AN ALIGNMENT-FREE METHOD FOR SEQUENCE IDENTIFICATION USING CHAOS GAME REPRESENTATION

by

MATTHEW D. HILL

A thesis submitted to the Graduate Faculty of
Elizabeth City State University
in partial fulfillment of the
requirements for the Degree of
Master of Science in Mathematics

Elizabeth City, North Carolina

May 2021

APPROVED BY

Julian A.D. Allagan, Ph.D.	Hirendra N. Banerjee, M.D.,Ph.D.
Malcolm DCosta, Ph.D.	Kenneth L. Jones, Ph.D.

Dipendra C. Sengupta, Ph.D.
Chair, Thesis Committee

ABSTRACT OF THESIS

AN ALIGNMENT-FREE METHOD FOR SEQUENCE IDENTIFICATION USING CHAOS GAME REPRESENTATION

Recent events in the area of public health have lead to the need for advancements in techniques to better understand viruses. A method of graphically representing biological sequences known as chaos game representation (CGR) was proposed by H.J. Jeffrey in 1990 [1] and has proved useful even today in the field of bioinformatics. CGR uses the midpoint distance formula to transform a sequence of characters into a graph that can help distinguish between biological sequences through pattern recognition. Initially, CGR was apllied to DNA sequences, but in our case we apply it to protein sequences. For this report CGR is used for the identification of several hundred protein sequences into their respective viral groups through feature extraction using python programming language. These features include, CGR centroid, amino acid frequency, compounded frequency, Shannon entropy, and Kullback-Lieber Discrimination Information. In turn better classification and identification of viruses is achieved.

DEDICATION

I would like to dedicate this thesis to my family that has been with me from the beginning and has helped me along the way. Without them I do not know where I would be in life.

ACKNOWLEDGEMENT

I would like to thank Dr. Dipendra Sengupta as his help has been instrumental in the completion of my thesis. I would also like to thank my family for supporting me along the way. A huge thank you to my thesis committee and the rest of the Math Department for your teaching and guidance during my time in the masters program.

Contents

List of Figures

1 Introduction and Review of Literature

The Central Dogma of Biology revolves around the transcription of deoxyribonucleic acid (DNA) into ribonucleic acid (RNA) and the translation of that RNA into proteins. DNA serves as the language in which organisms are written and studying features about it along with RNA and proteins can help to answer many biological questions. 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 transport, storage to structural components, and enzymes [10]. By studying the structure and function of proteins we can hurdle some of the obstacles in understanding evolutionary relationships of organisms.

The 20 amino acids that occur naturally in nature 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) [10]. Each amino acid has certain physical and chemical properties which distinguish it from others and in this report we focus the polarity and charge. This method of grouping of proteins was shown to be useful for sequence identification in

comparison to random grouping of proteins [2]. It was noted by Rigden [9] that similar protein sequences have similar functions. This leads to difficulty when comparing closely and distantly related sequences.

As mentioned above, DNA is transcribed into RNA and RNA is then translated into proteins, but some viruses use an enzyme known as reverse transcriptase to reverse transcribe their RNA into complimentary DNA (cDNA) for the host to use. These viruses that belong to the viral family Retroviridae are referred to as retroviruses and some of the common examples that impact humans include Human T-Cell Leukemia Virus Type 1 (HTLV 1), Human Immunodeficiency Virus Type 1 (HIV 1), and Human Immunodeficiency Virus Type 2 (HIV 2) [7]. For this report, HTLV 1, HIV 1, HIV 2, Ebola, Dengue, Middle Eastern Respiratory Syndrome (MERS), Severe Acute Res- piratory Syndrome Coronavirus (SARS-COV), and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2) were used for protein sequence comparison. SARS-COV2 has been detrimental to the human population over the past year. At the time of this report, more than 90 million people have contracted the virus with over 50 million recoveries, and over 2 million deaths. The first pathogenic novel coronavirus, discovered in 2003 and named SARS-CoV, caused SARS, a 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 [29]. SARS-COV2 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 genomic sequence, for planning, containment, and treatment. Motivated by the above need, we report several alignment-free methods combining with CGR to perform clustering analysis and create a phylogenetic tree based on it.

Viral sequences and other biological sequences tend to have variation within a species and this leads to variation in representing a particular group of viruses. In order to create dendrograms within python, certain parameters must be met such as the distance measure used in unweighted pair group method using arithmetic averages (UPGMA). These measures are referred to as distance based, which differ from other methods such as maximum likelihood and maximum parsimony [22], [23]. It is because of this variation multiple sequences from a particular group must be examined as well as sequences from different regions [4]. Comparing sequences can help to study the variation of viruses as well as the structure and function of proteins [6]. One method of comparison is alignment based in which a scoring system picks the best alignment. Such methods have shown some progress in sequence identification using global and local alignment [11] [12], but it has been noted this method has several drawbacks [2], [6], [5]. These include

sequence variation from distantly related sequences and high computation costs when aligning multiple sequences. Other methods of phylogenetic tree construction involve gene order [13] or gene family content and are helpful when complete genomes of the sequences are available. Some setbacks to this method include the small size of viral genomes and sequences that do not have a complete genome available.

Finally, another method of sequence comparison, which overcomes these drawbacks is alignment-free (AF). AF methods have been particular useful for sequence comparison due to their low computation cost and speed of analysis. The field of bioinformatics has seen an uptick in the use of these methods due to advancements in sequencing technology which have allowed for access to far more biological data than previously obtainable. Current AF methods in use today include iterated-function systems and chaos theory for sequence representation such as chaos game representation [1], information theory such as shannon information index [19], Fourier transformations such as digital signal processing [18], and moments of the positions of the nucleotides [20],[21]. Of the alignment-free methods mentioned, graphical representations have proved to be the most useful [24] [14], [15], [16],[6], [4], [5], [3], [2]. Descriptors are used for each protein based on numerical characterizations obtained from these graphical representations. The k-mer based methods are AF and have been among the most used [3], [5]. K-mer refers to subsequences of length k of a biological sequence. Applications of these methods have been utilized for phylogenetic analysis of viral and bacterial

genomes. The frequency feature profile (FPP) is an example of such methods and has been found to perform well when compared to natural vector methods. [5].

For this report, CGR was applied to protein sequences to distinguish between several species of viruses. It is used as a basis to obtain information about the viruses being studied. The protein sequences of the viruses were obtained from the National Center for Biotechnology Information (NCBI) website. 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 application of our proposed methods we apply multidimensional scaling (MDS) to the data. With this 2D and 3D projections of the data can be obtained for clustering analysis. Kruskal[36] 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 was has been previously shown using different methods [31].

2 Previous Work

2.1 Capstone

Computational biology deals with the use of mathematical tools to extract useful information from biological data. In this report we aim to use chaos game representation (CGR) as a means to identify organisms based on similarities they show in their graphs. The CGR graph can have recog- nizable patterns in the nucleotide sequences, obtained from NCBI website. The graphs are constructed by considering a DNA sequence as strings com- posed of four units, A, T, C and G. Similarities and differences in the CGR graphs can be quantified mathematically. This mathematical formula being used is the distance between points. The CGR graphs can also be a way to visualize fractals. Several different order Markov Chains were applied to genomes to help predict the occurrence of oligonucleotides with varying lengths. Probability matrices as well as kmer heat maps and DNA centroids were additionally extracted from the CGR graphs.

2.1.1 Introduction

Nucleic acids are the genetic code in which all organisms are comprised. One specific nucleic acid is known as deoxyribonucleic acid(DNA). Studying the structure and patterns in DNA can help with understanding the functionality of different genes. DNA has a double helix structure comprised of nitrogenous bases, phosphate groups, and sugar molecules. The four nitrogenous bases

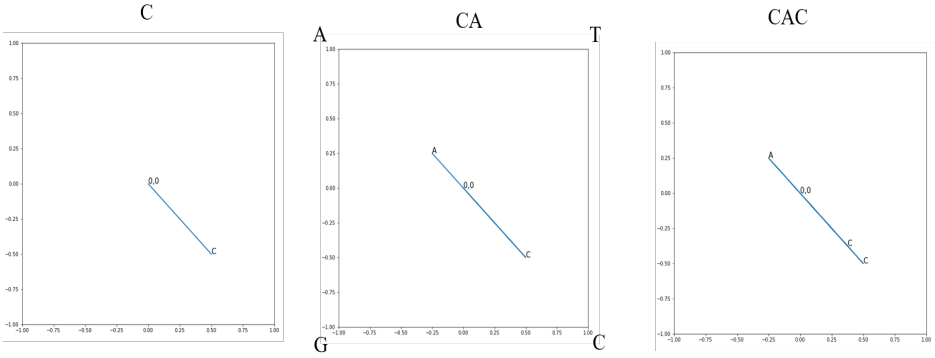
are adenine(A), thymine(T), guanine(G), and cytosine(C). These serve as the alphabet in which DNA is written as well as the focus of this research.

2.1.2 Methods - CGR

The recursive function being used to produce the CGR is

$$(x_{i+1}, y_{i+1}) = (\frac{x_i \pm 1}{2}, \frac{y_i \pm 1}{2})$$

where $(x_0, y_0) = (0, 0)$ and depending on the current letter in the sequence of DNA the next coordinate is half the distance between that letter and the previous coordinate. In our research we let A = [-1,1] , T = [1,1], G = [-1,-1], and C = [1,-1] be our vertices of the unit square for CGR on the xy plane. Figure 1 shows the CGR of the sequence CACGTT.



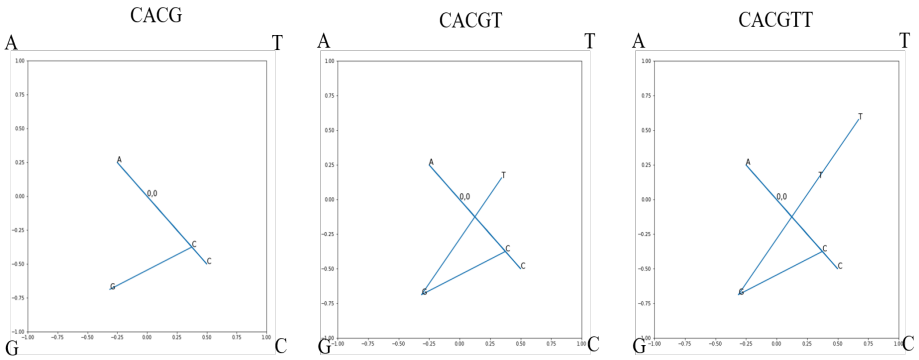


Figure 1: CGR of CACGTT

This method of recursive midpoint application to the mapping is what creates the fractals and due to their nature, CGR allows for pattern recognition as a mechanism for organism identification. Several other metrics can be obtained from the CGR mappings of different organisms and in turn be used for genome comparison. These include DNA centroid, probability distance, and kmer heat maps. Looking at the CGR of *Ustilago maydis*, a member of the Fungi kingdom we can see the pattern formed from CGR. In Figure 2, this organism shows less points toward the top of the CGR graph, showing a lack of AT within its genome.

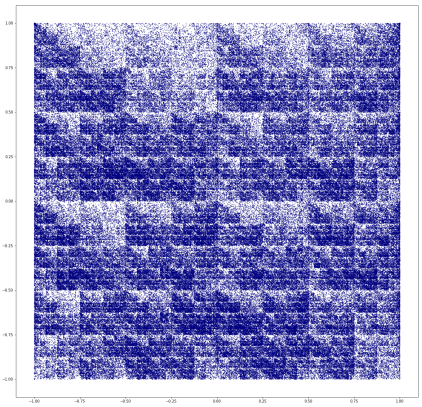
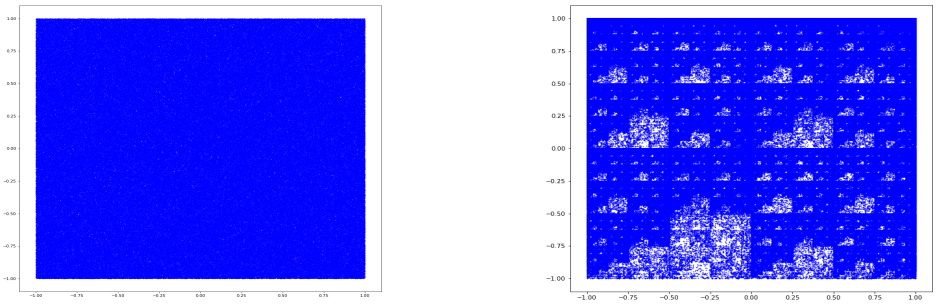


Figure 2: *Ustilago maydis* CGR

Initially we show that the pattern of CGR is unique for a particular organism and that a random sequence of DNA will not generate such a pattern. This is shown in Figure 3, the random sequence of DNA, 3a has no real distinct pattern or structure whereas the sequence of human chromosome 21, 3b shows a distinct double scoop pattern which is common in the CGR of vertebrates. The double scoop pattern was confirmed by Nick Goldman [3]. This also shows that while CGR can create fractals, it is not guaranteed to occur with every organism. Patterns differ between organisms as well which is shown in Figure 2. We note that the patterns formed by the CGR of prokaryotes differs from that of eukaryotes due to less variation in DNA.



(a) Random Sequence CGR (b) Human Chromosome 21 CGR

Figure 3: CGR

Python programming language was used for the implementation of CGR and other techniques applied to the DNA sequences. These include the calculation of frequency chaos game representation,centroids, difference between centroids and markov chain implementation. The sequences were obtained from the National Center for Biotechnology Information(NCBI) website.

