Scientific Research Publishing

Computational Molecular Bioscience, 2022, 12, 123-143
https://www.scirp.org/journal/cmb
ISSN Online: 2165-3453
ISSN Print: 2165-3445

Evolutionary Relationship of Protein Sequences of SARS-CoV-2 and Other Viruses through Chaos Game Representation

Matthew D. Hill, Kevin E. Simmons, Dipendra C. Sengupta*

Department of Mathematics, Computer Science, and Engineering Technology, Elizabeth City State University, Elizabeth City, NC, USA
Email: *dcsengupta@ecsu.edu

How to cite this paper: Hill, M.D., Simmons, K.E. and Sengupta, D.C. (2022) Evolutionary Relationship of Protein Sequences of SARS-CoV-2 and Other Viruses through Chaos Game Representation. *Computational Molecular Bioscience*, **12**, 123-143.
https://doi.org/10.4236/cmb.2022.123008

Received: June 25, 2022
Accepted: September 27, 2022
Published: September 30, 2022

Abstract

Comparison between different biological sequences is a key step in bioinformatics when analyzing similarities of sequences and phylogenetic relationships. A method of graphically representing biological sequences known as Chaos Game Representation (CGR) has achieved many applications in the studies of bioinformatics. The key issue in the application of CGR is to extract as many useful features as possible from CGR. Initially, CGR was applied to DNA sequences, but in this paper, a CGR-based approach is used to extract suitable features for comparing protein sequences of SARS-CoV-2 and other viruses. For this aim, several viral protein sequences from 12 groups are considered and CGR centroid, amino acid frequency, compounded frequency, Shannon entropy, and Kullback-Lieber Discrimination Information are applied to find the inter-relationship among the sequences. The experimental results demonstrate the potential strengths of CGR-based method for examining the evolutionary relationship of protein sequences. Our method is powerful for extracting effective features from protein sequences, and therefore important in classifying proteins and inferring the phylogeny of viruses.

Keywords

Chaos Game Representation (CGR), Protein, Multi-Dimensional Scaling (MDS)

1. Introduction

Proteins are complex molecules that play a critical role in several functions of the body as well as the structure of tissue and organs. They are comprised of amino acids which are connected in long chains ranging from a few hundred

to several thousand depending on the protein. These chains of amino acids determine the structure and function of a protein, which include the transport and storage of structural components, and enzymes. By studying the structure and function of proteins, we can hurdle some of the obstacles in understanding the evolutionary relationships of organisms. The 20 amino acids are Alanine (A), Arginine (R), Asparagine (N), Aspartic Acid (D), Cysteine (C), Glutamic acid (E), Glutamine(Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y), and Valine (V) [1]. Each amino acid has certain physical and chemical properties which distinguish it from others. In general, the biological function of a protein is determined by its 3-dimensional structure which is dependent on the linear sequence of amino acids. Rigden [2] presented that one of the fundamental principles of molecular biology is that proteins having similar sequences possess similar functions. This leads to difficulty when comparing closely and distantly related sequences. Similarity analysis of protein sequences plays an important role in protein sequence studies, e.g. the prediction or classification of protein structures and functions. In recent years, many numerical representation methods have been proposed and then applied in protein classification.

Apart from representing biological sequences into numerical expression directly, many other numerical representations are constructed by first giving the sequence a graphical representation and then studying the image numerically [3]. Chaos game representation (CGR) was originally applied to bioinformatics as an image representation of DNA sequences by Jefferey in 1990 [4]. The four nucleotides {A, T, G, C} were put on 4 vertices of the unit square, and every DNA sequence was mapped to a series of points inside the unit square in 2-dimensional space. Being capable of discovering the inner pattern of gene sequences, CGR has been widely used in the investigation of DNA sequences in [5]-[11]. Encouraged by the CGR of DNA sequences, the CGR of protein sequences has also been extensively studied by many researchers. While DNAs are composed of four kinds of nucleotides, proteins are made up of twenty kinds of amino acids. Thus, it remains to decide the distribution of the 20 amino acids when promoting CGR to the image representation of proteins.

Fiser [12] was one of the first to find a method to improve such techniques by creating a 20-sided polygon with each vertex representing one of the 20 amino acids. Another representation of the 20 amino acids was applied by Randic [13] in which the CGR exists within the unit circle. This approach ordered the amino acids alphabetically in comparison to organization based on their physiochemical properties. The properties of the amino acids serve as vital information for the characterization of protein sequences and this was noted by Randic. Considering the limitation that a 20-vertex CGR cannot be used to demonstrate the similarity of protein sequences with conservative substitution, Basu [14] proposed a 12-vertex CGR, with each vertex of a regular 12-sided polygon representing an amino acid with its conservative substitutions. The number of the vertices in CGR

was then reduced to four [15] [16], with each vertex of the square representing one of the four groups of amino acids, that is, the non-polar, uncharged polar, negative polar, and positive polar groups. The reduction in the vertices of CGR images can help represent the similarity in protein sequences.

Up to now, CGR method has achieved many applications in the studies of bioinformatics. The key issue in the application of CGR is to extract as many useful features as possible from CGR and several studies showed that those extracted features play important roles in protein studies [17]-[23]. One of the most used feature extraction methods is the so-called FCGR, in which the CGR image is split into small grids and the frequencies of points falling into each grid are taken as the feature of the corresponding protein sequence. In our previous work [24], we used FCGR to study the similarity of coronavirus sequences. While FCGR has been used mainly for coronavirus genome sequence encoding and classification, we modified it in this study to work also for protein sequences.

For this study, HTLV 1, HIV 1, HIV 2, Ebola, Dengue, Middle Eastern Respiratory Syndrome (MERS), Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) were used for protein sequence comparison. SARS-CoV-2 has been detrimental to the human population over the past year. At the time of this report, more than 179 million people have contracted the virus and over 3.8 million of those have been fatal. The first pathogenic novel coronavirus, discovered in 2003 and named SARS-CoV, caused SARS, serious and atypical pneumonia. The second, MERS-CoV, emerged a decade later in the Middle East and caused a similar respiratory ailment called Middle East respiratory syndrome (MERS). Since its identification, 2494 cases of MERS-CoV infection and nearly 900 deaths have been documented. The SARS-CoV epidemic proved larger but less deadly, with approximately 8000 cases and nearly 800 deaths. There are other four coronaviruses that cause colds in humans-known as HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1 [24]. SARS-CoV-2 is the third pathogenic novel coronavirus. Identifying ways to better understand such viruses is of grave importance to the human population. Such major outbreaks demand classification and origin of the virus genome sequence, for planning, containment, and treatment. Motivated by the above need, we report a method combining with CGR to perform clustering analysis and create a phylogenetic tree based on it.

For this report, CGR is used for the identification of several hundred protein sequences into their respective viral groups through feature extraction. These features include CGR centroid, amino acid frequency, compounded frequency, Shannon entropy, and Kullback-Lieber Discrimination Information. Due to the scale independence of CGR, smaller components of the CGR graph can be used to help explain the bigger picture. This points to the potential of extracting smaller features of the graph and using them to better explain the protein sequence as a whole. After the application of our proposed method, we apply multidimension-

al scaling (MDS) to the data. With this 2D and 3D projections of the data can be obtained for clustering analysis. Kruskal [25] first introduced this method of information visualization which takes the distance matrices computed from our methods as input. In turn, a representation of each viral sequence is created in euclidean space with corresponding distances between sequences that are equivalent to their distance given in the matrix. Therefore, similar viral sequences should be relatively close in this representation which has been previously shown for DNA sequences [26] [27].

2. Methods

In this section, we describe the dataset used for our experiments, then discuss the proteins version of CGR, give an overview of the three main steps of our experiments, and conclude with a description of features that we considered.

2.1. Dataset

Data acquisition: All protein sequences were downloaded in FASTA format from the database for our analysis: NCBI (https://www.ncbi.nlm.nih.gov/). The data sets shown in **Tables 1-6** are the accession numbers of 510 viral strains in 12 groups that we used for our experiments.

The HIV_1 group consisted of Gag-pol, Gag-pol poly, Gag-pol fusion, and Gag-pol fusion poly proteins. For HIV_2, pol, pol poly, Gag-pol poly, and Gag-pol fusion poly proteins, the Dengue group consisted of only polyproteins. HTLV group contained pol, polymerase, Gag/pol precursor, Gag-pro-pol poly, and Gag-protease proteins. SARS_CoV and SARS_CoV-2 encompassed six groups, two for ORF1a polyprotein, and two for ORF1ab polyprotein, one spike glycoprotein and one surface glycoprotein. MERS group contained 1a poly, 1ab poly, ORF1a, ORF1ab, OR1ab poly, and replicase poly proteins. Lastly, the Ebola group consisted of RNA-dependent RNA polymerase, RNA-directed RNA polymerase, polymerase, and L proteins.

