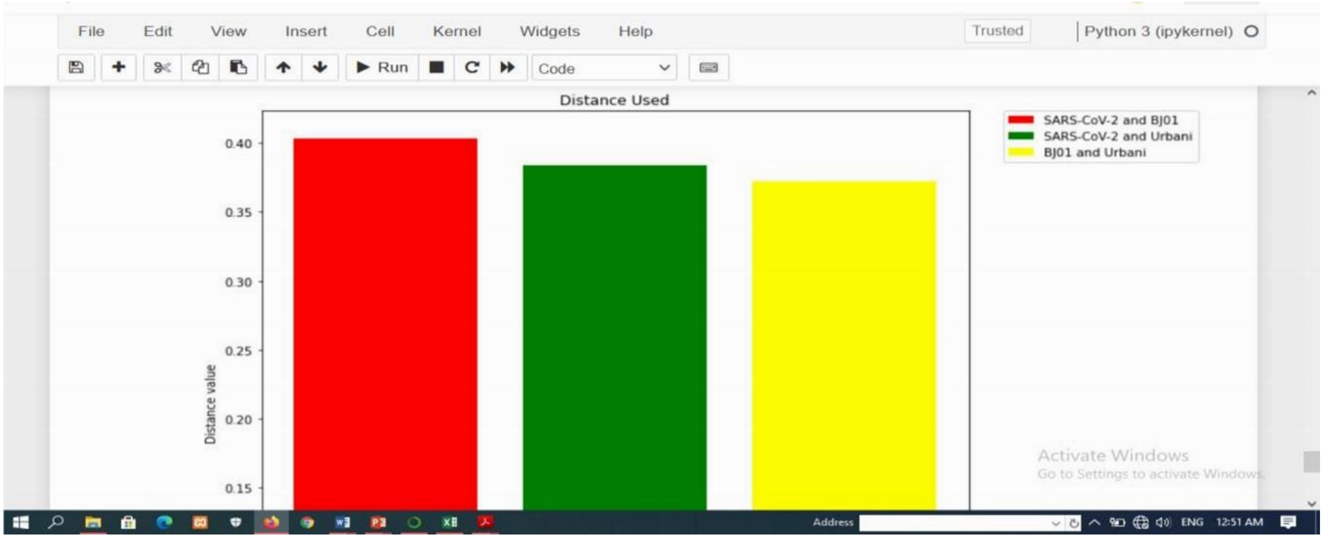
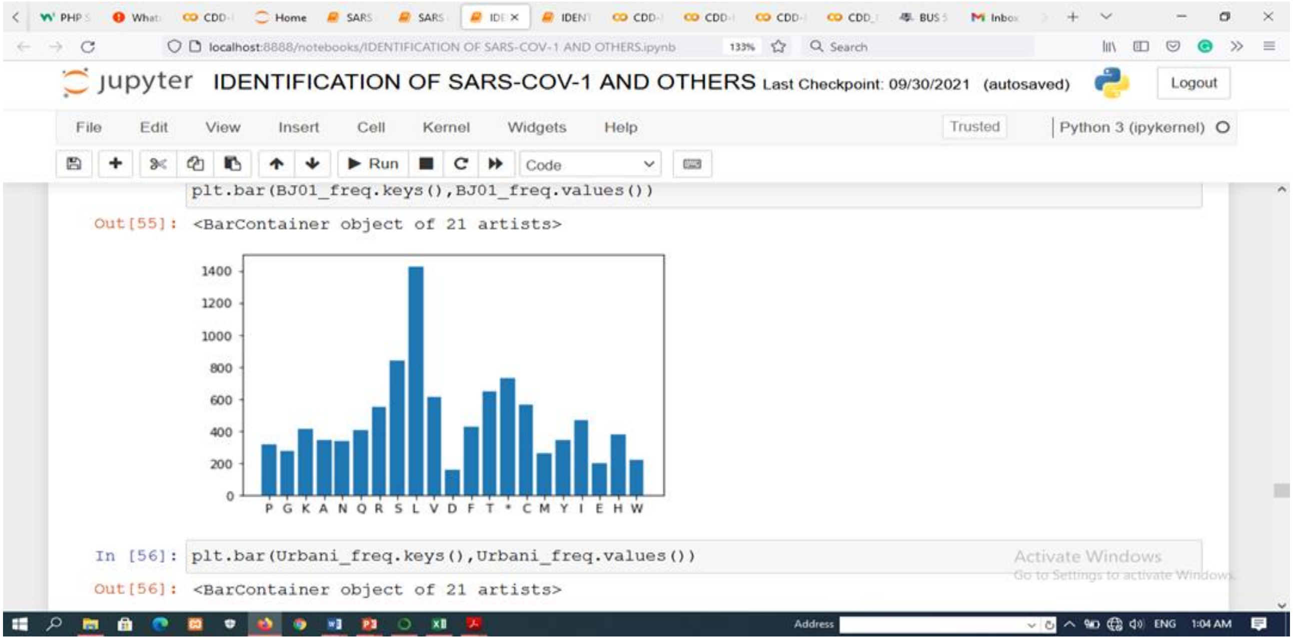