2.1.3 Methods - DNA Centroid

The centroid of a cluster of points is the mean of those points and is denoted by

$$C_{ij} = \left(\frac{\sum_{k=1}^{n} x_k}{n}, \frac{\sum_{k=1}^{n} y_k}{n} \right)$$

, $0 \leq i, j \leq 9$. We first partition our CGR mapping of DNA into a 10 x 10 grid. This gives 100 cells of points and the centroid is then calculated for each cell. The distance between the centroids of a two CGR mappings is

denoted as

$$d_{ij} = \sqrt{(x - x^{'})^2 + (y - y^{'})^2}$$

where x,y are the coordinates of the centroid at position i,j in a CGR mapping and x',y' are the coordinates of the centroid at that same position of another CGR mapping. The resulting distance from the 100 comparisons is then summed into,

$$D = \sum_{i=0}^{9} \sum_{j=0}^{9} d_{ij}$$

The method of calculating the centroid of CGR was previously applied by [30]

2.1.4 Methods - FCGR

Frequency chaos game representation (FCGR) is a method of graphically representing the kmers of a genome. A kmer is a nucleotide of length k and there are 4^k possible kmers. The probability of kmers can be used for frequency chaos game representation. Initially, the CGR mapping is divided into a 2^k x 2^k grid populated by frequency or probability of kmers. This is denoted by

$$p_{ij} = \frac{\text{number of kmer occurrences}}{\text{total number of kmers}}$$

,$0 \leq i, j \leq 4^k - 1$, k = kmer length. Using the distance denoted by

$$d_{ij} = |p_{ij} - p_{ij}^{'}|$$

, $i, j \geq 0$. D is found by summing up all d_{ij} as the same done for DNA centroid. A sequence of DNA is read through and a count is kept of each possible kmer, or nucleotide sequence. An example of FCGR representation of the sequence CACGTTA is shown step by step below in Figure 4. A similar approach was used in 2016 [31].

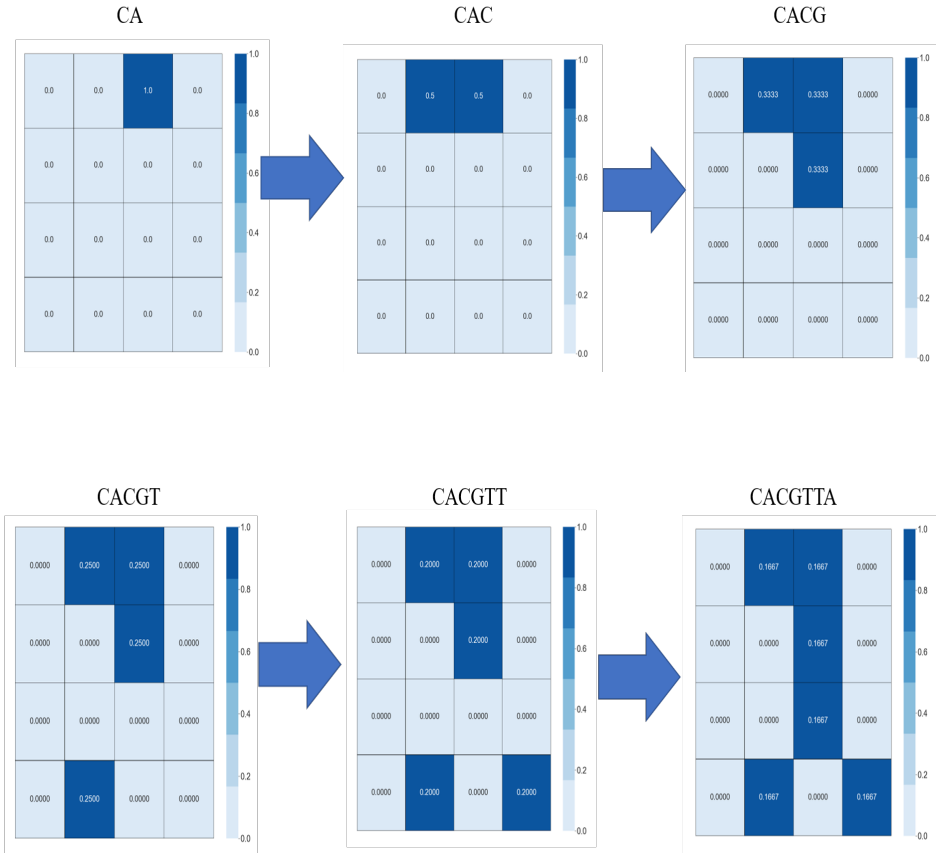


Figure 4: FCGR of CACGTTA

The FCGR of Severe Acute Respiratory Syndrome Cornonavirus 2 (SARS-COV2) Wuhan is shown in Figure 5.

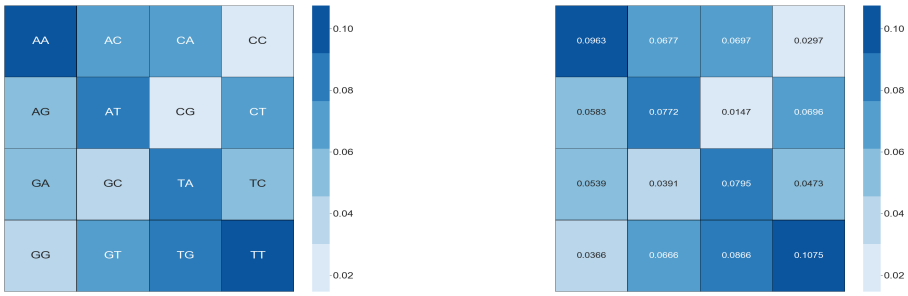


Figure 5: SARS-COV2 Wuhan 2mer Heat Map

2.1.5 Methods - Markov Model

Markov models are used as a means for predicting the state in a system given certain transition probabilities. These models are special because they exhibit the Markov property, which is that the transition to the next state is based solely on the current state and not any states prior. For DNA, the states are the kmers of length k. The order of a Markov model is denoted by m, where the m preceding residues determine the probability of each residue, r_i at position i

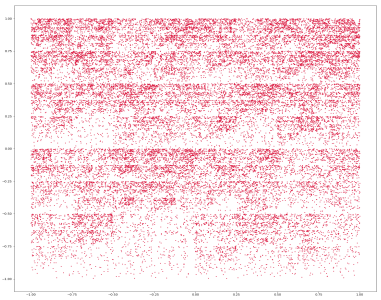
$$= P(r_i|S_{i-m,i-1}) = \frac{F(r_i|S_{1..m})}{\sum_{j\in A} F(r_j|S_{1..m})} = \frac{F(S_{1..m}r_i)}{\sum_{j\in A} F(S_{1..m}r_j)}$$

, $1 \leq i$, $m \geq 0$. r_i is the suffix while $S_{i-m,i-1}$ is the prefix. These probabilities are used to populate the transition matrix, which is necessary for determining the next state of the system. A model of order 0 is referred to as a Bernoulli model and this model assumes independence between successive nucleotides

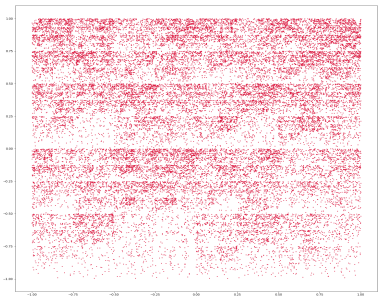
and uses the probability of kmers of length one. The Bernoulli model is simple, but not realistic when it comes to patterns in DNA. To find the genomic signature, a higher order markov model is needed. A model of order 1 uses the probabilities of the prefixes of length 1 to determine the resulting suffix, while a 2nd order model uses prefixes of length 2. The frequency of kmers of higher lengths can in turn be predicted by training the background markov model with a portion of the organisms genome.

2.1.6 Results - Similarities in CGR

When comparing the CGR of several different genomes we tend to find a distinguishable difference from the CGR of eukaryotes, prokaryotes, and viruses. Below are the CGR graphs of SARS COV2 Wuhan, SARS COV2 HKU, *Cricetulus griseus*, *Rousettus aegyptiacus*, *Natronomonas pharaonis*, and *Haloferax volcanii* in Figures 6, 7, and 8 respectively. We notice a low GC content in the two viral strains and their graphs look almost identical. Next the CGR of eukaryotes shows the double scoop pattern mentioned earlier as well as a lack of GC content within the genome. In comparison the CGR graphs of prokaryotes show a lack of AT content with a high content of GC.

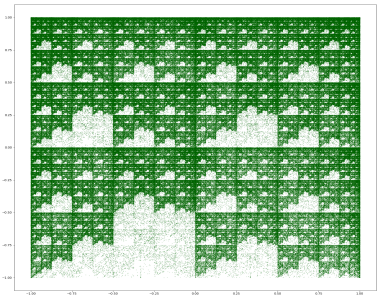


(a) SARS COV 2 Wuhan CGR

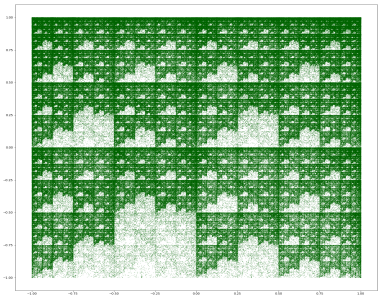


(b) SARS COV 2 HKU CGR

Figure 6: Viruses

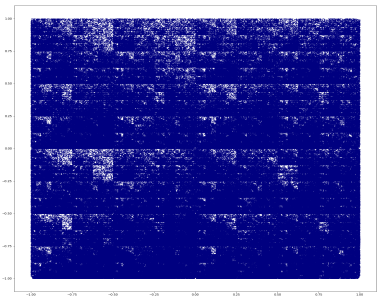


(a) *Cricetulus griseus CGR*

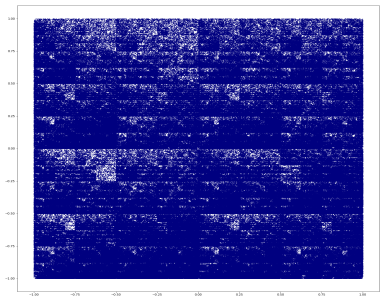


(b) *Rousettus aegyptiacus CGR*

Figure 7: Eukaryotes



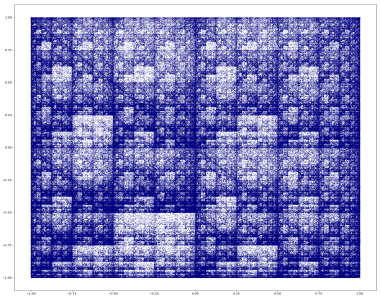
(a) *Natronomonas pharaonis CGR*



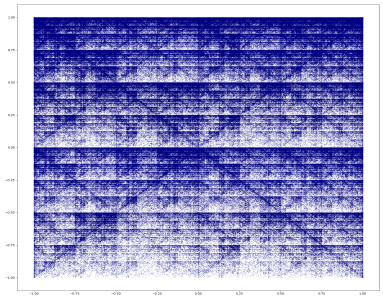
(b) *Haloferax volcanii CGR*

Figure 8: Prokaryotes

It is important to note that two eukaroytes can have vastly different CGR graphs. For example, take the CGR of *Bos taurus* on the left and *Candida albicans* on the right in Figure 9. This is mostly due to the differences in kingdoms of the two organisms as *Bos taurus* belongs to Animalia while *Candida albicans* belongs to the kingdom Fungi.



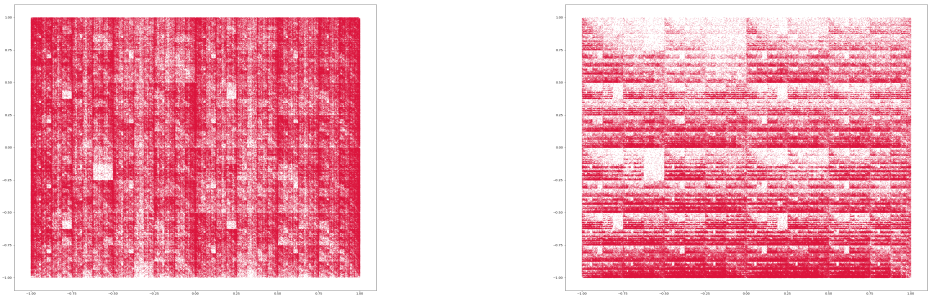
(a) *Bos taurus CGR*



(b) *Candida albicans CGR*

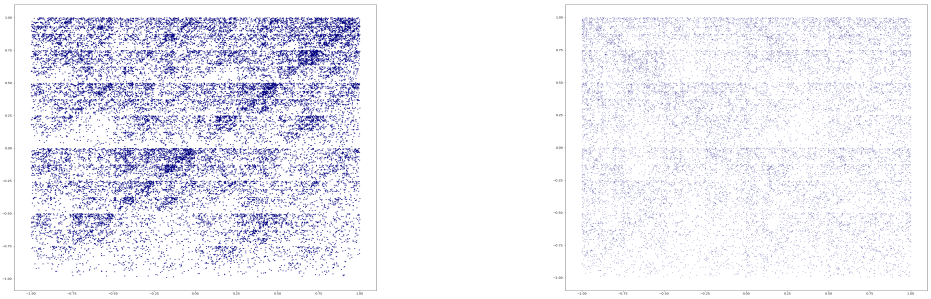
Figure 9: Differences in CGR of Eukaryotes

These differences can also be seen in the CGR of prokaryotes belonging to different kingdoms. *Archaeoglobus fulgidus* is a memeber of the Archaea kingdom and *Acidovorax citrulli* is a bacteria. Both organisms are prokaryotes, yet their graphs in Figure 10 are different. Lastly we show the differences in the CGR of viruses in Figure 11.



(a) *Archaeoglobus fulgidus CGR* (b) *Acidovorax citrulli CGR*

Figure 10: Differences in CGR of Prokaryotes

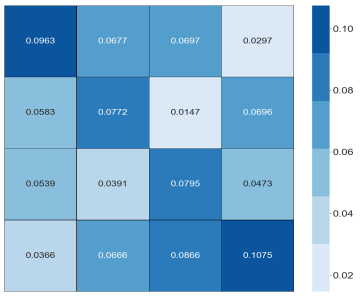


(a) Human Coronavirus 229E CGR (b) Ebola CGR

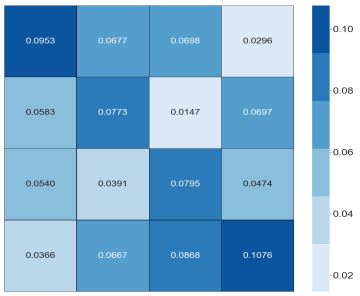
Figure 11: Differences in CGR of Viruses

2.1.7 Results - Similarities in FCGR Heat Maps

The FCGR graphs also show similarities for comparison. Figures 12a, 12, and 13 show the FCGR heat maps of SARS COV2 Wuhan, SARS COV2 HKU, *Cricetulus griseus*, *Rousettus aegyptiacus*, *Natronomonas pharaonis*, and *Haloferax volcanii* respectively.