2.2. CGR of Proteins

CGR is an algorithm that uses iterations in order to generate a pattern by utilizing the nucleotides in DNA or amino acids in protein sequences. CGR assigns a coordinate value to each alphabet in a sequence and hence a characteristic visual pattern is generated for each sequence. In the case of a DNA sequences, CGR assigns each of the four possible nucleotides A, T, G and C to one of the four vertices of a square. In our study, we used protein sequences; the 20 amino acids were divided into 4 groups, and each of these groups (designated A, B, C and D) was assigned to one of the four vertices of the square. We used groups based on amino acid residue chemical properties (charge and polarity): A = D, E (negatively charged); B = K, R, H (positively charged); C = S, T, N, C, Y, Q (neutral/polar); D = G, A, V, L, I, M, P (neutral/nonpolar).

Let the vertices of the unit square be: $A=(0,0)$, $B=(0,1)$, $C=(1,1)$, and

$D=(1,0)$. Successive points in the CGR were generated by an iterated function system defined by the following formula

$$(x_{i+1}, y_{i+1}) = \left(\frac{x_i + T_x(i)}{2}, \frac{y_i + T_y(i)}{2} \right)$$

Table 1. HIV_1 & HIV_2 data sets.

HIV_1		HIV_2	
CAD59561	CAD48441	ALQ56957	Q89928.3
AZI72458	CAD48455	2120212B	P18042.4
CAT00576	P03366.3	AIA59459	ATU79162
P04587.3	AZI72417	AAF82029	Q74120.3
P04588.3	AZI72491	ACH73021	P20876.3
AUO72800	AAN73511	BAH97695	P17757.3
AAD03225	AAN73835	ANG59323	AAC95341
Q9IDV9.3	AZI72386	ATU79172	APJ01827
AFB39387	AAD17072	APJ01785	ANG59330
BAC77486	BBC08805	AAT37062	ABV83026
Q79666.3	P12499.3	APJ01810	APJ01769
BAC77511	NP_057849	BAH97704	AAA64576
CAC86564	AZI72433	AAA43933	QLK12568
P20875.3	AZI72558	BAM76182	AYA94959
AAD03191	AUO72809	AAR98760	APJ01776
AAD03200	CAY83134	AIA59452	ALA65437
AAW68124	P0C6F2.1	QGV16580	AIA59451
AUO72845	O41798.3	Q76634.3	AIA59453
ABV00730	O93215.4	AAA43942	QGV16534
BBC08787	AAD03316	ATU79192	QGV16537
CAC38421		P18096.4	
AZI72408		AAA76841	
P03369.3		P12451.3	
AAG30116		ANG59316	
BBC08796		ALX35369	
AUO72688		QGV16583	
AAD03241		AYA94966	
CAB96338		BAA00710	
AAN73709		APJ01819	
AAD03184		AIA59450	

Table 2. SARS_CoV & SARS_CoV-2 ORF1ab polyprotein data sets.

SARS_CoV ORF1ab polyprotein		SARS_CoV-2 ORF1ab polyprotein	
QLG75207	QOF14847	QQI07512	QLG76455
QPN97028	QOU98004	QPI70323	QJR91795
QOU93276	QOQ14978	QPF58140	QPM28262
QQJ94670	QJX74509	QIA98605	QPM28286
QPZ45698	QIK02963	QPJ72410	QPJ72398
QPP19202	YP_009724389	QIC53203	QPJ72422
QQH18637	QPJ58632	QHD43415	QPI70311
QPZ33349	QQJ94682	QQJ95078	QPG83249
QPZ33508	QQJ95306	QHZ87591	QPG83261
QPZ75589	QPZ56528	QHO62876	QPG02368
QPN97040	QPZ56540	QHU79171	BCN28299
QQI07500	QPZ56564	QHN73809	BCN28311
QKS66638	QPZ75577	QPI75812	QPG00682
QQJ95318	QPV51018	QHZ00378	QPF21470
QOU87996	QPX60397	QHO60603	QHN73794
QMJ01339	QPP19226	QIB84672	QIH45022
QOQ07719	QPN97052	QPF58152	QHS34545
QPZ56552	QPN97064	BCA87360	BCB15089
QPF54048	QPN53402	QPF49350	QIA98553
QPV51042	QPN53415	QPI71724	QII57267
QPV51030		QQJ95090	
QPZ33361		QJR91771	
QPP19214		QOU97164	
QLJ57697		QNO98001	
QQI07488		QHR84448	
QMI94679		QPF49362	
QMI93420		QPI70335	
QQJ94103		QIG55993	
QLJ57685		QMJ01279	
QPJ58620		QPM28274	

Table 3. MERS & Dengue data sets.

MERS		Dengue	
AVN89429	AWH65952	QPZ88405	ANC57575
AID50417	AGR87639	QFS19562	ANC57576
ANC28665	YP_007188577	QPB40131	ANC57581
AKM76247	QGW51400	QFS19150	ANC57582

Continued

AJD81449	QOU08495	ACK28184	ANC57584
QFQ59585	QLD98092	QHR82546	ANC57591
AKJ80135	QEJ82213	QCZ25008	QGQ59490
ARQ84744	QDI73607	QFS19149	QGQ59491
QBM11746	QAT98897	ACL99188	QPU83821
ATQ39389	QAT98908	QQC97219	QPI70486
QOU08506	ANC28676	QPZ88403	QPI11926
AIZ48758	AMO03400	BBH51315	QPB40126
AKM76237	ALD51902	AEF01518	QPB40128
ANI69822	AHY21468	AAW23164	QPB40129
AKS48060	AHB33324	ANC57587	QOW96372
AZU90729	AVN89311	QPZ88404	QIB99388
AYM48029	AVN89418	ANC57579	QCZ25007
AWH65941	AUM60013	QPU83820	QIS48855
QGV13489	AUM60023	QPB40125	QBQ58384
QGV13494	AWH65953	QFS19134	QCE20685
AVN89300		QPB40127	
AHX71944		QGQ59492	
AHZ64055		ANC57577	
ANI69844		ANC57580	
ANI69833		QPI11922	
QGW51390		QIB99387	
QKX95935		ANC57586	
QBM11735		QBQ58385	
AHZ58509		ANC57578	
QJX19955		BBH51316	

Table 4. Ebola & HTLV data sets.

Ebola		HTLV	
ARU80343	QCH40643	ABM66546	P0C211.2
APT36405	QCH40651	AER08530	AAC82581
AWZ62332	AYP10283	ABM66560	P14078.3
AQA27316	QCF40472	QIZ31287	P03362.3
APA16576	ASU06439	QIZ31293	QIZ31284
APA16540	AXF48918	QIZ31278	QIZ31290
AYP66825	AXF48927	BAH85786	AAC00186
ATY51149	AXF48945	AYN25329	AAA85843
ARU80319	AXF48963	AOT98555	AAA96673
QEU56421	ARG43235	ABM66542	AYN25340

Continued

APT36396	APW30156	QIZ31299	AYN25351
ARC95311	APW30174	AOT98549	ATV90697
ASU06448	ARV89896	ABM66584	BAX76690
QNF60339	ARU80303	BBL33033	BAX76706
AYI50378	ARU80351	AOT98550	AHX00005
SCD11539	BAX08105	AAA85327	APR72307
AXE75594	AQS26699	AER08534	APR72311
ARU80359	AMY60341	AYN25362	ABM66540
APW30165	AMY60350	AOT98554	ABM66544
ARG43928	AMY60359	ATV90703	ABM66562
ARU80327		QIZ31296	
AXF48954		AAB20769	
ARG43937		BAA02931	
ALR82674		QNL15179	
AVQ09636		BAX76714	
AVQ09627		QIZ31281	
ARU80311		AAD50663	
AXH37632		ABM66574	
ALR82665		ABM66556	
ARU80335		ATV90700	

Table 5. SARS_CoV & SARS_CoV-2 ORF1a polyprotein data sets.

SARS_CoV ORF1a polyprotein		SARS_CoV-2 ORF1a polyprotein	
QRW47702	QRW76846	QLG76384	QLG76456
QRW78869	QRW99524	QJR91760	QSB33764
QRW99824	QQX03241	QSB33775	QRY28991
QRF77371	QRF77383	QRY29002	QRY29015
QRF77395	QRF77406	QRY29025	QRY29037
QOU96541	QOU97165	QRW37736	QRW37748
QOU98005	QOQ07720	QRW37760	QRW37772
QOQ14979	QMI93421	QRW37784	QPV15466
QMI94680	QMJ01280	QPV15478	QPV15490
QMJ01340	QLG75208	QPV15502	QPV15514
QKS66639	QJX74510		
QJR87032	QJR91124		
QJR91244	QSB33705		
QSB33717	QSB33729		
QSB33741	QSB33753		

Table 6. SARS_CoV spike glycoprotein & SARS_CoV-2 surface glycoprotein data sets.