Figure 4.2: Distance Measure Of Sars-Cov-2 Vs Viruses.

The sequences (SARS-CoV-2 BJ01), and (SARS-CoV-2 Urbanni) and (BJ01 and Urbanni) calculated and plotted using LZ for the selected viruses for a pair selectively segment of minimal length three hundred nucleotide shown from the DNA genome. LZ accurately display the least average measure for those pair SARS-CoV-1 virus (BJ01 and Urbanni) among those three pair. This clearly indicate that the presence of SAR-COV-2 with other SAR-COV-1 make the average distance measure higher and clearly shows the distinguished features of SARS-COV-2 from SAR-COV-1 viral sequence.

4.3 Identification of Target Amino Acid

Having identify the SARS-CoV-2, the DNA sequences was converted to protein


jupyter IDENTIFICATION OF SARS-COV-1 AND OTHERS Last Checkpoint: 09/30/2021 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
1400
1200
1000
800
600
400
200
0
P G K A N Q R S L V D F T * C M Y I E H W
Activate Windows
Go to Settings to activate Windows.

```
        plt.bar(BJ01_freq.keys(),BJ01_freq.values())
Out[55]: <BarContainer object of 21 artists>

In [56]: plt.bar(Urbani_freq.keys(),Urbani_freq.values())
Out[56]: <BarContainer object of 21 artists>
```

Figure 4.3: Bar plot of highly express protein.

sequences in order to identify the target amino acid. The target amino acid is the highly express amino acid in the SAR-CoV-2 virus, which can either inhibit or allow modular solubility of oral vaccine. The bioactivity chemical compound that can either activate the highly express protein for possible oral solubility or to inhibit the highly express protein, using the Lipinski rule of five [13] will then be determine.

4.4 Prediction of Drug-Likeness

The bioactivity of the highly expressed protein is downloaded from chemble database. We performed data Pre-Processing on the downloaded bioactivity compound, all redundant chemical was removed and the curated data was divided into two major activity "active and intermediate". The chemical data are label as follows: The bioactivity dataset is in the IC50 unit. Chemical Compounds having bioactivities of less than 1000 nM will be determine to be active while bioactivities higher than 10,000 nM will be team to be inactive. As for those values among 1,000 and 10,000 nM will be tag intermediate

4.4.1 Lipinski descriptors

I Christopher Lipinski, a chemist at Pfizer, he developed a set of rule-of-five for concluding the drug-likeness of compounds. Such drug-likeness is based on the Absorption, Distribution, Metabolism and Excretion (ADME) that is also known as the pharmacokinetic profile**.**

The Lipinski's Rule-of-five state:

Ii Molecular weight < 500 Dalton

Ii Octanol-water partition coefficient (LogP) < 5

Iii Hydrogen acceptors < 10

IV Hydrogen donors < 5

Statistical analysis | Mann-Whitney U Test was later used to test the percentage significant. The rule of the significant factor is as follows:

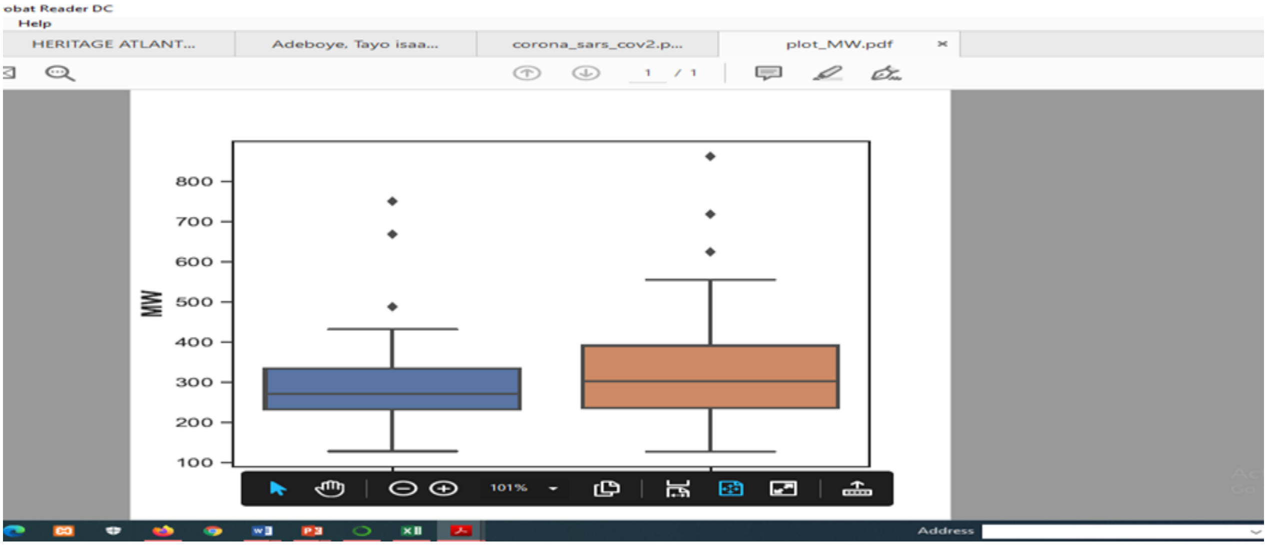


Figure 4.4.1: (Molecular weight).

Different

MW 1058779.0 2.072256e-57 0.05 distribution (reject H0)

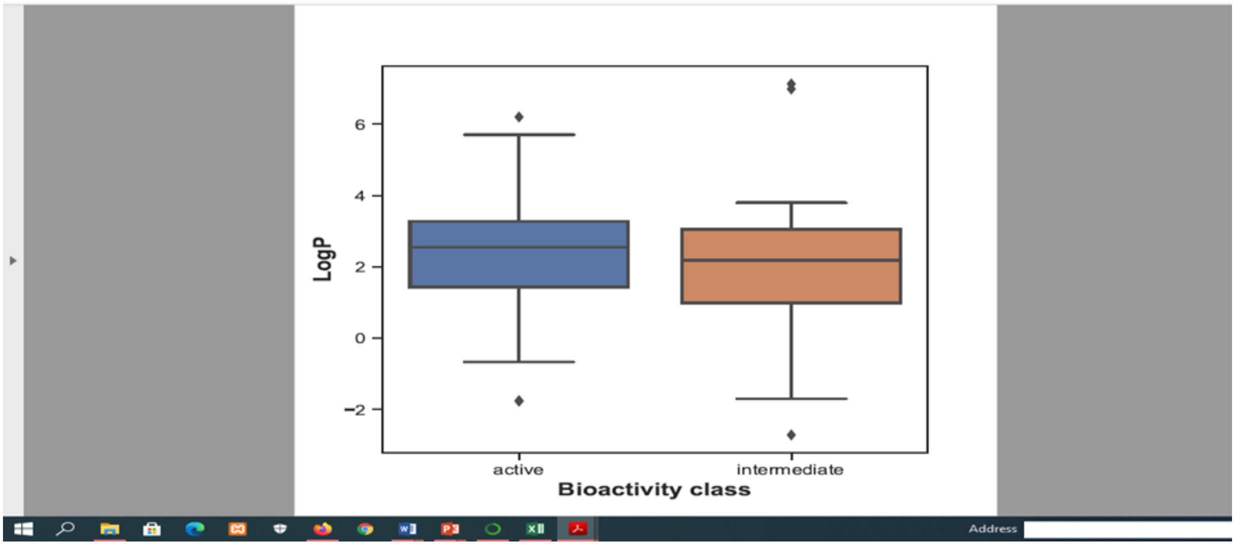


Figure 4.4:1i (Log P).