(a) SARS COV2 Wuhan FCGR



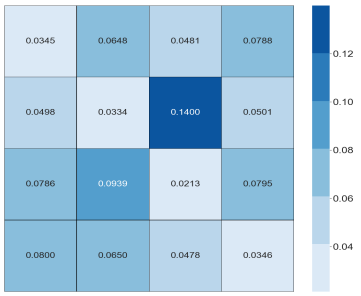
(b) SARS COV2 HKU FCGR

Figure 12: FCGR of Viruses

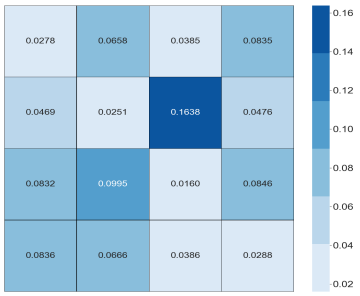
(a) *Cricetulus griseus FCGR*

(b) *Rousettus aegyptiacus FCGR*

Figure 13: FCGR of Eukaryotes



(a) *Natronomonas pharaonis FCGR*



(b) *Haloferax volcanii FCGR*

Figure 14: FCGR of Prokaryotes

The FCGR graphs for prokaryotes show both organisms have a high CG content. In contrast, the FCGR graphs of Eukaryotes show an abundance of the 2mers AA, AT, TA, and TT in their genomes whie CG content is the lowest amongst all other 2mers. The FCGR of viruses tells a similar story of an abundance of AA, AT, TA, and TT with a lack of CG content.

2.1.8 Results - DNA Centroid Comparison

Comparison of the centroids of several genomes was useful for distinguishing between them. In Figure 15, we see that the centroids of the two prokaryotes were both over 1.6 away from the two eukaryotes. Also, the two strains of virus show a distance greater than 1 from the eukaryotes. In fact, the two strains of virus show very similar distances when compared to all other genomes. The lines in the graph are almost on top of each other. We can see that both eukaryotes and prokaryotes show similar distances from the

viruses. Overall, the seperation necessary for classification can be seen when comparing the centroids of genomes.

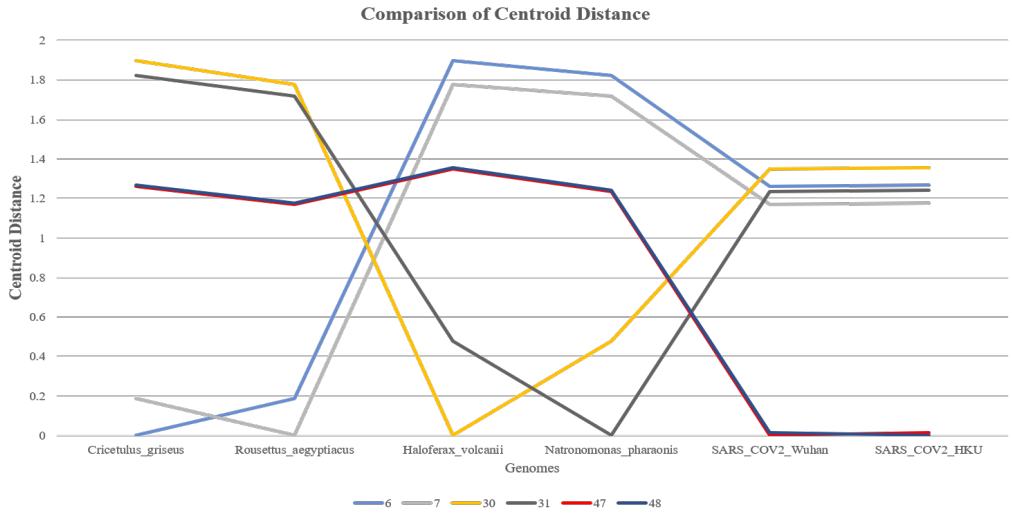


Figure 15: DNA Centroid Comparison

2.1.9 Results - Probability Distance

The probability distance was a useful method for distinguishing between genomes. The below figure shows the probability distance matrix for the genomes of several viruses and bacteria. Theses viruses are listed in Figure 17.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2	0.25197542	0													
3	0.2475856	0.011414	0												
4	0.21949078	0.13784	0.132572	0											
5	0.21949078	0.148376	0.143108	0.021071	0										
6	0.3213345	0.305531	0.302897	0.366989	0.37928	0									
7	0.32396839	0.307287	0.304653	0.367867	0.380158	0.007024	0								
8	0.21773486	0.08604	0.088674	0.110623	0.119403	0.295874	0.297629	0							
9	0.48902546	0.350307	0.345917	0.415277	0.424056	0.459175	0.455663	0.429324	0						
10	0.34152766	0.172081	0.172959	0.244952	0.251975	0.410009	0.411765	0.222125	0.330114	0					
11	0.40386304	0.242318	0.246708	0.304653	0.312555	0.438104	0.438104	0.28446	0.280948	0.138718	0				
12	0.47234416	0.303775	0.305531	0.334504	0.337138	0.529412	0.528534	0.351185	0.330114	0.149254	0.152766	0			
13	0.36698859	0.356453	0.352063	0.38367	0.39245	0.165057	0.162423	0.339772	0.491659	0.474978	0.496927	0.597893	0		
14	0.33187006	0.348551	0.345917	0.365233	0.373134	0.187006	0.188762	0.333626	0.523266	0.466198	0.498683	0.602283	0.085162	0	
15	0.27216857	0.237928	0.234416	0.267779	0.279192	0.175593	0.176471	0.208077	0.448639	0.385426	0.437226	0.53029	0.218613	0.238806	0

Figure 16: Probability Distance Matrix

ID	Genome
1	Escherichia_coli
2	SARS_COV2_HKU
3	SARS_COV2_Wuhan
4	MERS_NL140455
5	MERS_NL13892
6	InfluenzaA_H1N1_England673
7	InfluenzaA_H1N1_England719
8	SARS_COV
9	Haemophilus_influenzae
10	Human_Coronavirus_229E
11	Human_Coronavirus_OC43
12	Human_Coronavirus_NL63
13	HIV_1
14	HIV_2
15	Ebola

Figure 17: Genome ID

From the table we can see that similar probabilities tend to be relatively close to each other in terms of distances. SARS COV2 HKU, SARS COV, SARS COV are all within 0.1 of each other which is due to how close relation of the strains. They also show close relation with MERS NL140455, MERS

NL13892, and Human Cornonavirus 229E as they are within 0.2. Both strains of HIV showed very similar probabilites as well. The two strains of influenzae showed a distance of 0.007. Overall, shorter distances equate to similar genomes.

2.1.10 Results - Markov Model

Using the transition probabilities shown in Figure 18, a Markov model of order 1 was created. The first 10,000 nucleotides of the sequence were used to train the model on 2mer frequency. The nucleotides of the sequence are selected based on the transition probabilities and the actual frequency of 3mers is compared with the test 3mer frequency. This is seen in the figure below.

Prefix/Suffix	A	C	G	T
A	0.330217	0.161937	0.211352	0.296494
C	0.34525	0.241062	0.02094	0.392748
G	0.295495	0.21982	0.190991	0.293694
T	0.248151	0.187518	0.199053	0.365277

Figure 18: Transition matrix of *Rousettus aegyptiacus*

0.036707	0.012903	0.019004	0.030306	0.018904	0.011602	0.0015	0.016503
0.021204	0.013903	0.011202	0.017003	0.024205	0.013203	0.017704	0.033707
0.019204	0.012603	0.016503	0.019304	0.016003	0.012703	0.0007	0.017804
0.0014	0.0008	0.0009	0.001	0.018204	0.016703	0.016603	0.025405
0.016703	0.008202	0.011902	0.012402	0.013503	0.007401	0.0009	0.014803
0.008302	0.006701	0.007101	0.009702	0.011802	0.008802	0.013003	0.015303
0.026305	0.014803	0.015903	0.026805	0.019104	0.015503	0.001	0.027806
0.018304	0.015203	0.012603	0.021204	0.029706	0.024705	0.020004	0.04911

Figure 19: Actual 3mer frequency of first 10,000 nucleotides of *Rousettus aegyptiacus*

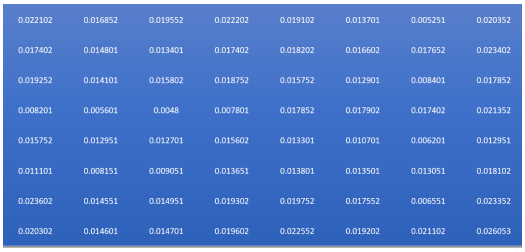


Figure 20: Test 3mer frequency of 10,000 nucleotides

We find our markov model to be simple, but on the right path in terms of predicting 3mer probability. By comparing the above probabilities we find a distance of 0.2111. This distance isn't too far from the actual genome but more training of the model is needed to further decrease this distance.

2.1.11 Conclusion

It has been shown that the pattern of CGR can help to classify organisms based on how diverse these graphs are between eukaryotes, prokaryotes, and viruses. The eukaryotes CGR showed recognizable patterns with some having the double scoop feature. Prokaryotes show very different CGR graphs as a pattern can be hard at times to recognize depending on the genome. The CGR of some viruses shows a somewhat similar pattern to the organisms they attack. In the case of SARS COV2, both strains showed a lack in GC content which is also found in the CGR of human chromosome 21. FCGR also proved useful for comparing genomes, but a longer kmer may prove necessary for better identification. DNA centroid calculations did allow for distinguishing between the organisms, but more is needed to be done in opti-

mizing this. Some of these include adjusting the length of the sequences being used as well as applying different partitioning to the CGR graph. Probability distance was also helpful for comparisons. A longer kmer however could help better correlate the distance with the genomes. Finally, a higher order markov model will be applied moving forward to more accurately predict kmer frequency within genomes.

2.2 Publication

Chaos Game Representation (CGR) was first proposed by H. J. Jeffrey in 1990 [1] as a novel scale independent graphical representation of a biological sequence. This representation is created through the use of an iterative process in which a one dimensional sequence is converted into a two dimensional array of points within a confined space. This method has proved to be helpful in the area of bioinformatics as it allows for more efficient sequence storage as well as sequence identification through pattern recognition of the CGR image. The efficient storage is due to the iterative process that allows for an entire sequence to be obtained through the last coordinate of the CGR. In Jeffrey's report CGR was applied to DNA sequences, so the string of characters being used included adenine (A), cytosine (C), guanine (G), and thymine (T) [1].

2.2.1 Background

DNA sequences analysis, is one of the most important parts of bioinformatics, which was considered to reveal the essence of all life phenomenon, has been developing rapidly in recent years. Sequence comparison is crucial to understand the evolutionary relationships among organisms. Using the analysis of the similarity/dissimilarity of biological sequences has been shown useful in understanding organisms [6]. CGR has mostly been restricted to a visualization tool representing nucleotide sequences, in which patterns like over-or underrepresentation of nucleotides, dinucleotides, trinucleotides, etc. can be visually ascribed. Goldman concluded that the patterns exhibited by CGR are sufficient to evaluate word length composition of three, i.e., the frequencies of nucleotides, dinucleotides and trinucleotides. However, it was shown later that longer oligonucleotide frequencies also influence the patterns seen in CGR. Later, a spectrum of word lengths, in addition to nucleotide and dinucleotide, in CGRs were identified as factors that can differentiate between genomes of different species. Several distance measures were proposed to compare two or more CGRs and it was employed for studying phylogenetic relationships among diverse species. However, it is not clear if intra-species genomic variability, which is much less than between-species variation, can be resolved using CGRs with similar word lengths. Later it was found in, the value k = 7 achieved the highest accuracy scores for HIV-1 subtypes classification [29].

CGR was performed on complete genomes of 15 corona viruses and two

alignment-free methods emerged which clustering analysis was applied to create a phylogenetic tree. A list of these viruses is found in Table 12. To each DNA sequence we associate a matrix then define distance between two DNA sequences to be the distance between their associated matrix. These methods are being used for phylogenetic analysis of coronavirus sequences. Our approach provides a powerful tool for analyzing and annotating genomes and their phylogenetic relationships. We also compare our tool to ClustalX algorithm which is one of the most popular alignment methods. Our alignment-free methods are shown to be capable of finding closest genetic relatives of coronaviruses. The two methods, probability matrix method and centroid matrix method are combined with CGR to construct distance matrix between two genomes, and then create dendrogram using Hierarchical Agglomerative Clustering (HAC) analysis. Our dendrogram can accurately identify the genetic relationship of different biology, and this method is generally applicable to various organisms [29].