SARS_CoV spike glycoprotein		SARS_CoV-2 surface glycoprotein	
AGT21273	AGT21288	BCN86425	BCN86437
AGT21303	AGT21318	QRK43459	QRX53181
BAF42873	ABD72999	QRX62421	QRU92034
ABD73000	ABD73001	QLG76853	QRG48189
ABD73002	ABB29898	QNO97727	QNP06175
AAV98002	AAV98003	QSB33742	QRW53372
AAU93318	AAU93319	QRX10874	QRX20730
AAU93320	AAV49722	QRC46729	QQY95746
AAV49723	AAT76147	QQV41531	QQN01053
AAS10463	AAS00003	QQN92410	QPC41132
AAP82968	AAP73417	QSB35551	QSB35563
AAP41037	AIA62320	QSB35575	QSB35587
AID16716	ADE34779	QSB35599	QQX20575
ADE34790	ADE34801	QQX20587	QQX20599
ADE34812	ADE34823	QQX20611	QQX20623

where $T_x(i)$ is the x coordinate and $T_y(i)$ is the y coordinate of the vertex of the corresponding group of the next amino acid in the sequence. To create a CGR image, we first began with an initial point $(0.5, 0.5)$, the center of a unit square in quadrant 1 of the xy-plane. A point is plotted half the distance from this vertex and the previous coordinate. The output file contained x, y coordinate values for each amino acid present in the input sequence. These x and y coordinate values were plotted as scatterplots. Some examples of the CGR of several viruses used in this report are shown in Figures 1-5.

2.3. Overview

The method we used to analyze and classify protein sequences has three steps: 1) generate graphical representations (images) of each Protein sequence using Chaos Game Representation (CGR), 2) compute all pairwise distances between these images using one of the following features, and 3) visualize the interrelationships implied by these distances as two- or three-dimensional maps, using Multi-Dimensional Scaling (MDS).

2.4. CGR Centroid

Once the CGR is created for a protein sequence, the CGR square is divided into four cells. Each cell represents one of the four groups, $\{A_i, B_i, C_i, D_i; i = 1, 2, \cdots, n\}$ where n is the length of the sequence. These cells correspond to the vertex located in that cell. The points in each cell are then averaged to find the centroid of each cell denoted by

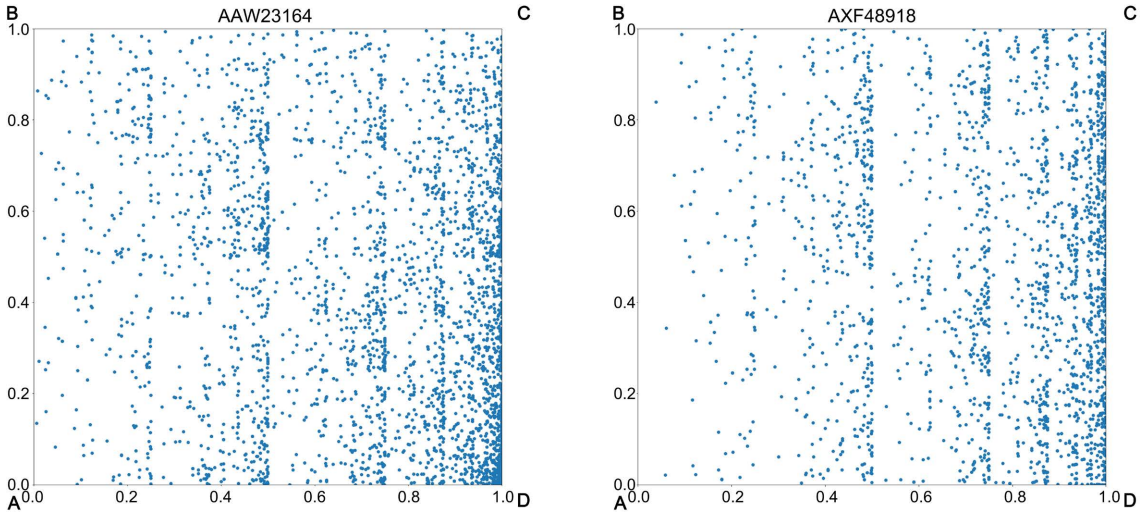

Figure 1. CGR of Dengue (left) and Ebola (right).

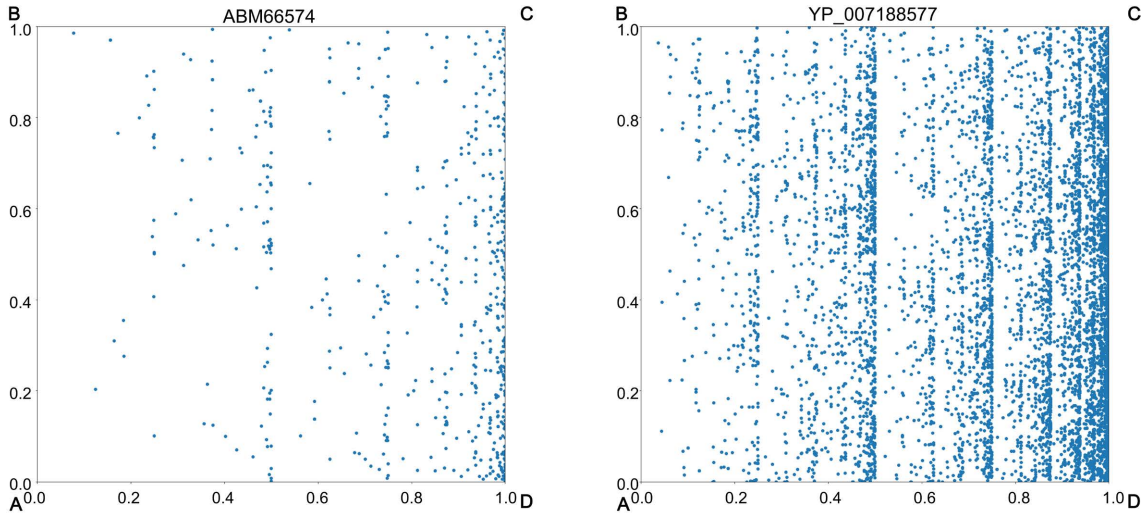

Figure 2. CGR of HTLV (left) and MERS (right).

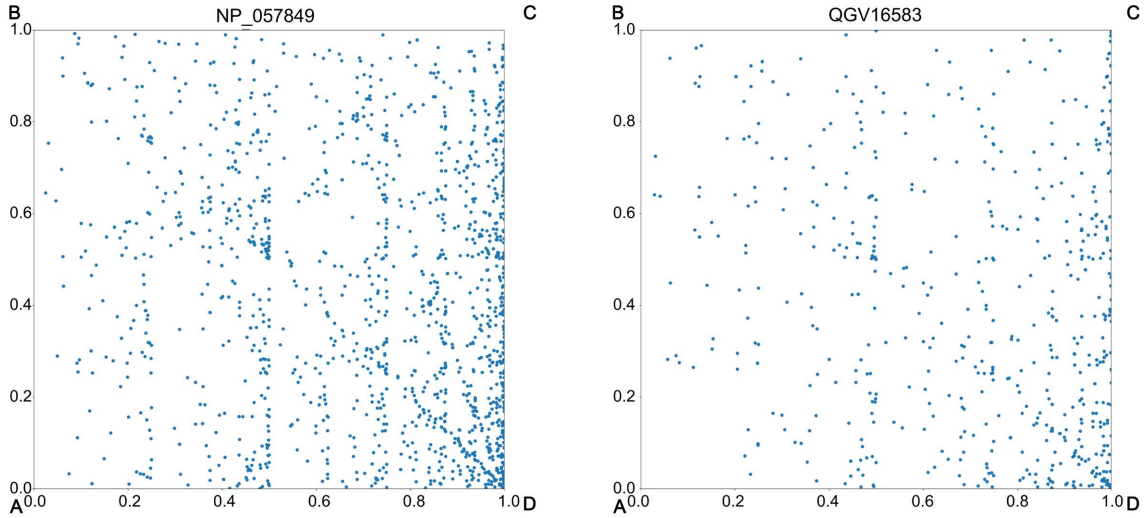

Figure 3. CGR of HIV_1 (left) and HIV_2 (right).