Lipinski rule 2: (Log P < 5)

Different Log P 10419900.0 2.318667e-61 0.05 distribution (reject H0)

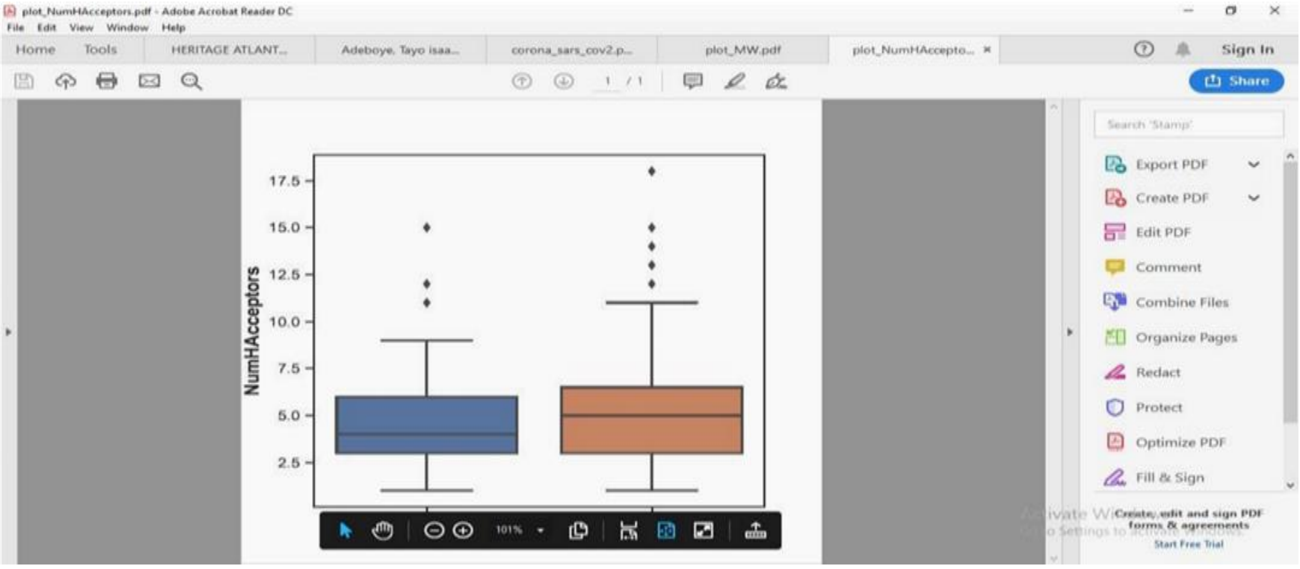

plot_NumHAcceptors.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home Tools HERITAGE ATLANT... Adeboye, Tayo isaa... corona_sars_cov2.p... plot_MW.pdf plot_NumHAccepto...
Sign In
Share
1 / 1
NumHAcceptors
17.5
15.0
12.5
10.0
7.5
5.0
2.5
101%
Search 'Stamp'
Export PDF
Create PDF
Edit PDF
Comment
Combine Files
Organize Pages
Redact
Protect
Optimize PDF
Fill & Sign
Create, edit and sign PDF forms & agreements
Start Free Trial

Figure 4.4.Iii: (NumHAcceptos).