2.2.2 Methods

The method we used to analyze and classify the 15 sequences of the dataset has three steps: 1) generate graphical representations (images) of each DNA sequence using CGR and define FCGR probability matrix and CGR centroid method using the features of CGR; 2) compute all pairwise distance to obtain two distance matrices; and 3) create the dendrogram of the distance matrix using Hierarchical Agglomerative Clustering (HAC) analysis. CGR is an

Virus name	NCBI/GISAID Accession number
1) hCov-19/bat/Yunnan	EPI_ISL_412976
2) hCov-19/pangolin/Guangdong	EPI_IS_410721
3) hCov-19/bat/Yunnan/RaTG13	EPI_ISL_402131
4) hCov-19/India	EPI_ISL_431117
5) hCov-19/Italy	EPI_ISL_417446
6) hCov-19/Iran	EPI_ISL_437512
7) hCov-19/Spain	EPI_ISL_428684
8) hCov-19/USA	EPI_ISL_431086
9) hCov-19/Wuhan	EPI_ISL_412980
10) Human Coronavirus-229E	KF-514433
11) Human Coronavirus-HKU1	KF-430201
12) Human Coronavirus-NL63	KF-530114
13) Human Coronavirus-OC43	KF-530099
14) SARS-Cov	NC_004718
15) MERS	KT-026456

Table 1: Dataset for experiment

iterative method introduced by Jeffery [1] to visualize the structure of a DNA sequence. A CGR associates an image to each DNA sequence as follows: starting from a square with corner labeled four nucleotides C, G, A and T, and the center of the square as the starting point, the image is obtained by successively plotting nucleotide as the middle point between the current point and the corner labeled by the nucleotide to be plotted. If the generated square image has a size of 2^k 2^k pixels, then every pixel represents a distinct k-mer: A pixel is color red if the k-mer it represents appears in the DNA sequence, otherwise it is white. CGR images of generating DNA sequences coming from various species show pattern such as squares, parallel lines, rectangles, triangles, and also complex fractal patterns. We have created

CGR of all 15 virus genomes and visually they look similar (see Figure 1 below). For step (1), we will use a slight modification version of the original CGR, a k-th order FCGR (Frequency Chaos Game Representation) is a 2^k 2^k matrix that can be constructed by dividing the CGR plot into a 2^k 2^k grid, and defining the element $|a_{ij}|$ as the number of points that are situated in the corresponding grid square. A first-order FCGR and a second-order FCGR have the structure shown below, where N_w is the number of occurrences of the k-mer w, in the sequence s is

$$FCGR_1(s) = \begin{pmatrix} N_C & N_G \\ N_A & N_T \end{pmatrix} \text{ and } FCGR_2(s) = \begin{pmatrix} N_{CC} & N_{GC} & N_{CG} & N_{GG} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AA} & N_{TA} & N_{AT} & N_{TT} \end{pmatrix}$$

The (k+1)th order $FCGR_{k+1}(s)$ can be obtained by replacing each element N_X in $FCGR_k(s)$ with four elements $\begin{pmatrix} N_{CX} & N_{GX} \\ N_{AX} & N_{TX} \end{pmatrix}$ where X is the sequence of length k over the alphabet {A,C,G,T}. For each $k \geq 1$, we can define a probability matrix of $FCGR_k(s)$ by taking each entry of $FCGR_k(s)$ dividing by the total counts of all k-mers. We denote the FCGR probability matrix by $(P_ij), 1 \leq i, j \leq 2^k$. Note that $\sum_{i,j} P_{ij} = 1$. Probability matrix can be interpreted as probability of distribution.

Since the CGR captures the information of the whole genome data, extracting the global features from the CGR may not be efficient enough to

distinguish the genomes. In CGR Centroid method, we concentrate on extracting the local features as shown in [30]. We partition the CGR into sub-regions so that it re- veals local information of the interested areas. If two dots are within the same quadrant, they correspond to sequences with the same last mononucleotide; if they are in the same sub-quadrant, the sequences have the same last dinucleo- tides; and so on. This can demonstrate the structure of the sequences yielding the points in the CGR. Chaos Centroid method utilizes this biological signi- ficance by computing the centroid of the distributed points of each sub-region.

For Chaos Centroid method, the CGR is partitioned into 1010 equal subre- gion. The choice of 10 is to minimize the computation time. For each partition, we compute the centroidas follows. Let $(x_k, y_k$ be the coordinates of a point in the CGR. We define the centroid in each of the 1010 grid as follows:

$$c_{ij} = (\frac{\sum_{k=1}^{|a_{ij}|} x^k}{|a_{ij}|}, \frac{\sum_{k=1}^{|a_{ij}|} y^k}{|a_{ij}|}), 1 \leq i, j \leq 10.$$

For step (2), after computing FCGR probability matrices and computing cen- troid for each of the sequences in the dataset, the goal was to measure "distance" between two CGR images. There are many distances as it is given in [31],[30] that can be defined for our purpose. One of the goals of this study was to identify what distance is better able to differentiate the structural differences of various genomic DNA sequences. In this paper we use two different distances: FCGR Probability Matrix distance and CGR Centroid

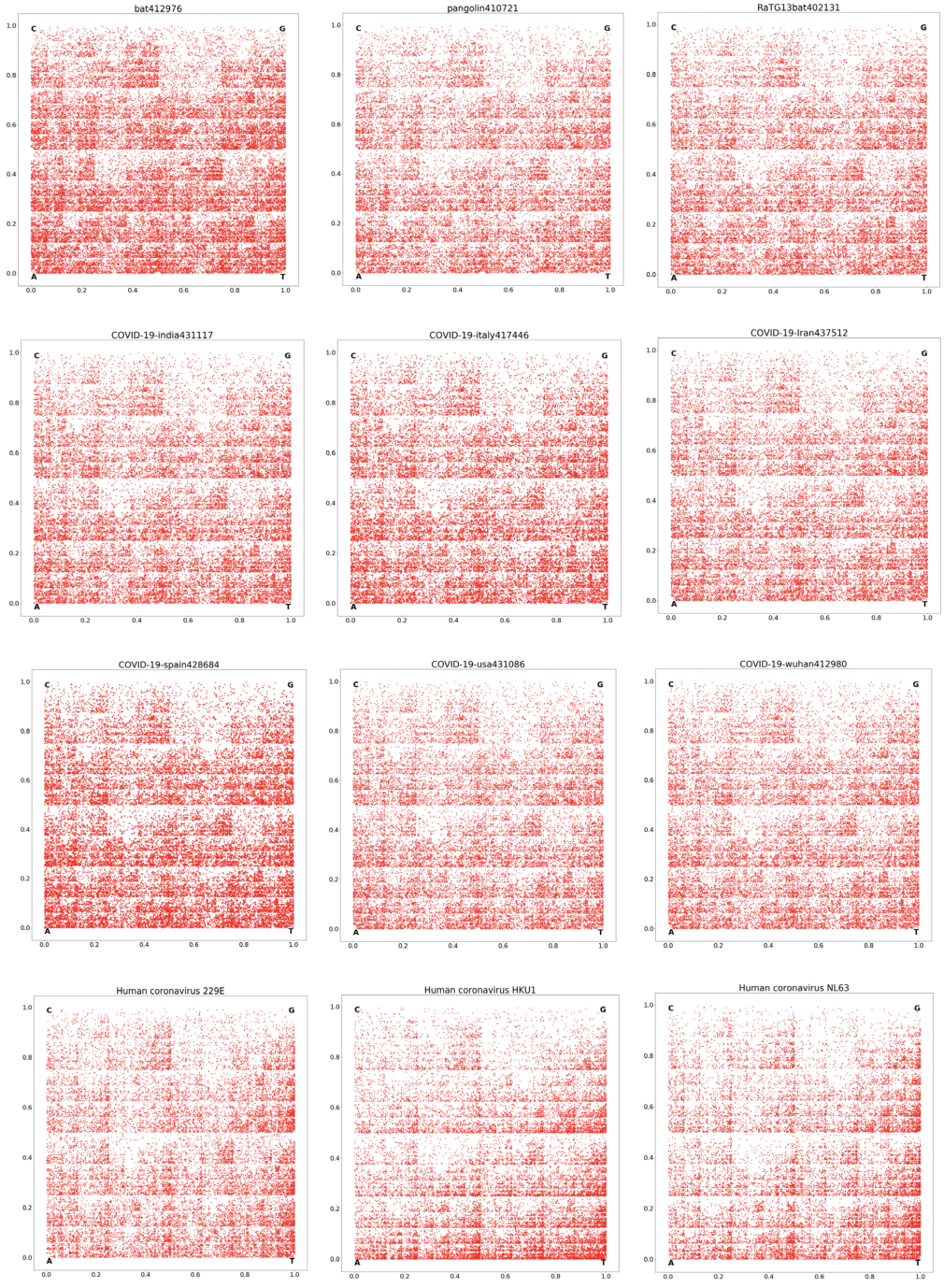


Figure 21: CGR of some of the viruses from Table 12

distance. Both use the Euclidean distance. For step (3), after computing all pairwise distances we obtained two different distance matrices. Then, we created the dendrogram of the distance ma- trices using Hierarchical Agglomerative Clustering (HAC) analysis.

In this section we formally define each of two distances. For two FCGR probability matrices (p_{ij}) and (p'_{ij}) we define $d_{ij} = |p_{ij} - p'_{ij}|$. The distance between two probability matrices denoted by $D_{PM} = \sum_{i=1}^{2^k} \sum_{j=1}^{2^k} d_{ij}$. For two genomes, we calculate 100 centroids $c_{ij} = (x_{ij}, yij)$ and $c'_{ij} = (x'_{ij}, y'_{ij})$ respectively for $1 \leq i, j \leq 10$. Then we found Euclidean distance between them $d_{ij} = \sqrt{(x_{ij} - x'_{ij})^2 + (y_{ij} - y'_{ij})^2}$. Then calculated the centroid distance between two genomes denoted by $D_{cd} = \sum_{i=1}^{10} \sum_{j=1}^{10} d_{ij}$.

2.2.3 Results

For our dataset we used k = 7, that is, each DNA sequence represented as a $2^7 \quad 2^7$ FCGR matrix. In [32], it was found highest accuracy in HIV-1 classification and this value is being used here as it is relevant for our viral analysis. Table 2 display the pairwise distance among 15-virus genomes in the dataset using probability matrix distance while Table 3 display the same using centroid distance.

Table 2. Probability distance matrix of 15 viruses listed in Table 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	0.3079														
3	0.4900	0.4606													
4	0.5129	0.6301	0.6303												
5	0.7076	0.7548	0.7506	0.7436											
6	0.7342	0.7737	0.7602	0.7969	0.7858										
7	0.8657	0.8700	0.8443	0.9420	0.8850	0.8406									
8	0.8074	0.8299	0.8037	0.8828	0.8587	0.7247	0.7237								
9	0.7578	0.7904	0.7744	0.8132	0.7894	0.7612	0.7067	0.7470							
10	0.4920	0.7671	0.2929	0.6313	0.7441	0.7714	0.8531	0.8123	0.7846						
11	0.4947	0.4750	0.0600	0.6408	0.7608	0.7614	0.8519	0.8029	0.7827	0.3143					
12	0.4930	0.4677	0.0321	0.6341	0.7553	0.7602	0.8477	0.8028	0.7783	0.3024	0.0299				
13	0.4905	0.4644	0.0180	0.6311	0.7529	0.7601	0.8456	0.8032	0.7757	0.2972	0.0492	0.0200			
14	0.4901	0.4646	0.0179	0.6318	0.7524	0.7595	0.8451	0.8030	0.7748	0.2978	0.0530	0.0254	0.0168		
15	0.4907	0.4623	0.0095	0.6306	0.7514	0.7599	0.8444	0.8037	0.7748	0.2953	0.0583	0.0320	0.0192	0.0192	

Table 3. CGR Centroid distance matrix of 15 viruses listed in Table 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	0.4531														
3	0.5567	0.4439													
4	0.5408	0.6281	0.6188												
5	0.9029	0.9255	0.8784	0.7598											
6	0.8845	0.8718	0.8409	0.8615	0.8762										
7	1.4297	1.3203	1.2682	1.3924	1.300	1.2339									
8	1.2246	1.0924	1.0161	1.2011	1.200	0.9157	0.9635								
9	1.0256	0.9862	0.9295	0.9869	0.9310	0.8623	0.9538	0.9123							
10	0.5581	0.4575	0.3303	0.6356	0.9271	0.8824	1.2759	1.0163	0.9912						
11	0.5915	0.4816	0.1350	0.6525	0.9115	0.8667	1.2682	1.0432	0.9391	0.3694					
12	0.5654	0.4591	0.0969	0.6312	0.8839	0.8518	1.2604	1.0403	0.9217	0.3446	0.0670				
13	0.5607	0.4576	0.0702	0.6247	0.8837	0.8450	1.2644	1.0326	0.9291	0.3367	0.1156	0.0636			
14	0.6113	0.5127	0.1596	0.6785	0.9097	0.8583	1.3064	1.0584	0.9613	0.3859	0.2254	0.1793	0.1558		
15	0.5460	0.4416	0.0454	0.6167	0.8783	0.8332	1.2680	1.0221	0.9235	0.3290	0.1295	0.0943	0.0721	0.1586	

Figure 22 shows the phylogenetic tree obtained using Table 2 distances by py- thon Hierarchical Agglomerative Clustering (HAC) analysis. Similarly Figure 23 shows the phylogenetic tree using Table 3. Figure 4 is the Neighbor Joining Phylogenetic tree using traditional Clustal X method. From

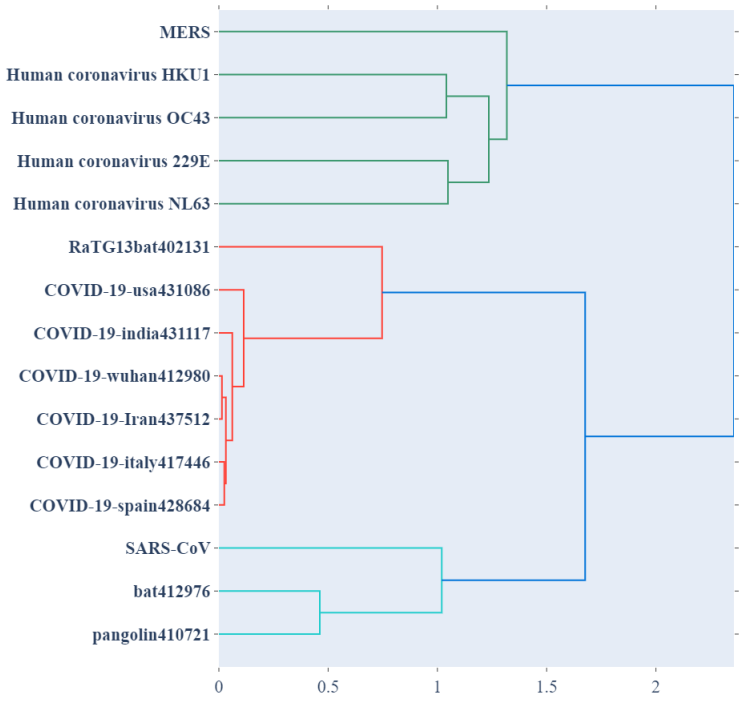


Figure 22: HAC phylogenetic tree using probability matrix distance.

Figure 22 and Figure 24, we can see that the cluster results between Clustal X method and probability distance method are essentially same. Similar Phylogenetic analysis of bat coronaviruses with other coronaviruses and the phylogenetic tree was constructed using Clustal W also done in [33].