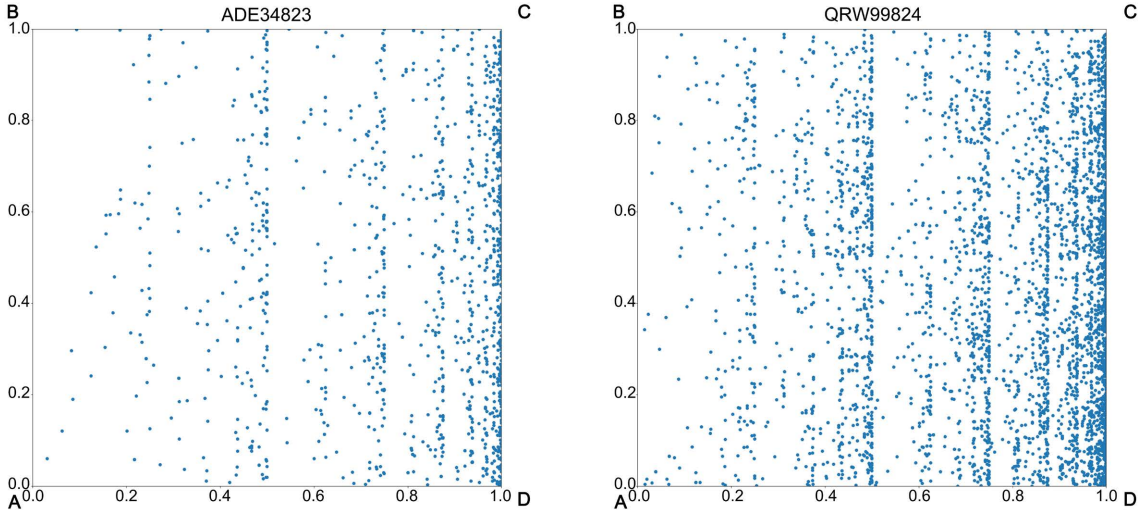

Figure 4. CGR of SARS_CoV spike glyco (left) and SARS_CoV ORF1a (right).

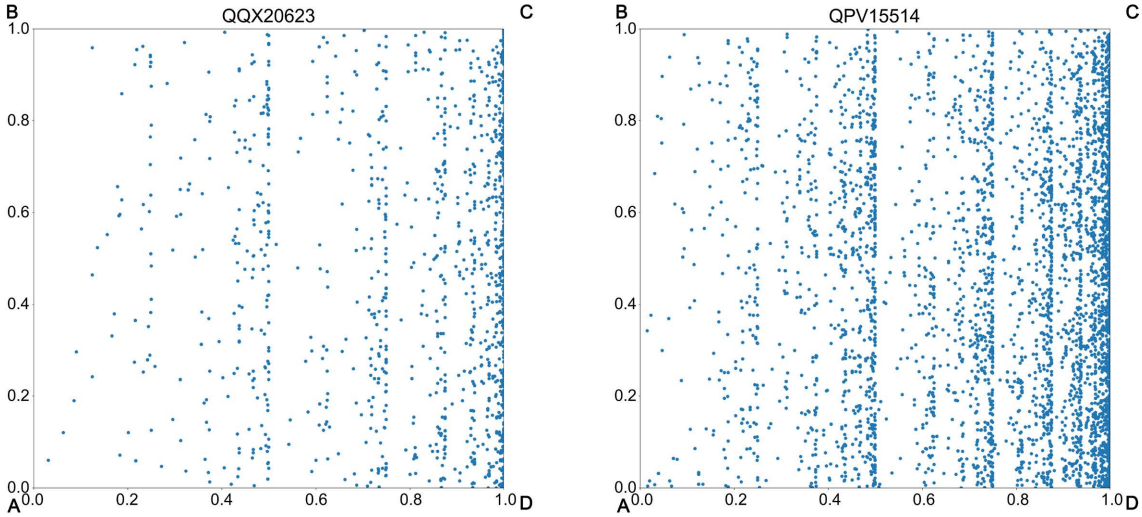

Figure 5. CGR of SARS_CoV-2 surface glyco (left) and SARS_CoV-2 ORF1a (right).

$$C_k = \left(\frac{\sum_{i=1}^{n} a_i(x)}{n}, \frac{\sum_{j=1}^{n} a_j(y)}{n} \right)$$

where $a_i(x)$ and $a_i(y)$ are the x and y coordinates respectively in a cell and $k = 1,2,3,4$. This gives four centroids C_1, C_2, C_3 and C_4 for comparison of viral sequences.

2.5. CGR Centroid Bisection

Upon calculation of the four CGR centroids, a rectangle is created from these vertices. Next, the diagonals of this rectangle are constructed and their intersection is taken as the CGR Centroid Bisection denoted $B_C(x)$ of viral sequence x.

$$B_C(x) = \frac{C_1 + C_4}{2}$$

2.6. Amino Acid Frequency

The next method of sequence comparison examined is the amino acid frequency (AAF) of 2mers. A 2mer is subsequence of length 2 of a string of characters and they are found by taking the cross product between the set of amino acids and itself. This yields $20^2 = 400$ possible 2mers and some of these include: DE, MA, AR, HE, and RT. The frequency of each 2mer is calculated as follows

$$p_{ij} = \frac{\text{Number of occurences of 2mer}}{400}$$

$1 \le i \le j \le 400$. Several distance measures can then be obtained by comparing the amino acid FCGR of viral sequences. One distance metric that encompasses two others is the minkowski distance and is derived as follows

$$\sum_{i=1}^{n}\left(\left|p_{ij} - p'_{ij}\right|^{t}\right)^{\frac{1}{t}}$$

Note that when $t = 1$, we have

$$M = \sum_{i=1}^{n}\left(\left|p_{ij} - p'_{ij}\right|\right)$$

which is manhattan distance and when $t = 2$, we have

$$E = \sqrt{\sum_{i=1}^{n}\left(\left|p_{ij} - p'_{ij}\right|^{2}\right)}$$

euclidean distance.

2.7. Group Frequency Chaos Game Representation

Each cell in the CGR of protein contains an x amount of points and by dividing this amount by four for the four cells; we have the group frequency chaos game representation (GFCGR). With this the frequencies are defined for the four groups of amino acids as opposed to just one. The GFCGR is defined as follows:

$$\text{GFCGR}(z) = \frac{\text{Number of occurences of amino acid in a group } z}{\text{Length of the sequence}}$$

where $z = \{A, B, C, D\}$.

2.8. Kullback-Leibler Discrimination Information

A previous method introduced by Li [20] utilized the Kullback-Leibler Discrimination Information for sequence comparison. This comparison proved useful and in this report we further extend this method to be applicable with our previously mentioned methods. Given a discrete random variable Y, different distribution laws can be applied. For example under Hypothesis 1, we have

$$\begin{pmatrix} Y \\ p_1(y) \end{pmatrix} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ p_1(y_1) & p_1(y_2) & \cdots & p_1(y_n) \end{pmatrix}$$

Under Hypothesis 2, we have

$$\begin{pmatrix} Y \\ p_2(y) \end{pmatrix} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ p_2(y_1) & p_2(y_2) & \cdots & p_2(y_n) \end{pmatrix}$$

These distributions can be compared by using the Kullback-Leibler Discrimination Information denoted by

$$I(p_1, p_2) = \sum_{i=1}^{n} p_1(a_i) \log \frac{p_1(a_i)}{p_2(a_i)}$$

In this report, we let these distributions be the 2mer AAF of viral genomes. So for viruses *x* and *y*, we have $I(x,y)$, but due to its directed divergence $I(x,y)$ might not necessarily equal $I(y,x)$. For this reason, the metric $J(a,b)$ is defined as follows

$$J(x,y) = I(x,y) + I(y,x)$$

Note that when $x = y$, $J(x,y) = 0$. We also note that for any two viral sequences *x* and *y*, $J(x,y) = J(y,x)$. Li [20] noted that this method can accurately measure the dissimilarity between two sequences.

2.9. Compounded Frequency

Another method for sequence comparison that has been previously examined is the compounded frequency. This method was proposed by Almeida [3] for comparison of biological sequences. First we denote the compounded frequency nw as follows

$$nw = \sum_{i=1}^{k} x_i * y_i$$

The compounded frequency is then used in conjunction with the Pearson correlation coefficient, *rw* for sequence comparison.

$$rw = \frac{\sum_{i=1}^{k} \frac{x_i - \mu x}{\sqrt{sx}} * \frac{y_i - \mu y}{\sqrt{sy}} * x_i * y_i}{nw}$$

where

$$sx = \frac{\sum_{i=1}^{k} (x_i - \mu x)^2 * x_i * y_i}{nw}$$

and

$$sy = \frac{\sum_{i=1}^{k} (y_i - \mu u)^2 * x_i * y_i}{nw}$$

with μx and μy defined as follows

$$\mu x = \frac{\sum_{i=1}^{k} x_i^2 * y_i}{nw}$$

$$\mu y = \frac{\sum_{i=1}^{k} y_i^2 * x_i}{nw}$$

Previous studies used this method for comparison of the FCGR of two sequences. Similarly, we use the 2mer AAF to find the rw between two sequences. By using the weight of nw, each 2mer is proportional to its frequency. Now we define the sequence distance as $d = 1 - rw$, which has values from 0 - 2. For $d > 1$, a negative correlation exists and for $d < 1$ a positive correlation exists. When $d = 0$, the sequences are exactly similar.