Lipinski rule 4: (Hydrogen bond acceptor < 10)

1407572.5 0.000004 0.05

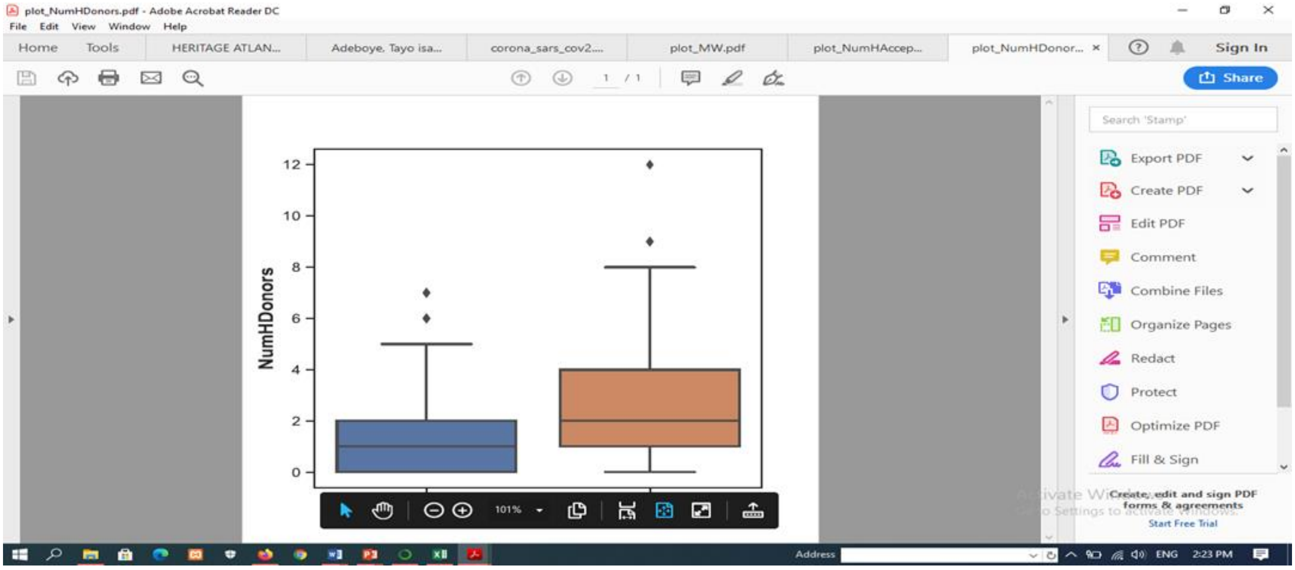

plot_NumHDonors.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home Tools HERITAGE ATLAN... Adeboye, Tayo isa... corona_sars_cov2.... plot_MW.pdf plot_NumHAccep... plot_NumHDonor...
Sign In
Share
1 / 1
NumHDonors
12
10
8
6
4
2
0
101%
Search 'Stamp'
Export PDF
Create PDF
Edit PDF
Comment
Combine Files
Organize Pages
Redact
Protect
Optimize PDF
Fill & Sign
Create, edit and sign PDF forms & agreements
Start Free Trial
Address
ENG
2:23 PM

Figure 4.4.Iii: (NumHDonors).

Lipinski rule 3: (Hydrogen bond donors < 5)

Different

NumHDonors 1361005.5 2.520096e-10 0.05 distribution (reject H0)

5. Conclusion

Compression-complexity metrics and grouping methods can be quite useful in the domain of NGS sequences. An entire genome sequence contains a number of intrinsic characteristics that aid in the detection of expressed genes and protein sequences, which can help with the development of oral vaccines while also allowing for proper taxonomic classification. The Lipinski description concept lays out the different scenarios for the emergence of any viral disease everywhere on the earth. This study's regressor algorithm were evaluated and equated to the performance of several algorithm approaches.

Our approach appears to be more accurate than support vector machines, decision trees, and other models in terms of accuracy, which is crucial for predicting drug-likeness genomic sequences. The regressor model shows that it is more accurate than other comparison models, and thus may be enhanced and used as a backend model not just in protein sequence oral vaccine discovery, but also in other fields of bioinformatics and computational biology where data integrity is a crucial problem. This research used a clustering and complexity technique to detect viral genomes, and then used the k-means and Lempel-Ziv (LZ) algorithm to create a regressor model that might provide biological interpretation, making drug-likeness prediction and vaccine development easier.

6. Funding

This research was funded by the National Institutes of Health, grant number U41HG006941; H3Africa Bioinformatics Network, H3ABioNet,.