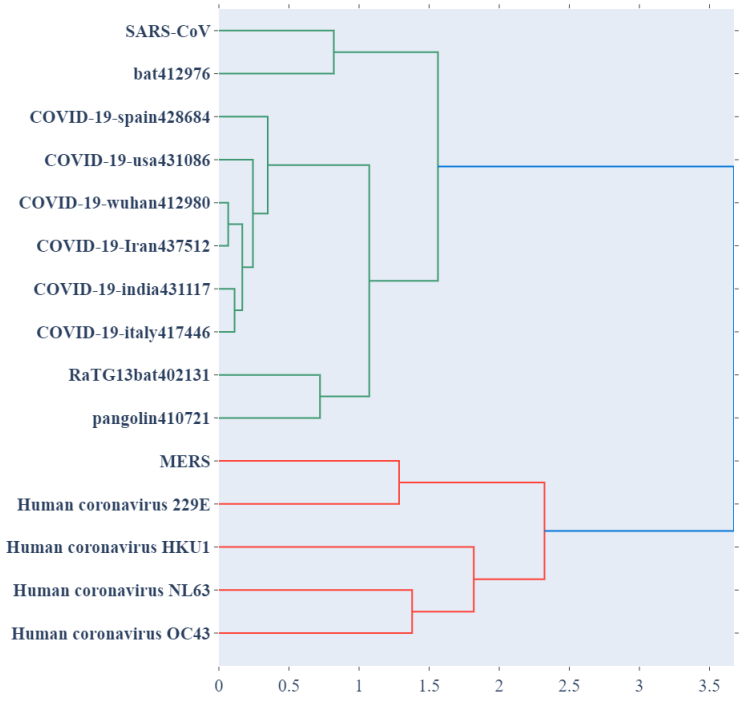


Figure 23: HAC phylogenetic tree using CGR centroid distance.

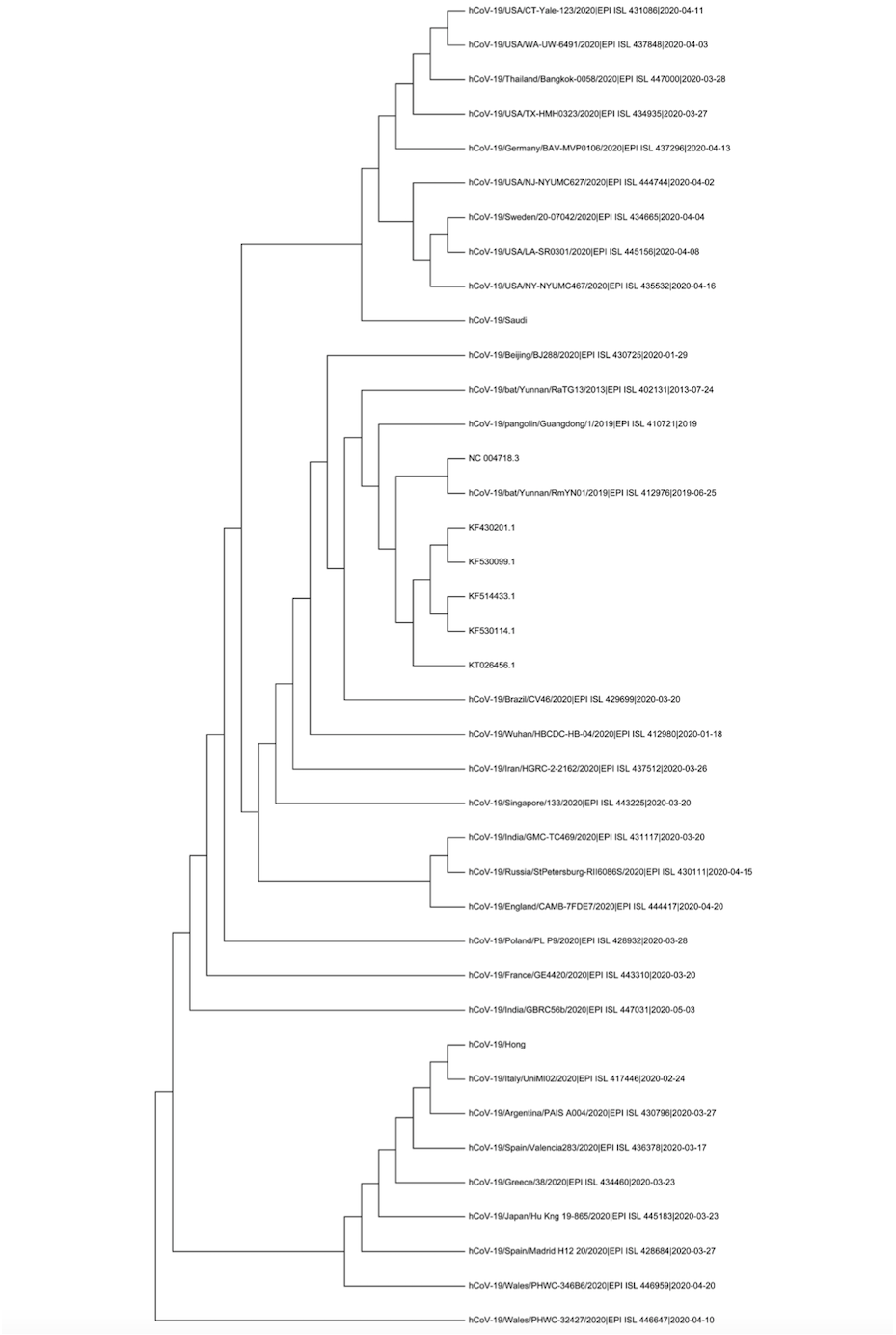


Figure 24: Phylogenetic Tree was created by Clustal X by aligning 15 DNA sequences using Neighborhood Joining Method.

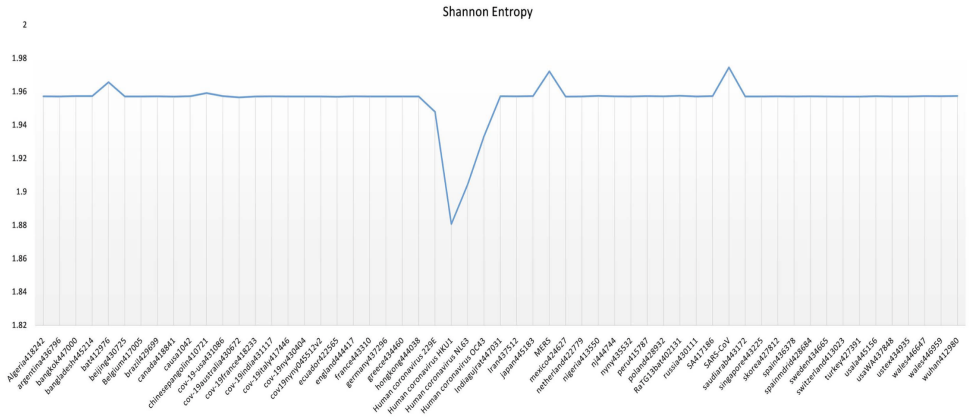


Figure 25: Shannon Entropy of 57-virus genomes.

All sequence data contain inherent information that can be measured by Shan- non's uncertainty theory. Measuring uncertainty may be used for rapid screen- ing for sequences with matches in available database, prioritizing computational resources, and indicating which sequences with no known similarities are likely to be important for more detailed analysis as seen in [34]. We started with 57 genome sequences and then reduced to 15 based on the Shannon Entropy and Shannon Entropy of 7-mers of the sequences, see Figure 25 and Figure 26. All Covid-19 sequences have entropy close to 1.957. We choose only six Covid-19 sequences in the dataset along with all other sequences with deviated entropy from 1.957 for our analysis of corona viruses.

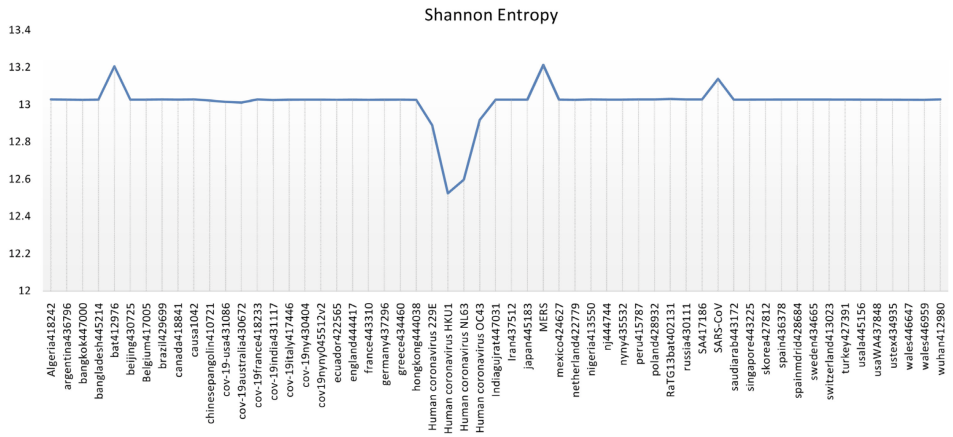


Figure 26: 7-mers Shannon Entropy of 57 virus sequences.

2.2.4 Discussion and Conclusions

Our methods are comparable to many other alignment-free methods as shown in [15], [30]. The proposed methods i.e. FCGR Probability and Chaos Centroid, are based on Chaos game representation, which provides a unique and scale-independent representation of DNA sequences through the statistical distribution of k-mers along DNA sequences. An advantage of CGR over alignment is that it has the potential to reveal the evolutionary and/or functional relationships between the sequences having no significant homology, as explained in [25]. Furthermore, it does not require prior knowledge of consensus sequences, nor does it involve exhaustive searches for sequences in databases. The limitation of CGR is that it takes a computational time to generate the representations from DNA sequences. In conclusion, results show that our method can accurately classify different genomic sequences.

In terms of classification accuracy, our method is basically the same as the state-of-the art Clustal X and compare with the traditional Clustal X phylogenetic tree construction method [18], our method is much faster. Furthermore, our dendrogram construct method can be widely applicable for various kinds of organisms. This research may contribute to reveal the biological evolu- tion process to some extent, as well as promote the further development of bio- informatics. We may make efforts in our future work to provide a webserver for the methods presented in this paper. All the codes in this paper are written in python and can be available upon request.

3 Chaos Game Representation (CGR)

Biological systems tend to have quite a bit of entropy or chaos. In order to represent such dynamical systems, one of the crux of statistical methods, chaos theory is applied [4]. Chaos theory helps to sort out such dynamical systems and lend potential information to better understand these processes. Since Jeffrey's report, several other applications of CGR to biological sequences have been explored including arbitrary sets of characters [26]. Other studies applied CGR for studying such dynamical systems using interger length resolutions [4]. Further use of the applications of CGR on protein sequences have shown promising results[25]. One obstacle is deciding how to represent the amino acid using CGR as there are 20 characters to represent as opposed to 4. Fiser [27] was one of the first to find a method to improve

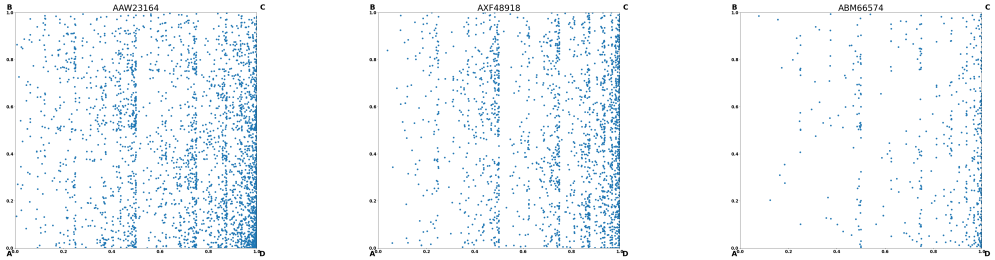
such techniques by creating a 20-sided polygon with each vertex representing one of the 20 amino acids. Another representation of the 20 amino acids was applied by Randic [28] in which the CGR exists within the unit circle. This approach ordered the amino acids alphabetically in comparison to organization based on their physiochemical properties. The properties of the amino acids serves as vital information for characterization of protein sequences and this was noted by Randic. Another arrangement of the amino acids was proposed by Basu [25] and included the separation of the 20 amino acids into 12 groups. His proposition also referred to as the 12-CGR was fruitful in sequence comparison. Bhoumik [2] utilized the 4-CGR method which places the amino acids into 4 groups based on their physiochemical properties. All of the previously mentioned methods are graphical representations and can be advantageous for sequence comparison [4], [6]. Goldman [3] noted that the frequency of nucleotides plays a role in determining the complex patterns in CGR of DNA. The similarity/dissimilarity of sequences has also been successful in genome comparison as such vectors can be used for representation of a group as opposed to an individual organism [6]. Other methods have been proposed in the field of bioinformatics, to study the features of viral sequences some of which include frequency chaos game representation (FCGR), positional distribution, and adjacency vectors [3],[5],[24].

3.1 CGR of Proteins

To create a CGR graph, we first began with an initial point $(0.5, 0.5)$, the center of a unit square in quadrant 1 of the xy-plane. Let the vertices of the unit square be: A = $(0, 0)$, B = $(0, 1)$, C = $(1, 1)$, and D = $(1, 0)$. The 20 amino acids are divided into four groups (A,B,C,D). Group A contains the negatively charged amino acids Aspartic Acid (D) and Glutamic acid (E). Group B consists of the positively charged amino acids Lysine (K), Arginine (R), and Histidine (H). Group C contains the neutral polar amino acids Serine (S), Threonine (T), Asparagine (N), Cysteine (C), Tyrosine (Y), and Glutamine(Q). Lastly, group D consists of the neutral non-polar amino acids Alanine (A), Glycine (G), Isoleucine (I), Leucine (L), Methionine (M), Phenylalanine (F), Proline (P), Tryptophan (W), and Valine (V). These vertices are arbitrary and can have any label, such as A, U, C, and G for RNA and in the case of DNA A, T, C, and G. We denote the next coordinate in the CGR graph,

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

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. A point is plotted half the distance from this vertex and the previous coordinate. Some examples of the CGR of several viruses used in this report are shown in figure 27.