2.10. Shannon Entropy

The Shannon information index has been used in some of our past work as well as other studies. It is denoted

$$S_2 = -\sum_{i=1}^{k} p_i * \log_2(p_i) = \sum_{i=1}^{k} p_i * \log_2\left(\frac{1}{p_i}\right)$$

where 2mer AAF $= p_1, p_2, \cdots, p_n, 1 \le i \le n$. This method has been used in some of our past works for sequence comparison. In this report we use this method as a measure of the amount of information contained within a sequence of proteins.

3. Results

For the Shannon entropy, 2mer AAF and GFCGR the manhattan distance is used. The euclidean distance is applied to both the CGR centroids and CGR centroid bisections, while $J(x, y)$ and Pearson correlation have the respective distance measures. MDS is then applied to the distance matrices to create 2D and 3D projections shown in Figures 6-12.

To rank the effectiveness of each distance metric, we define the $\delta(x, y)$ function as in [26] of two viral sequences x and y as follows

$$\delta(x, y) = \begin{cases} 0, \text{ if } x \text{ and } y \text{ belong to same viral group} \\ 1, \text{ otherwise} \end{cases}$$

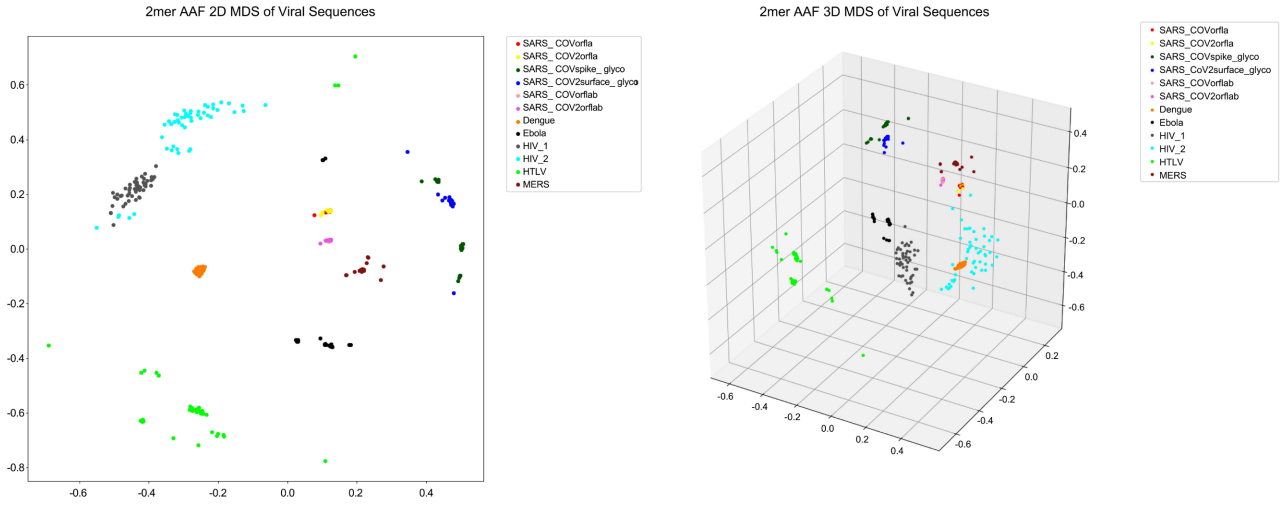

Figure 6. 2mer AAF 2D and 3D MDS charts.

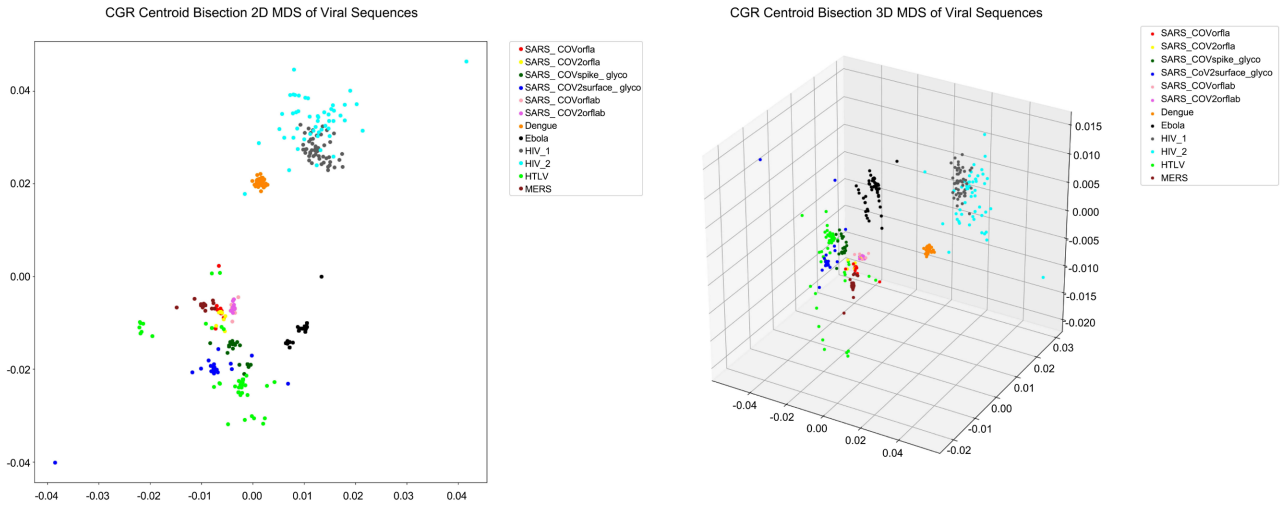

Figure 7. CGR Centroid Bisection 2D and 3D MDS charts.

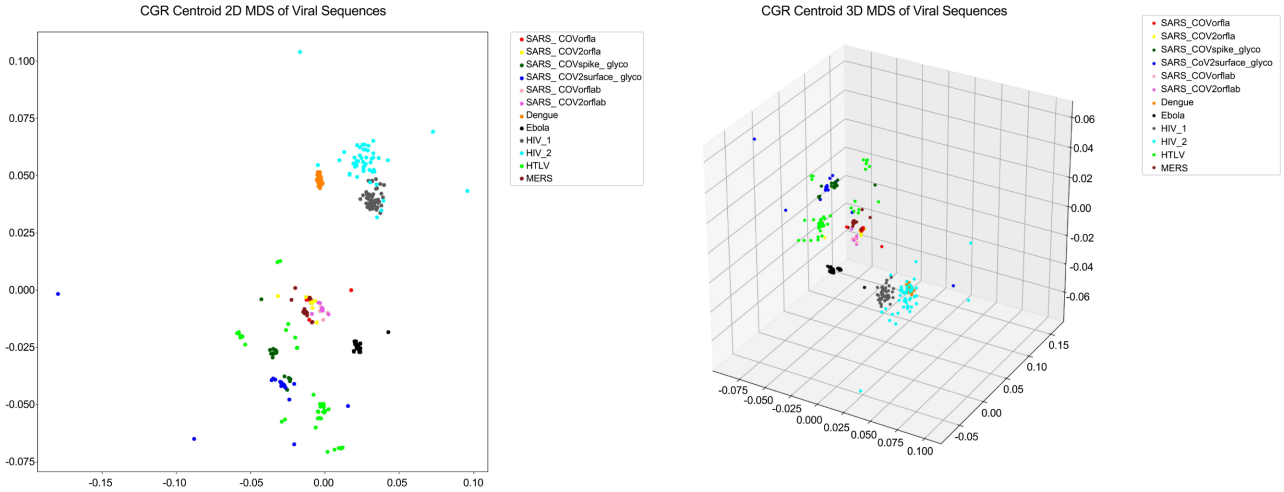

Figure 8. CGR Centroid 2D and 3D MDS charts.