Acknowledgements

The authors acknowledge the suggestions of Dr Temitayo Olagunju and Dr Taiwo Adigun in the initial review of the manuscript

7. Conflicts of Interest

The authors declare no conflict of interest. The funders or the H3Africa Initiative had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Abraham, L. & Jacob, Z. (2019). The difficulties of discrete sequencing is discussed. IEEE Transactions on Cognitive Science, 22(1):75–81

[2] Alagaili, AN., Briese, T., Mishra, N., Kapoor, V., Sameroff, SC., & deWit, E., et al.(2014). Middle east respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. MBio. 2014; 5. https://doi.org/10. 1128/mBio.00884-14

[3] Bin Li, Yi-Bing Li, and Hong-Bo He (2005). LZ Nucleotide sequence functionality position and its implementation in phylogenetic analysis restructuring. Genomics Bioinformatics & Proteomics,

3(4):206–212.

[4] DR,P., Bose, P. (2021). Comparative study of Sars, Mers, Bat-sars and Sars- cov-2. News medical Life sciences

[5] Gurjit, R., & Maximillian P. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel. International Journal of natural Science and Engineering Research Council of Canada.: 10:1371.

[6] *Hafiz., A.,* & *Farheen., R.* (2021). Comprehensive comparative genomic and microsatellite analysis of SARS, MERS, BATSARS, and COVID-19 coronaviruses. International Journal of natural Science and Engineering Research Council of Canada, Vol 10 issue 1002.

[7] Hasan, H., & Khalid, S,. (2003). A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 19(16):2122–2130, 2003.

[8] Karthi, B., Nithin, N. (2020). Compression-complexity measurements: Automatic identification of SARS coronavirus. International Journal of natural Science and Engineering Research Council of Canada.: Volume 3 issue 24.

[9] Vijayaragavan., S.P. Kumar., B. & Ajay., p. (2000). Prediction of genetic structure in eukayotic DNA using refrence point logistic regression and sequence alignment. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). Vol 16 Issue 5

[10] Mitchell, T,. (1999). Machine learning and data mining. *Communications of the* ACM, 42(11), 30-36.

[11] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22),:4673–80.

[12] Vinga, S. & Almeida, J. (2003). Alignment-free sequence comparison—a review. Bioinformatics; 19(4):513–23.

[13] Christopher, L,. (2002). Capture the Untapped Value of Therapeutics. Melior Pharmaceuticals. https://www.meliordiscovery.com/christopher-lipinski/

[14] Karthi, B. & Nithin, N. (2020). Compression-complexity measurements: Automatic identification of SARS coronavirus bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006007.

[15] Kumar, S., Stecher, G., Tamura, K,. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol; 33(7):1870–4.

[16] Lempel, A., Ziv, J. (1976). On the complexity of finite sequences. IEEE Transactions on information theory;22(1):75–81. LI, M. & Paul, M. (2014). Kolmogorov complexity and its applications. Algorithms and Complexity, 1:187. Liwei Liu, Dongbo Li, and Fenglan Bai. Application of a relative Lempel-Ziv complexity to comparing biological nucleotide. Letters in Chemical Physics, 530:107–112.

[17] Lu, H., Yang, L., Yan, K., Xue, Y. & Gao, Z. (2017). A cost-sensitive rotation forest algorithm for gene expression data classification. Neurocomputing; 228:270–6

[18] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B. & Wu, H., et al (2020). Genomic characterization and epidemiology of novel coronavirus: implications for virus origins and receptor binding. Lancet;. https://doi.org/10.1016/S0140-6736(20)30251-8 Luk, H., Li, X., Fung, J., Lau, S. & Woo, P. (2019). Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and

Evolution; 71: 21–30. https://doi.org/10.1016/j.meegid.

[19] Maguire, P., Moser, P., Maguire, R. & Griffith, V. (2014). Is it possible to program consciousness? Using algorithmic information theory to quantify integrated data. arXiv preprint arXiv:14050126.

[20] Mitchell, T,. (1999). Machine learning and data mining. *Communications of the* ACM, 42(11), 30-36. Ming, Li., Jonathan, H., Badger, X., Chen., Sam, K., Paul, K., & Haoyong, Z. (2001). The application of an information- based sequencing distance to the phylogeny of the entire mitochondrial genome. Bioinformatics, 17(2):149–154.