(a) Dengue (b) Ebola (c) HTLV

Figure 27: CGR of Proteins

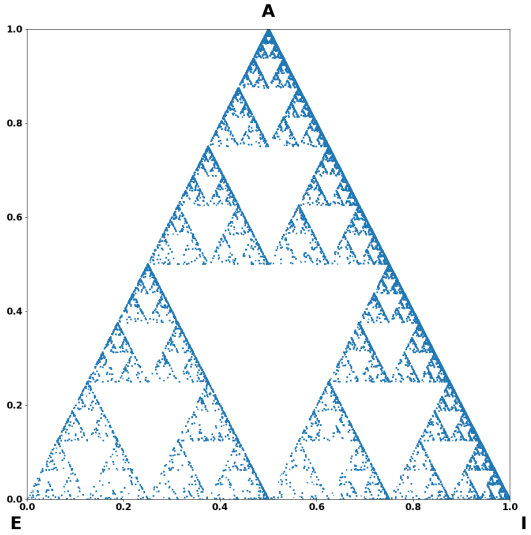


Figure 28: CGR of Bible

CGR can, but is it not always guaranteed to create fractals which are infinitely complex self-similar shapes on varying scales [35]. They are the images of dynamical systems and are driven by an ongoing feedback loop. Many common fractals include the Sierpinski Triangle, leaves, seashells, and snowflakes. The Sierpinski triangle was first described by Polish mathematician Waclaw Sierpinski, a leading figure in point set topology in 1915. An example of a Sierpinski Triangle created by looping through the Holy Bible is shown in Figure 28. We denote the occurrences of the letters A, I, and E by mapping a point half the distance to their vertex. Vertex A is located at $(0.5, 1)$, vertex I is located at $(1, 0)$ and vertex E is located at $(0, 0)$ on the xy coordinate plane. The Sierpinski Triangle can also be made by repeatedly removing the middle triangle of an equilateral triangle. A diagram of this repeated process is shown in Figure 29. The nature of fractals allows for continuous magnification to gather more detail as they are infinitely complex. This magnification is limited to the processing power of the computer being used for magnification, which for current technology is about 10^16 or ten quadrillion. To better understand the CGR, Goldman [3] showed that the frequency of nucleotides plays a role in determining the complex patterns in CGR of DNA.

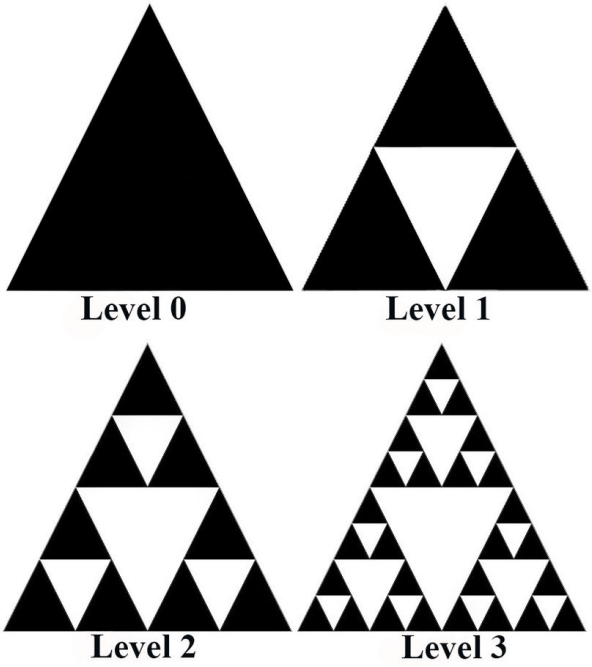


Figure 29: Sierpinski Triangle Creation

4 Methods

For this report, CGR was applied to protein sequences to distinguish between several species of viruses. It is used as a basis to obtain information about the viruses being studied. Extracting as much information from the CGR is one issue faced and is key in sequence identification and classification [24]. The protein sequences of the viruses were obtained from the National Center for Biotechnology Information (NCBI) website. 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 use them to better explain the protein sequence as a whole. To accomplish the goal of sequence identification, first a means of grouping the amino acids to allow for CGR was decided. The choice of Bhoumik's method [2] was made on the basis of the results obtained from this grouping. Other potential methods of grouping that have been previously studied, include Li's method [21] of a 12-sided polygon as well as a random grouping of 4 and 5 [2].

4.1 CGR Centroid

Once the the CGR is created for a protein sequence, the CGR square is divided into four cells. Each cell represents one of the four groups, $\{A_i, B_i, C_i, D_i; i = 1, 2, ..., n\}$ 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

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

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.

4.2 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.

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

4.3 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:

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

, $1 \leq i \leq j \leq 400$. 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

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

. Note that when t = 1, we have

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

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

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

, euclidean distance.

4.4 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). This is the same as the FCGR defined

previously with the only difference being the frequencies are defined for a group of amino acids as opposed to just one. The GFCGR is defined as follows:

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

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

4.5 Kullback-Lieber Discrimination Information

A previous method introduced by Li [8] utilized the Kullback-Lieber 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

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

. Under hypothesis 2 we have

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

. These distributions can be compared by using the Kullback-Lieber 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 it's 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 [8] noted that this method can accurately measure the dissimilarity between two sequences.

4.6 Compounded Frequency

Another method for sequence comparison that has been previously examined is the compounded frequency. This method was proposed by Almeida [4] 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.

4.7 Shannon Entropy

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

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

where $2merAAF = p_1, p_2, ...p_n, 1 \leq i \leq n$. 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.

5 Data and Results

The data sets shown in figures 2, 3, 4, 5 consists of the accession numbers of the 400 strains of 8 viral groups, so 50 strains per group.

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: HIV_1 & HIV_2 data sets

SARS_COV		SARS_COV2	
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: SARS_COV & SARS_COV2 data sets

MERS		Dengue	
AVN89429	AWH65952	QPZ88405	ANC57575
AID50417	AGR87639	QFS19562	ANC57576
ANC28665	$YP_0$07188577	QPB40131	ANC57581
AKM76247	QGW51400	QFS19150	ANC57582
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: MERS & Dengue 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
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: Ebola HTLV data sets

First we construct the CGR graph for all 400 viruses and calculate the 2mer AAF. Next the pairwise distances between each of the viruses is com-

puted for all of the previously mentioned methods. For the Shannon entropy, 2mer AAF and GFCGR the manhattan distance is used. From this several distance matrices are obtained, snapshots of these are shown in figures 30, 31, 32. The euclidean distance is applied to both the CGR centroids and CGR centroid bisections as shown in figures 33, 34 while $J(x, y)$ and Pearson correlation have the respective distance matrices 35, 36. MDS is then applied to the distance matrices to create 2D and 3D projections shown in figures 37, 38, 39, 40, 41, 42.

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.00315872	0													
QFS19150	0.00339468	0.00023596	0												
ACK28184	0.0034403	0.000281582	4.56221e-05	0											
QHR82546	0.00208286	0.00107586	0.00131182	0.00135745	0										
QCZ25008	0.00603372	0.002875	0.00263904	0.00259342	0.00395087	0									
QFS19149	0.00365417	0.000495449	0.000259489	0.000213867	0.00157131	0.00237955	0								
ACL99188	0.00467445	0.00151573	0.00127977	0.00123415	0.0025916	0.00135927	0.00102028	0							
QQC97219	0.00335267	0.000193949	4.20106e-05	8.76328e-05	0.00126981	0.00268105	0.000301499	0.00132178	0						
QPZ88403	0.0031922	3.34747e-05	0.000202485	0.000248107	0.00110934	0.00284153	0.000461974	0.00148226	0.000160475	0					
BBH51315	0.00315872	0	0.00023596	0.000281582	0.00107586	0.002875	0.000495449	0.00151573	0.000193949	3.34747e-05	0				
AEF01518	0.00348257	0.000323848	8.78881e-05	4.2266e-05	0.00139971	0.00255115	0.000171601	0.00119188	0.000129899	0.000290373	0.000323848	0			
AAW23164	0.00291568	0.000243047	0.000479007	0.000524629	0.000832816	0.00311805	0.000738496	0.00175878	0.000436997	0.000276522	0.000243047	0.000566895	0		
ANC57587	0.00474989	0.00159117	0.00135521	0.00130958	0.00266703	0.00128383	0.00109572	7.54344e-05	0.00139722	0.00155769	0.00159117	0.00126732	0.00183421	0	
QPZ88404	0.00427273	0.00111401	0.000878051	0.000832429	0.00218987	0.00176099	0.000618562	0.000401722	0.000920062	0.00108054	0.00111401	0.000790163	0.00135706	0.000477156	0
ANC57579	0.00438666	0.00122794	0.000991979	0.000946356	0.0023038	0.00164706	0.00073249	0.000287794	0.00103399	0.00119446	0.00122794	0.00090409	0.00147099	0.000363229	0.000113928
QPU83820	0.00320583	4.71071e-05	0.000188853	0.000234475	0.00112297	0.00282789	0.000448342	0.00146863	0.000146842	1.36324e-05	4.71071e-05	0.000276741	0.000290154	0.00154406	0.0010669
QPB40125	0.00239329	0.000765432	0.00100139	0.00104701	0.000310431	0.00364043	0.00126088	0.00228116	0.000959382	0.000798907	0.000765432	0.00108928	0.000522385	0.0023566	0.00187944
QFS19134	0.00283777	0.000320957	0.000556916	0.000602539	0.000754907	0.00319596	0.000816405	0.00183669	0.000514906	0.000354431	0.000320957	0.000644805	7.79092e-05	0.00191212	0.00143497
QPB40127	0.0037701	0.000611377	0.000375417	0.000329795	0.00168724	0.00226363	0.000115928	0.000904356	0.000417427	0.000577902	0.000611377	0.000287529	0.000854424	0.00097979	0.000502634
QGQ59492	0.00219024	0.000968485	0.00120444	0.00125007	0.000107379	0.00384349	0.00146393	0.00248422	0.00116243	0.00100196	0.000968485	0.00129233	0.000725438	0.00255965	0.0020825
ANC57577	0.00451514	0.00135642	0.00112046	0.00107484	0.00243228	0.00151858	0.000860969	0.000159315	0.00116247	0.00132294	0.00135642	0.00103257	0.00159946	0.000234749	0.000242407
ANC57580	0.00279299	0.000365729	0.000601689	0.000647311	0.000710135	0.00324073	0.000861177	0.00188146	0.000559678	0.000399203	0.000365729	0.000689577	0.000122681	0.0019569	0.00147974
QPI11922	0.0034092	0.000250474	1.4514e-05	3.11082e-05	0.00132634	0.00262453	0.000244975	0.00126526	5.65246e-05	0.000216999	0.000250474	7.33741e-05	0.000493521	0.00134069	0.000863537
QIB99387	0.00482913	0.00167041	0.00143445	0.00138883	0.00274627	0.00120459	0.00117496	0.000154677	0.00147646	0.00163694	0.00167041	0.00134656	0.00191346	7.92431e-05	0.000556399
ANC57586	0.00231819	0.000840537	0.0010765	0.00112212	0.000235327	0.00371554	0.00133599	0.00235627	0.00103449	0.000874011	0.000840537	0.00116438	0.000597489	0.0024317	0.00195455
QBQ58385	0.00362068	0.000461955	0.000225996	0.000180373	0.00153782	0.00241305	3.34932e-05	0.00105378	0.000268006	0.000428481	0.000461955	0.000138107	0.000705003	0.00112921	0.000652055
ANC57578	0.00173563	0.00142309	0.00165905	0.00170467	0.000347226	0.00429809	0.00191854	0.00293882	0.00161704	0.00145656	0.00142309	0.00174694	0.00118004	0.00301426	0.0025371
BBH51316	0.0034092	0.000250474	1.4514e-05	3.11082e-05	0.00132634	0.00262453	0.000244975	0.00126526	5.65246e-05	0.000216999	0.000250474	7.33741e-05	0.000493521	0.00134069	0.000863537
QPZ88405	0.00477832	0.0016196	0.00138364	0.00133802	0.00269546	0.0012554	0.00112415	0.000103868	0.00142565	0.00158613	0.0016196	0.00129575	0.00186265	2.84334e-05	0.000505589