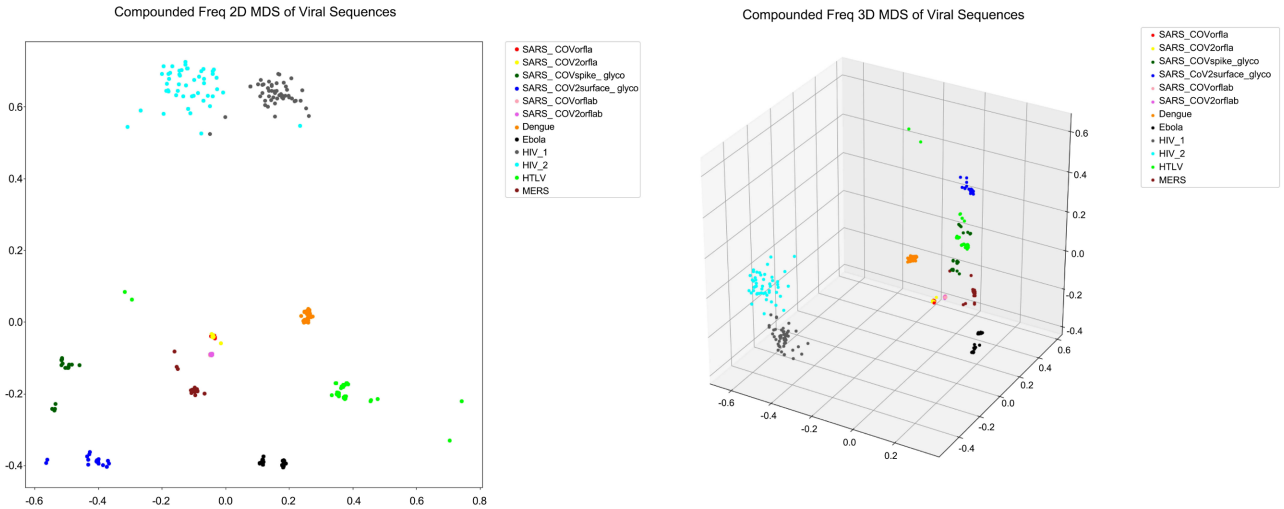

Figure 9. Compounded Freq 2D and 3D MDS charts.

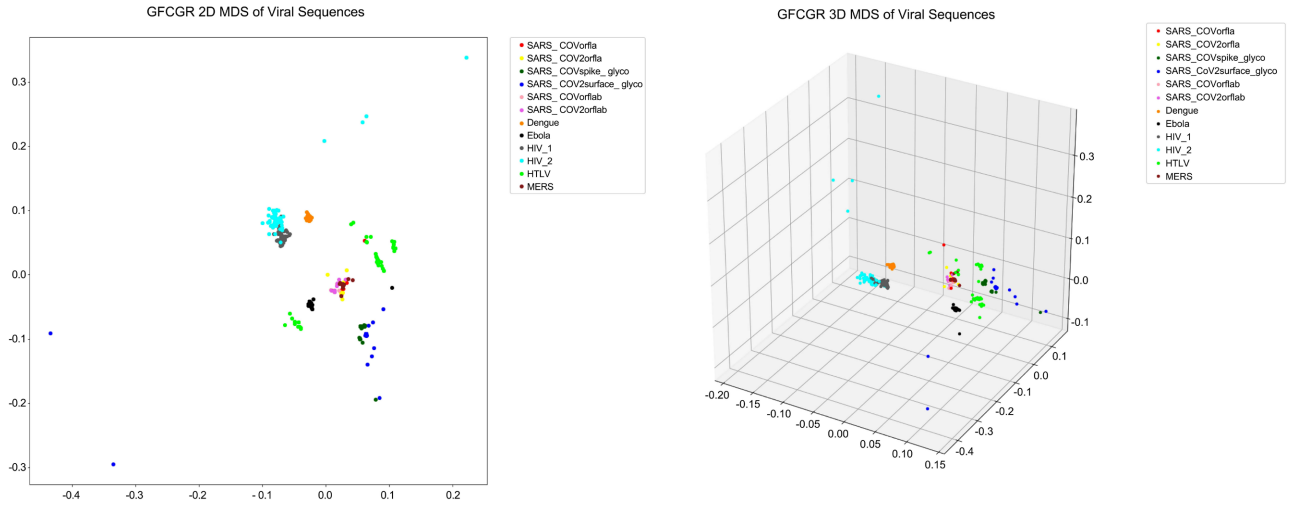

Figure 10. GFCGR 2D and 3D MDS charts.

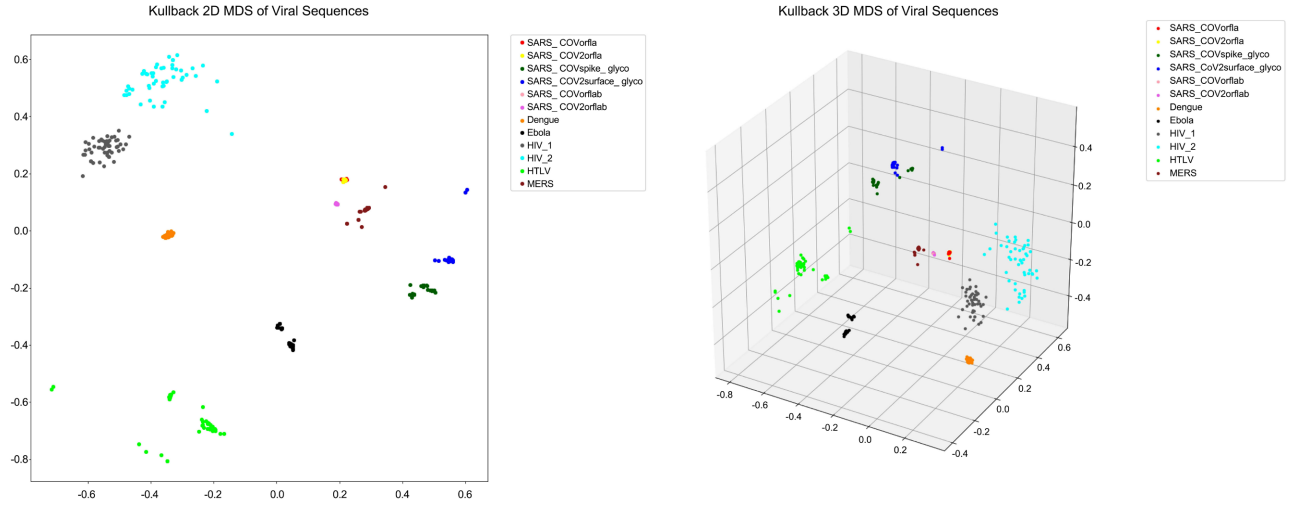

Figure 11. Kullback-Leibler 2D and 3D MDS charts.

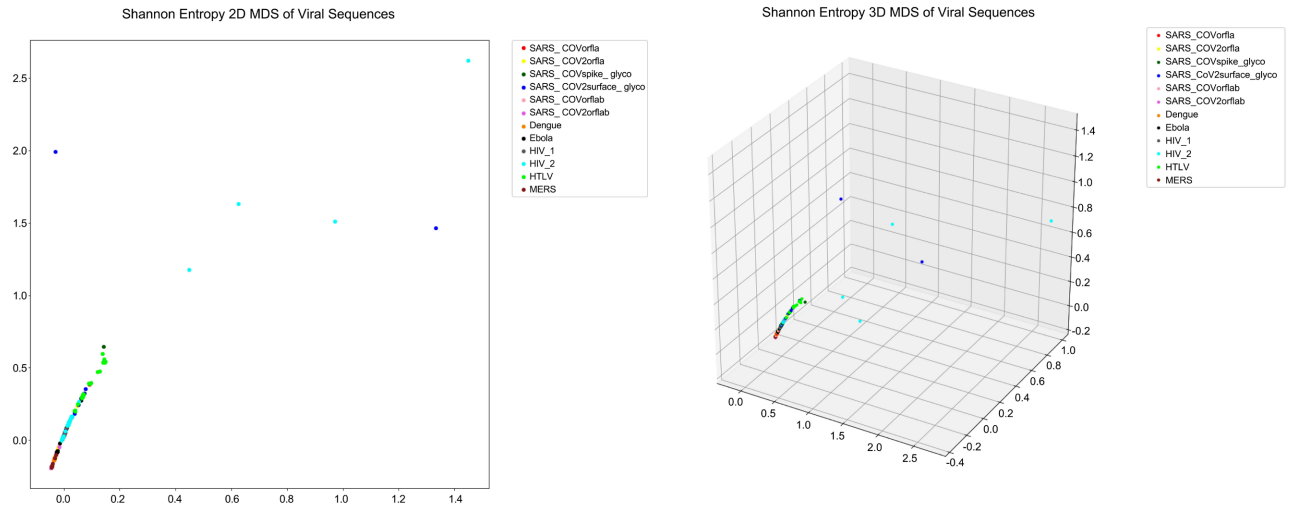

Figure 12. Shannon Entropy 2D and 3D MDS charts.

With this function we create a 400 × 400 distance matrix of the viruses and take the upper triangular matrix as a vector U_δ. Next, we take the upper triangle matrix, $U_\alpha, \alpha \in$ 2mer AAF, $J(x,y)$, S_2, $D = 1 - rw$, GFCGR, CGR Centroid, CGR Centroid Bisection of each of the 7 distance matrices for comparison with U_δ. The Pearson correlation coefficient is used to establish how well a distance measure fits a particular viral sequence to its corresponding group cluster. We denote this coefficient as

$$P_\alpha = \frac{\sigma_{\alpha\delta}}{\sigma_\alpha \sigma_\delta}$$

with a range of $[-1,1]$. Values of 1 indicate a linear correlation between U_δ and U_α while a value of 0 indicates the pair is unrelated. The values of P_α for each distance measure are shown in Table 7.