Figure 30: Distance matrix of Shannon Entropy

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.088027	0													
QFS19150	0.0498082	0.0660377	0												
ACK28184	0.0534901	0.0641054	0.0193815	0											
QHR82546	0.0874346	0.0247642	0.0678066	0.0652843	0										
QCZ25008	0.0894778	0.0221108	0.0704599	0.0679138	0.0232901	0									
QFS19149	0.0474872	0.0648585	0.00235849	0.0182157	0.067217	0.0698703	0								
ACL99188	0.0511437	0.0641013	0.0170202	0.0212202	0.0670409	0.0679095	0.0146833	0							
QQC97219	0.0569097	0.0678066	0.0300708	0.0182136	0.071934	0.073408	0.0277123	0.0311462	0						
QPZ88403	0.102181	0.0683962	0.0908019	0.0900056	0.0742925	0.0716392	0.0902123	0.0876551	0.0955189	0					
BBH51315	0.0862793	0.00825472	0.0648585	0.0623395	0.0200472	0.0179835	0.0636792	0.0629234	0.0666274	0.0678066	0				
AEF01518	0.0605462	0.0711785	0.0287854	0.0265252	0.07412	0.0749844	0.0270383	0.0300619	0.0358583	0.0953109	0.0705831	0			
AAW23164	0.0558529	0.0641075	0.0217261	0.0229885	0.0676371	0.0696749	0.0193893	0.0200413	0.0323242	0.0894198	0.0629244	0.0294724	0		
ANC57587	0.0723219	0.0670487	0.054089	0.0536398	0.0717424	0.0702575	0.0535004	0.0536398	0.0588046	0.0888317	0.0652734	0.0607132	0.0571765	0	
QPZ88404	0.0898306	0.0518868	0.0766509	0.0764618	0.0595519	0.0568986	0.0760613	0.0735257	0.0819575	0.0294811	0.0518868	0.0817642	0.0764698	0.0741063	0
ANC57579	0.0546743	0.0670515	0.0270148	0.0247569	0.0723443	0.0732032	0.0252642	0.0247569	0.0346688	0.0917701	0.0664547	0.0371353	0.0294724	0.0589449	0.0794075
QPU83820	0.0545578	0.0666274	0.0241745	0.0123405	0.0695755	0.0710495	0.021816	0.0229177	0.0229953	0.0925708	0.0648585	0.028808	0.0246845	0.0535129	0.0778302
QPB40125	0.0904181	0.0477594	0.0754717	0.0758744	0.0560142	0.056309	0.0748821	0.0764748	0.0813679	0.0678066	0.0501179	0.082357	0.0735338	0.0699872	0.0536557
QFS19134	0.0874745	0.0459906	0.0701651	0.0705849	0.0560142	0.0510024	0.0695755	0.0694053	0.0754717	0.0518868	0.0465802	0.0758887	0.0729435	0.0693969	0.0294811
QPB40127	0.0868474	0.00412736	0.0654481	0.0629262	0.026533	0.0238797	0.0642689	0.0635132	0.0666274	0.0695755	0.00943396	0.0699991	0.0635165	0.0652843	0.0536557
QGQ59492	0.0845136	0.0188679	0.0654481	0.0623343	0.0200472	0.0138561	0.0648585	0.0617456	0.0683962	0.0725236	0.0135613	0.0688268	0.0629225	0.0664471	0.0560142
ANC57577	0.0947108	0.0640974	0.0794024	0.0783967	0.0705602	0.0696653	0.078235	0.0795756	0.0841175	0.0593867	0.0640941	0.0854701	0.0783967	0.0789861	0.0440916
ANC57580	0.0847034	0.0570272	0.0746771	0.0742706	0.0646666	0.0614162	0.0735022	0.0736811	0.0776214	0.0505587	0.0570267	0.0789861	0.07486	0.0730917	0.0287991
QPI11922	0.0868491	0.00117925	0.0654481	0.0635158	0.0235849	0.0209316	0.0642689	0.0635116	0.067217	0.0678066	0.00707547	0.0705888	0.0635179	0.0658701	0.0512972
QIB99387	0.076295	0.0571934	0.0512972	0.0552745	0.0666274	0.0663325	0.0501179	0.0552687	0.0619104	0.084316	0.057783	0.062343	0.0564516	0.0558698	0.0660377
ANC57586	0.057033	0.0694119	0.0323218	0.0330091	0.0747032	0.0732282	0.0305659	0.0294724	0.0417423	0.08884	0.0682432	0.040672	0.0318302	0.0583554	0.0776504
QBQ58385	0.0516021	0.0654481	0.0182783	0.0211532	0.0678066	0.0716392	0.0165094	0.0199776	0.0306604	0.0913915	0.0642689	0.0317453	0.0229191	0.0576161	0.0748821
ANC57578	0.0876383	0.052924	0.0735145	0.07486	0.0629133	0.060259	0.0729287	0.0707339	0.0817468	0.0535066	0.0541013	0.0807545	0.0760389	0.0701444	0.0335073
BBH51316	0.0868655	0.00825472	0.0654481	0.0629298	0.0229953	0.020342	0.0642689	0.0629281	0.0666274	0.0701651	0.00648585	0.0711786	0.0635202	0.0652805	0.053066
QPZ88405	0.0915863	0.0542453	0.0778302	0.076466	0.0642689	0.0616156	0.0772406	0.0752898	0.0813679	0.0341981	0.0554245	0.0823568	0.078234	0.0770551	0.00943396

Figure 31: Distance matrix of 2mer AAF

	QFS19562	ANC57575	ANC57576	ANC57577	ANC57578	ANC57579	ANC57580	ANC57581	ANC57582	ANC57584	ANC57586	ANC57587	ANC57591	AEF01518	ACL99188
QFS19562	0														
ANC57575	0.00285937	0													
ANC57576	0.00624114	0.00353669	0												
ANC57577	0.00580661	0.00294724	0.00412614	0											
ANC57578	0.00757496	0.00648394	0.00648394	0.00589449	0										
ANC57579	0.0103673	0.00766284	0.00648394	0.00825228	0.0129679	0									
ANC57580	0.00580661	0.00294724	0.00412614	0.0023578	0.00353669	0.00943118	0								
ANC57581	0.00521716	0.00294724	0.00412614	0.00294724	0.00353669	0.00943118	0.000589449	0							
ANC57582	0.00683059	0.00530504	0.00412614	0.00589449	0.0106101	0.00353669	0.00707339	0.00707339	0						
ANC57584	0.00388335	0.00176835	0.0023578	0.00353669	0.00825228	0.00648394	0.00471559	0.00471559	0.00353669	0					
ANC57586	0.00565169	0.00294724	0.00176835	0.00412614	0.00825228	0.00471559	0.00471559	0.00471559	0.0023578	0.00176835	0				
ANC57587	0.00617428	0.00353669	0.00353669	0.00353669	0.00943118	0.00589449	0.00589449	0.00589449	0.00412614	0.0023578	0.00294724	0			
ANC57591	0.00794262	0.00530504	0.00530504	0.00589449	0.011789	0.00294724	0.00825228	0.00825228	0.00294724	0.00412614	0.00353669	0.00294724	0		
AEF01518	0.0121356	0.0106101	0.00825228	0.0111995	0.0147362	0.00294724	0.0123784	0.0123784	0.00530504	0.00884173	0.00766284	0.00884173	0.00589449	0	
ACL99188	0.00742004	0.00471559	0.00294724	0.00294724	0.00825228	0.00530504	0.00471559	0.00471559	0.00412614	0.00353669	0.00294724	0.0023578	0.00353669	0.00825228	0
ACK28184	0.00683059	0.00412614	0.00353669	0.00471559	0.0100206	0.00353669	0.00648394	0.00648394	0.00176835	0.00294724	0.00176835	0.0023578	0.00176835	0.00648394	0.0023578
AAW23164	0.00506224	0.0023578	0.00176835	0.00294724	0.00766284	0.00589449	0.00412614	0.00412614	0.00353669	0.0011789	0.0011789	0.00176835	0.00412614	0.00884173	0.0023578
QGQ59490	0.00454117	0.0025804	0.00428167	0.00325639	0.00390353	0.00958671	0.00142097	0.000831519	0.00722892	0.00434875	0.00487112	0.00552765	0.00788544	0.012534	0.00502474
QGQ59491	0.00454117	0.0025804	0.00428167	0.00325639	0.00390353	0.00958671	0.00142097	0.000831519	0.00722892	0.00434875	0.00487112	0.00552765	0.00788544	0.012534	0.00502474
QGQ59492	0.00454117	0.0025804	0.00428167	0.00325639	0.00390353	0.00958671	0.00142097	0.000831519	0.00722892	0.00434875	0.00487112	0.00552765	0.00788544	0.012534	0.00502474
QQC97219	0.00658974	0.00388529	0.00161455	0.00279293	0.00744127	0.0055266	0.00397183	0.00397183	0.00316881	0.00270639	0.00168181	0.00242505	0.00434771	0.0084066	0.001487
QPZ88403	0.00572042	0.00286105	0.00428167	0.00207715	0.00390353	0.00958671	0.000522371	0.00102563	0.00722892	0.00434875	0.00487112	0.00552765	0.00788544	0.012534	0.00434875
QPZ88404	0.00558665	0.00279415	0.00397305	0.00201024	0.00626203	0.00899709	0.00272533	0.00272533	0.00663929	0.0033836	0.0042815	0.00316916	0.00604984	0.0119443	0.00436804
QPZ88405	0.00558665	0.00279415	0.00397305	0.00154678	0.00744127	0.00781784	0.00390458	0.00390458	0.00546005	0.0033836	0.00397305	0.00198991	0.0048706	0.0107651	0.00318879
QPU83820	0.00541049	0.00270604	0.00170127	0.00251159	0.00744127	0.00604898	0.00390458	0.00390458	0.00369118	0.00168163	0.00161455	0.00198991	0.00434771	0.00899622	0.00207663
QPU83821	0.00541049	0.00270604	0.00170127	0.00251159	0.00744127	0.00604898	0.00390458	0.00390458	0.00369118	0.00168163	0.00161455	0.00198991	0.00434771	0.00899622	0.00207663
QPI70486	0.00631004	0.00493889	0.00605054	0.00434944	0.0017011	0.0114228	0.00199165	0.00199165	0.00906504	0.00670724	0.00670724	0.00788614	0.0102439	0.0143028	0.00670724
QPI11922	0.00558665	0.00310312	0.00546092	0.00318949	0.00390353	0.010766	0.00201059	0.00170162	0.00840816	0.00487147	0.00605037	0.00552765	0.00788544	0.0137132	0.00554728
QPI11926	0.00558665	0.00310312	0.00546092	0.00318949	0.00390353	0.010766	0.00201059	0.00170162	0.00840816	0.00487147	0.00605037	0.00552765	0.00788544	0.0137132	0.00554728
QPB40125	0.00717936	0.00447491	0.000938218	0.00369083	0.00685165	0.00611623	0.00456145	0.00456145	0.00375843	0.00329601	0.00168181	0.00360429	0.00493733	0.00788457	0.00251193

Figure 32: Distance matrix of GFCGR

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.00434716	0													
QFS19150	0.00453397	0.00531696	0												
ACK28184	0.00343235	0.00447939	0.00279599	0											
QHR82546	0.0052009	0.00332665	0.00659164	0.00479608	0										
QCZ25008	0.00559249	0.00401136	0.007052	0.005546	0.00358406	0									
QFS19149	0.00392149	0.00494781	0.00117951	0.00195508	0.00601655	0.0067888	0								
ACL99188	0.00365136	0.00407407	0.00343246	0.00348135	0.00578705	0.00666116	0.00278703	0							
QQC97219	0.0039557	0.00346646	0.00242747	0.00257031	0.00491626	0.00504787	0.00247798	0.00371074	0						
QPZ88403	0.00627407	0.00299956	0.00676243	0.00663219	0.0049401	0.00594866	0.00655289	0.00489258	0.0055913	0					
BBH51315	0.00446123	0.00153819	0.00509496	0.00430439	0.00284603	0.00367672	0.00487747	0.00459154	0.00326127	0.00346178	0				
AEF01518	0.00551942	0.00616857	0.00318565	0.00305495	0.0068739	0.00797946	0.00251884	0.00432816	0.00393443	0.00794889	0.00634729	0			
AAW23164	0.00315765	0.00229589	0.00375703	0.002762	0.00430012	0.00496215	0.00315316	0.00275753	0.00237091	0.00468105	0.00291596	0.00433286	0		
ANC57587	0.0045057	0.00192215	0.00502136	0.00361717	0.00298359	0.00399764	0.0045402	0.00436001	0.00334918	0.00427736	0.00193559	0.00549318	0.00238967	0	
QPZ88404	0.00563257	0.00351764	0.00631755	0.00618476	0.00450098	0.00578227	0.00608585	0.004623	0.00537323	0.00208722	0.00355855	0.00759754	0.00485845	0.00460307	0
ANC57579	0.00557905	0.00647398	0.00187506	0.00346927	0.00765292	0.00786758	0.00220827	0.00432977	0.00357865	0.00797388	0.00641345	0.00278843	0.00481638	0.00594823	0.00753153
QPU83820	0.00324457	0.00321133	0.00340775	0.00177146	0.00418182	0.00505479	0.00254101	0.00238902	0.00268164	0.00519939	0.00350036	0.00366768	0.00166371	0.00265546	0.00497158
QPB40125	0.00377475	0.00335422	0.00318983	0.00201813	0.0040314	0.00520474	0.00243695	0.00281113	0.0024171	0.0051703	0.00347298	0.00328624	0.00226685	0.00285104	0.00467364
QFS19134	0.00381513	0.0037235	0.00317175	0.00345169	0.00442546	0.005388	0.00314026	0.00352253	0.00258924	0.00477963	0.00312094	0.00504785	0.00344134	0.0039999	0.00375171
QPB40127	0.00476944	0.00117264	0.00557762	0.00500478	0.00329158	0.00408191	0.00537067	0.00471478	0.00370977	0.00281139	0.00107391	0.00675546	0.00312322	0.00249334	0.00317408
QGQ59492	0.00346821	0.00222442	0.0047705	0.00331817	0.00215654	0.00354304	0.00422885	0.00388734	0.00319834	0.00422308	0.0019304	0.00554909	0.0025991	0.00229074	0.0036429
ANC57577	0.00575952	0.00416321	0.00602759	0.00570357	0.00360326	0.00551341	0.00585639	0.00561345	0.00487826	0.00430971	0.00346172	0.00715776	0.00514878	0.00480206	0.00302509
ANC57580	0.00480403	0.00377132	0.00571424	0.00582569	0.00484337	0.00621314	0.00554546	0.00452668	0.00479507	0.00356873	0.00367494	0.00719302	0.00448891	0.0050748	0.00262848
QPI11922	0.00445413	0.000744182	0.00537739	0.00455574	0.00290734	0.0038123	0.00505276	0.00428943	0.00355463	0.0028351	0.00102318	0.00638252	0.00272262	0.00200997	0.00309864
QIB99387	0.0041051	0.00357423	0.00289922	0.00323824	0.00519237	0.00612596	0.00247609	0.00145554	0.00285208	0.00442293	0.00386886	0.00407577	0.00258909	0.00389613	0.00408278
ANC57586	0.00285216	0.0042219	0.00306091	0.00190592	0.00534215	0.00590922	0.0021119	0.00199716	0.00307063	0.00602695	0.0045652	0.00342534	0.0023225	0.00392886	0.00565674
QBQ58385	0.00371707	0.00440495	0.00170191	0.001217	0.00532637	0.00600882	0.000954713	0.00291535	0.00206759	0.00625319	0.00426241	0.00277897	0.00266641	0.0037734	0.00587382
ANC57578	0.00426434	0.00323715	0.00498597	0.0044696	0.00322663	0.00480392	0.00467586	0.0042038	0.00388897	0.00388983	0.00276517	0.00615728	0.00390111	0.00382689	0.00247014
BBH51316	0.00447367	0.000887429	0.00541851	0.00448694	0.00275	0.00357103	0.00509466	0.00446656	0.0035421	0.00310336	0.000851699	0.00640514	0.00277226	0.00177981	0.00334888
QPZ88405	0.00507846	0.00361278	0.00545018	0.00545795	0.00445484	0.00587022	0.00522733	0.00397863	0.00476553	0.00280992	0.00347945	0.00683527	0.00447419	0.00449972	0.00110946