We see that of the distance measures, 2mer AAF is most closely related with U_δ. Further confirmation of this is shown in the 2D and 3D MDS charts for 2mer AAF Figure 6, which show a good separation of the viral sequences into their respective groups. It can also be noted that viruses belonging to the coronavirus family cluster close together as do viruses belonging to the HIV family. We expect this as these viruses are more closely related than say HTLV or Dengue. In fact, SARS_CoV ORF1a and SARS_CoV-2 ORF1a overlap as do SARS_CoV ORF1ab and SARS_CoV-2 ORF1ab. This is indicative of a distance measure of almost 0, which shows just how closely related they are. Other measures such as Shannon entropy and GFCGR which have the lowest correlation with U_δ, $P_\alpha = 0.147915$ and 0.36562 respectively, show a lack of separation between viral groups in their MDS charts Figure 10 and Figure 12.

4. Discussion and Conclusions

Feature extractions of protein sequences play an important role in protein sequence similarity studies. Although many methods have been proposed for extracting features of protein sequences, most of them showed great limits in practical applications. Many studies have shown that the CGR-based strategy would be one of the most useful approaches for protein feature extractions, and the so-called FCGR method is currently the most frequently used method-based CGR, however, a large amount of useful information, e.g. physicochemical properties of amino acids and the distribution information of points in the CGR image were not taken into consideration in the method of FCGR.

In this study, CGR was used for the identification of several hundred protein sequences into their respective viral groups through feature extraction. These features include CGR centroid, amino acid frequency, compounded frequency, Shannon entropy, and Kullback-Lieber Discrimination Information.

The method, we used to analyze and classify protein sequences, has three steps: 1) generate graphical representations (images) of each Protein sequence using Chaos Game Representation (CGR), 2) compute all pairwise distances between these images, and 3) visualize the interrelationships implied by these distances as two- or three-dimensional maps, using Multi-Dimensional Scaling (MDS).

Table 7. P_α of distance metrics.

Method	P_α
2mer AAF	0.556726
$J(a,b)$	0.537113
$D=1-rw$	0.486536
CGR Centroid	0.405884
CGR Centroid Bisection	0.36107
GFCGR	0.36562
S_2	0.147915

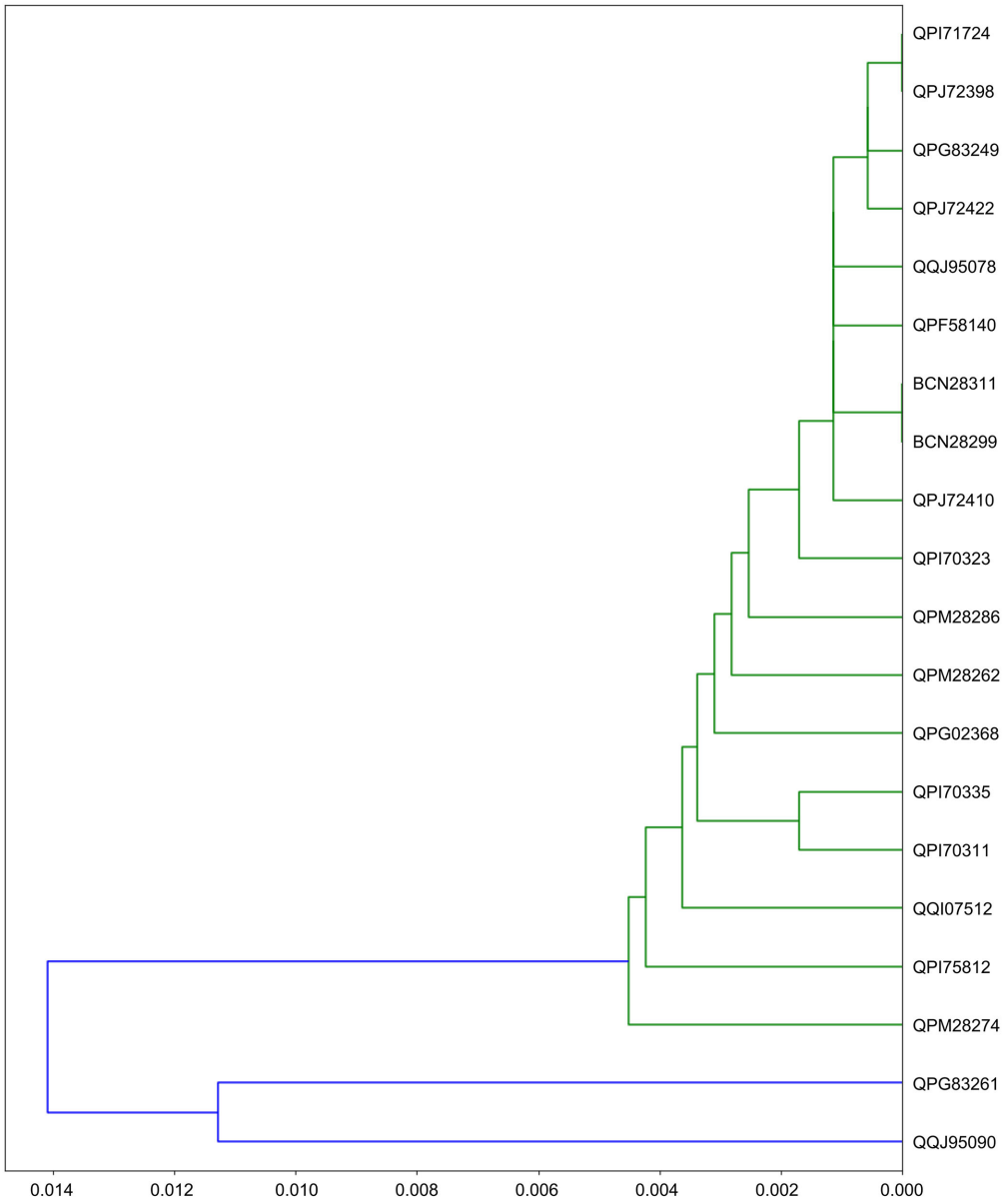

Figure 13. Dendrogram made using 2mer AAF of SARS_CoV-2 ORFlab.

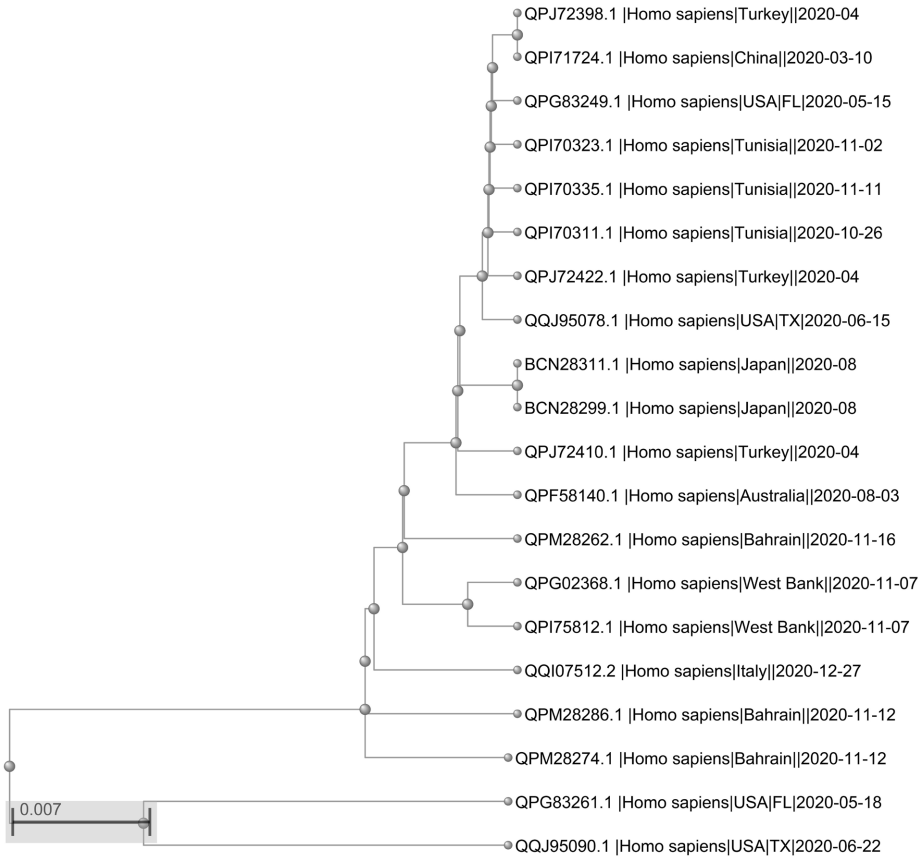

Figure 14. Phylogenetic tree of SARS_CoV-2 ORF1ab from NCBI website.

Several distance metrics were introduced for comparison as well as a method of ranking these metrics. Our quantitative comparison of seven different distances suggests that the Kullback-Lieber Discrimination Information as well as the manhattan distance of 2mer AAF outperform all other distances. Our findings suggest that the Kullback-Lieber Discrimination Information as well as the manhattan distance of 2mer AAF is best in clustering viruses into their respective groups. This shows the importance of the frequency of 2mers in correctly identifying viral sequences. We compare the results of the phylogenetic tree of SARS_CoV-2 ORF1ab obtained from our 2mer AAF distance method with those given in the NCBI site in **Figure 13** and **Figure 14**. The NCBI method performs equally well with our 2mer distance method. The two-dimensional and three-dimensional Molecular Distance Maps we obtain, which visualize the simultaneous interrelationships among the sequences in our dataset, show this method's potential. Further analysis is needed to explore this method's potential for the analysis of closely related sequences.

In conclusion, our distance comparison results on datasets illustrate the potential strengths of CGR-based method for examining the evolutionary relationship. Our method is powerful for extracting effective features from protein sequences, and therefore important in classifying proteins and inferring the phylogeny of viruses.

Acknowledgements

This work was done while D.C.S. mentored undergraduate student Kevin Simmons and a graduate student Matthew Hill. Kevin Simmons was funded by NIH-Minority Access to Research Career (MARC) Program Grant # NIH-NIGMS T34 GM 100831-09. This research was also partially funded by the University of North Carolina System's Covid-19-related grant. The authors also would like to thank Mr. Joel W. Perry for spending time to fix all the Latex errors.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] MedlinePlus. What Are Proteins and What do They Do? U.S. National Library of Medicine. https://medlineplus.gov/genetics/understanding/howgeneswork/protein

[2] Rigden, D.J. (2009) From Protein Structure to Function in Bioinformatics. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4020-9058-5

[3] Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A. and Fletcher, M. (2001) Analysis of Genomic Sequences by Chaos Game Representation. *Bioinformatics*, **17**, 429-437. https://doi.org/10.1093/bioinformatics/17.5.429

[4] Jeffrey, H.J. (1990) Chaos Game Representation of Gene Structure. *Nucleic Acids Research*, **18**, 2163-2170. https://doi.org/10.1093/nar/18.8.2163

[5] Olyaee, M.H., Khanteymoori, A. and Khalifeh, K. (2019) Application of Chaotic Laws to Improve Haplotype Assembly Using Chaos Game Representation. *Scientific Reports*, **9**, Article No. 10361. https://doi.org/10.1038/s41598-019-46844-y

[6] Olyaee, M.H., Pirgazi, J., Khalifeh, K. and Khanteymoori, A. (2020) RCOVID19: Recurrence-Based SARS-CoV-2 Features Using Chaos Game Representation. *Data Brief*, **32**, Article ID: 106144. https://doi.org/10.1016/j.dib.2020.106144

[7] Lchel, H.F. and Heider, D. (2021) Chaos Game Representation and Its Applications in Bioinformatics. *Computational and Structural Biotechnology Journal*, **19**, 6263-6271. https://doi.org/10.1016/j.csbj.2021.11.008

[8] Joseph, J. and Sasikumar, R. (2006) Chaos Game Representation for Comparison of Whole Genomes. *BMC Bioinformatics*, **7**, 243-252. https://doi.org/10.1186/1471-2105-7-243

[9] Tanchotsrinon, W., Lursinsap, C. and Poovorawan, Y. (2015) A High Performance Prediction of HPV Genotypes by Chaos Game Representation and Singular Value Decomposition. *BMC Bioinformatics*, **16**, Article No. 71.

[10] Hoang, T., Yin, C.C. and Yau, S.S.-T. (2016) Numerical Encoding of DNA Sequences by Chaos Game Representation with Application in Similarity Comparison. *Genomics*, **108**, 134-142. https://doi.org/10.1016/j.ygeno.2016.08.002

[11] Goldman, N. (1993) Nucleotide, Dinucleotide and Trinucleotide Frequencies Explain Patterns Observed in Chaos Game Representations of DNA Sequences. *Ucleic Acids Research*, **21**, 2487-2491. https://doi.org/10.1093/nar/21.10.2487

[12] Fiser, A., Tusndy, G.E. and Simon, I. (1994) Chaos Game Representation of Protein Structures. *Journal of Molecular Graphics*, **12**, 302-304. https://doi.org/10.1016/0263-7855(94)80109-6

[13] Randic, M., Butina, D. and Zupan, J. (2006) Novel 2-D Graphical Representation of Proteins. *Chemical Physics Letters*, **419**, 528-532. https://doi.org/10.1016/j.cplett.2005.11.091

[14] Basu, S., Pan, A., Dutta, C. and Das, J. (1997) Chaos Game Representation of Proteins. *Journal of Molecular Graphics & Modelling*, **15**, 279-289. https://doi.org/10.1016/S1093-3263(97)00106-X

[15] Bhoumik, P. and Hughes, A.L. (2018) Chaos Game Representation: An Alignment-Free Technique for Exploring Evolutionary Relationships of Protein Sequences. https://doi.org/10.1101/276915

[16] Yu, Z.G., Anh, V. and Lau, K.S. (2004) Chaos Game Representation of Protein Sequences Based on the Detailed HP Model and Their Multifractal and Correlation Analyses. *Journal of Theoretical Biology*, **226**, 341-348. https://doi.org/10.1016/j.jtbi.2003.09.009

[17] Qi, Z., Li, K., Ma, J., Yao, Y. and Liu, L. (2018) Novel Method of 3-Dimensional Graphical Representation for Proteins and Its Application. *Evolutionary Bioinformatics*, **14**, 1-8. https://doi.org/10.1177/1176934318777755

[18] Mu, Z., Yu, T., Qi, E., *et al.* (2019) DCGR: Feature Extractions from Protein Sequences Based on CGR via Remodeling Multiple Information. *BMC Bioinformatics*, **20**, Article No. 351. https://doi.org/10.1186/s12859-019-2943-x

[19] Mehri, M., Fatemeh, A. and Vahid, Z. (2018) A Novel Graphical Representation and Similarity Analysis of Protein Sequences Based on Physiochemical Properties. *Physica A*, **510**, 477-485. https://doi.org/10.1016/j.physa.2018.07.011

[20] Li, N., Shi, F., Niu, X. and Xia, J. (2009) A Novel Method to Reconstruct Phylogeny Tree Based on the Chaos Game Representation. *Journal of Biomedical Science and Engineering*, **2**, 582-586. https://doi.org/10.4236/jbise.2009.28084

[21] Sun, Z., Pei, S., He, R.J. and Yau, S. (2020) A Novel Numerical Representation for Proteins: Three-Dimensional Chaos Game Representation and Its Extended Natural Vector. *Computational and Structural Biotechnology Journal*, **18**, 1904-1913. https://doi.org/10.1016/j.csbj.2020.07.004

[22] Yu, L., *et al.* (2017) Protein Sequence Comparison Based on Physicochemical Properties and the Position-Feature Energy Matrix. *Scientific Reports*, **7**, Article No. 46237. https://doi.org/10.1038/srep46237

[23] Hannah, F., Lchel, D.E., Sperlea, T. and Heider, D. (2020) Deep Learning on Chaos Game Representation for Proteins. *Bioinformatics*, **36**, 272-279. https://doi.org/10.1093/bioinformatics/btz493

[24] Sengupta, D.C., Hill, M.D., Benton, K.R. and Banerjee, H.N. (2020) Similarity Studies of Corona Viruses through Chaos Game Representation. *Computational Molecular Bioscience*, **10**, 61-72. https://doi.org/10.4236/cmb.2020.103004

[25] Kruskal, J. (1964) Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, **29**, 1-27. https://doi.org/10.1007/BF02289565

[26] Karamichalis, R., Kari, L., Konstantinidis, S., *et al.* (2015) An Investigation into Inter and Intra-Genomic Variations of Graphic Genomic Signatures. *BMC Bioinformatics*, **16**, Article No. 246. https://doi.org/10.1186/s12859-015-0655-4

[27] Randhawa, G.S., Soltysiak, M.P.M., El Roz, H., de Souza, C.P.E., Hill, K.A. and Kari, L. (2020) Machine Learning Using Intrinsic Genomic Signatures for Rapid Classification of Novel Pathogens: COVID-19 Case Study. *PLOS ONE*, **15**, e0232391. https://doi.org/10.1371/journal.pone.0232391