Figure 33: Distance matrix of CGR Centroid

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.000607862	0													
QFS19150	0.00173826	0.0017835	0												
ACK28184	0.00159565	0.00170211	0.000207011	0											
QHR82546	0.00154207	0.000938801	0.00208772	0.00209926	0										
QCZ25008	0.000991285	0.000433382	0.00213491	0.00207749	0.000704194	0									
QFS19149	0.00181046	0.00191356	0.00020166	0.000216533	0.00227117	0.00228239	0								
ACL99188	0.00155276	0.00159772	0.000188794	0.000180595	0.0019388	0.00195531	0.000334803	0							
QQC97219	0.000755592	0.000953576	0.000985553	0.000840485	0.00162714	0.00137508	0.00105592	0.000802734	0						
QPZ88403	0.00189676	0.0012915	0.00233775	0.00237171	0.000356249	0.00101577	0.00252919	0.00220318	0.00196301	0					
BBH51315	0.00137212	0.000766353	0.00217368	0.00215839	0.000285418	0.000445766	0.0023454	0.00201061	0.00158491	0.000570059	0				
AEF01518	0.00288315	0.0031666	0.00158333	0.00154798	0.00364724	0.00357445	0.00138711	0.00171199	0.00221791	0.00391482	0.0036985	0			
AAW23164	0.000636445	0.00104032	0.00122232	0.00104659	0.00184061	0.00147351	0.00124996	0.00105394	0.000327087	0.00218889	0.00175509	0.0022468	0		
ANC57587	0.00106929	0.000813593	0.0010614	0.0010434	0.00106491	0.00109491	0.00122917	0.000894455	0.000676461	0.00136371	0.00111642	0.00258859	0.000979129	0	
QPZ88404	0.00229764	0.00169746	0.00253599	0.00259957	0.00075946	0.00143592	0.00273461	0.0024233	0.00229368	0.000420163	0.000990221	0.00411864	0.00254382	0.00165231	0
ANC57579	0.00281988	0.00298336	0.00123428	0.00128177	0.00331246	0.00335688	0.00107452	0.00140469	0.00208551	0.00354224	0.00340762	0.000640942	0.00220672	0.00229534	0.00369599
QPU83820	0.00150112	0.00159214	0.000251261	0.000111919	0.00199173	0.00196582	0.000322244	0.000122525	0.000745684	0.00226799	0.00204766	0.00165599	0.000971554	0.000933493	0.00250298
QPB40125	0.00167947	0.00170358	9.26086e-05	0.000231649	0.00199524	0.00204942	0.000289316	0.000130967	0.000931796	0.00224533	0.00208263	0.00167455	0.0011848	0.000971768	0.0024456
QFS19134	0.00152063	0.000987125	0.00168271	0.00171914	0.000454651	0.000950314	0.00187418	0.00154909	0.00138955	0.000655098	0.000688935	0.00326019	0.00165754	0.000739647	0.000912681
QPB40127	0.00125458	0.000706691	0.00236258	0.00231792	0.000651536	0.000273607	0.00251836	0.0021874	0.00164134	0.000891967	0.000366931	0.00382957	0.00174698	0.00130667	0.00130321
QGQ59492	0.00108161	0.000543258	0.00162091	0.00160267	0.0005501	0.000609186	0.00179032	0.00145561	0.00107705	0.000891707	0.000555733	0.00314399	0.00129787	0.000561155	0.00125331
ANC57577	0.00238006	0.00177531	0.00292997	0.00296674	0.00090829	0.00141409	0.00312261	0.0027977	0.00253532	0.000595182	0.00100898	0.0045089	0.00274089	0.00195566	0.000529086
ANC57580	0.00187706	0.00135428	0.00295976	0.00293289	0.000970172	0.000920923	0.00312618	0.00279207	0.00228701	0.000974545	0.000793751	0.00446037	0.00239401	0.00189869	0.0012648
QPI11922	0.00106853	0.000461881	0.00198685	0.00194859	0.000505232	0.000238288	0.00214603	0.00181365	0.00131339	0.00084503	0.000304603	0.00347177	0.00146255	0.000928584	0.00126031
QIB99387	0.00152569	0.00144706	0.000432253	0.000504744	0.00166036	0.00176011	0.000627167	0.000324151	0.000833211	0.00190559	0.00175975	0.00201402	0.00113973	0.000666304	0.00210834
ANC57586	0.0016416	0.0018625	0.00052017	0.000348101	0.00237069	0.00226654	0.000379623	0.00052436	0.000924627	0.002662	0.00240048	0.00130822	0.00102667	0.00130581	0.00291253
QBQ58385	0.00172432	0.0017895	5.80517e-05	0.000160235	0.00211872	0.00214818	0.000156014	0.00019292	0.000969471	0.00237408	0.00219819	0.00154298	0.0011948	0.00108324	0.0025787
ANC57578	0.0017846	0.00118673	0.00216504	0.00219982	0.000257361	0.000959552	0.00235658	0.00203087	0.00180794	0.000172713	0.00052673	0.00374237	0.00204368	0.00119685	0.000513537
BBH51316	0.001091	0.000483204	0.00196797	0.00193356	0.000471527	0.000278618	0.00212934	0.00179626	0.00131228	0.000815306	0.000288194	0.00346066	0.00146987	0.000907931	0.00122877
QPZ88405	0.0022434	0.00165956	0.00233349	0.00240848	0.000751777	0.00145518	0.00253371	0.0022301	0.00216377	0.000485791	0.00102005	0.00391272	0.00243186	0.00150312	0.000244138

Figure 34: Distance matrix of CGR Centroid Bisection

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.0254359	0													
QFS19150	0.0112397	0.0177588	0												
ACK28184	0.0128275	0.0165528	0.00392571	0											
QHR82546	0.0254362	0.00475565	0.0177564	0.0165455	0										
QCZ25008	0.0262042	0.00413997	0.0186935	0.0174774	0.00368981	0									
QFS19149	0.0106973	0.0172152	0.000324006	0.00368195	0.0173316	0.0183119	0								
ACL99188	0.0121415	0.0159829	0.0030774	0.00360232	0.0168456	0.0172027	0.00261236	0							
QQC97219	0.0152929	0.0196508	0.00679586	0.00366515	0.0204244	0.0212885	0.00647185	0.00621986	0						
QPZ88403	0.0331116	0.0178149	0.027359	0.0272477	0.0193872	0.0184929	0.0272266	0.0264376	0.0304666	0					
BBH51315	0.025153	0.00142445	0.01713	0.0155801	0.00394598	0.00318833	0.0166536	0.0155855	0.0189118	0.017697	0				
AEF01518	0.0148402	0.0196798	0.00541393	0.00471887	0.0202245	0.0206924	0.00528177	0.00570146	0.00782255	0.0296628	0.0189925	0			
AAW23164	0.0141819	0.0165217	0.00495194	0.00454656	0.0175652	0.017872	0.00445756	0.0038261	0.00740313	0.0269739	0.0156651	0.00693297	0		
ANC57587	0.0199091	0.0200288	0.0140645	0.0126882	0.0202241	0.0204383	0.0133105	0.0137942	0.0154404	0.029011	0.0187994	0.0152578	0.0138232	0	
QPZ88404	0.0267131	0.0128149	0.0208259	0.0201054	0.0136106	0.0130947	0.0205091	0.0196356	0.0237202	0.0067878	0.0124657	0.0223253	0.0205922	0.0221927	0
ANC57579	0.0134088	0.0180475	0.00539558	0.00464151	0.0193027	0.0191389	0.00522085	0.00445475	0.00722866	0.0276766	0.017369	0.00800219	0.00641943	0.0148266	0.0217028
QPU83820	0.0133433	0.0175266	0.00474015	0.00199416	0.0183198	0.0185762	0.00441615	0.0039745	0.00467796	0.0285857	0.0166387	0.0056471	0.00496145	0.0133519	0.0210002
QPB40125	0.0287762	0.0139846	0.0232261	0.0217052	0.015829	0.0147788	0.0228072	0.0228183	0.0258233	0.0181492	0.0136693	0.0249422	0.0219644	0.0217779	0.0147302
QFS19134	0.0254695	0.011634	0.0192516	0.0189363	0.0131754	0.0121788	0.0190107	0.0188571	0.0222478	0.0118629	0.0113165	0.0213956	0.0199155	0.0201545	0.00515972
QPB40127	0.0248164	0.00100689	0.0170957	0.0158538	0.00483857	0.00429395	0.0165521	0.0156295	0.0190914	0.0181591	0.00144434	0.0188201	0.0159362	0.0188281	0.0131315
QGQ59492	0.0244718	0.00328551	0.0169294	0.0154969	0.00310302	0.0020131	0.0165017	0.0154285	0.019452	0.019316	0.00246464	0.019103	0.0157522	0.0191322	0.0133005
ANC57577	0.0304355	0.0170235	0.0252356	0.0237699	0.0178472	0.0182935	0.0248609	0.0236897	0.0273518	0.0158679	0.0165616	0.0268236	0.0233364	0.0236455	0.00997914
ANC57580	0.0249558	0.0145011	0.020678	0.0202005	0.0152378	0.0147669	0.0202992	0.0202277	0.023058	0.0122176	0.0142603	0.022354	0.0207175	0.0208573	0.00569616
QPI11922	0.0247568	0.000461981	0.0173242	0.0161184	0.00429366	0.00367799	0.0167806	0.0155485	0.0194335	0.0175427	0.000962467	0.0192454	0.0160872	0.0193497	0.0125427
QIB99387	0.0208826	0.0146827	0.0128427	0.0142474	0.0155277	0.016138	0.0122742	0.0131659	0.0170539	0.0247101	0.0140123	0.01621	0.0138493	0.0142692	0.0172414
ANC57586	0.014185	0.0190017	0.00610139	0.00668496	0.0200255	0.020251	0.00580306	0.00518432	0.00919254	0.0284727	0.018857	0.00801183	0.0064011	0.0154097	0.0223811
QBQ58385	0.0121268	0.0179293	0.00401856	0.00406591	0.018271	0.0186978	0.00384965	0.00388772	0.00651733	0.0276598	0.0172331	0.00667696	0.00512042	0.0139876	0.0207275
ANC57578	0.0265556	0.0141282	0.0216219	0.0211075	0.0156638	0.0152127	0.0213369	0.0200255	0.0254026	0.0136125	0.0143721	0.0230388	0.0208165	0.0222944	0.00693206
BBH51316	0.0250285	0.00174357	0.0171224	0.0158712	0.00442153	0.00389946	0.0165988	0.0158671	0.0191732	0.0188989	0.00104552	0.0193502	0.0160722	0.0189997	0.0135067
QPZ88405	0.0267801	0.0131032	0.0209102	0.0202261	0.0150549	0.0144335	0.0206231	0.020038	0.0239422	0.00717507	0.0131297	0.0224824	0.0212684	0.023235	0.00203689

Figure 35: Distance matrix of J(x,y)

	QFS19562	QPB40131	QFS19150	ACK28184	QHR82546	QCZ25008	QFS19149	ACL99188	QQC97219	QPZ88403	BBH51315	AEF01518	AAW23164	ANC57587	QPZ88404
QFS19562	0														
QPB40131	0.0174852	0													
QFS19150	0.00822477	0.0136621	0												
ACK28184	0.0106547	0.0137769	0.00280613	0											
QHR82546	0.0191449	0.00344628	0.0139901	0.0135484	0										
QCZ25008	0.0220782	0.003296	0.0156855	0.0153141	0.00359133	0									
QFS19149	0.00742582	0.0121417	0.000475927	0.00279054	0.0129801	0.0145585	0								
ACL99188	0.00855879	0.0104826	0.00275758	0.00351915	0.0113564	0.0129691	0.00225192	0							
QQC97219	0.0106073	0.0120897	0.00494336	0.00244727	0.0129988	0.0145736	0.00441214	0.00478415	0						
QPZ88403	0.0218869	0.012089	0.021319	0.0204569	0.0151567	0.0136645	0.0199426	0.0181126	0.0187547	0					
BBH51315	0.0182898	0.00124591	0.012962	0.0124925	0.00296089	0.00272991	0.011797	0.0100656	0.0115704	0.012904	0				
AEF01518	0.0108745	0.0142675	0.00505007	0.00456593	0.0153263	0.0169706	0.00421806	0.00490282	0.00600308	0.0237609	0.0141004	0			
AAW23164	0.0103716	0.0129997	0.00283792	0.00275371	0.013404	0.0148685	0.00239244	0.00231018	0.00441572	0.0198705	0.0122236	0.00453395	0		
ANC57587	0.0155496	0.0122461	0.0113251	0.0102398	0.0134192	0.0145405	0.0108788	0.00905097	0.0109976	0.0182836	0.0112839	0.0130616	0.0113244	0	
QPZ88404	0.0195606	0.0087613	0.0172863	0.0170805	0.0121696	0.0104893	0.0161436	0.0147185	0.0161097	0.00342098	0.00932057	0.0192207	0.0165949	0.0137723	0
ANC57579	0.00978526	0.0131092	0.00411982	0.00350179	0.0138376	0.0147397	0.00364342	0.00382664	0.00412632	0.0173638	0.0122096	0.00622906	0.00434949	0.0113334	0.0148621
QPU83820	0.0106554	0.0107195	0.00439302	0.00243201	0.0119923	0.0133556	0.00390586	0.00285382	0.00344534	0.0186775	0.0101298	0.00525988	0.0030851	0.00898125	0.0148592
QPB40125	0.0187209	0.00683929	0.0168695	0.0159816	0.0101612	0.00961696	0.0150537	0.014572	0.0155282	0.0134301	0.00802448	0.0155665	0.0160266	0.0132585	0.00954161
QFS19134	0.0186432	0.00601944	0.0145094	0.0149261	0.00959307	0.00814321	0.013442	0.0124477	0.014217	0.00759711	0.00645574	0.015921	0.0147792	0.0124823	0.0036962
QPB40127	0.0169969	0.000225495	0.0134423	0.0135862	0.00362917	0.00350949	0.0119452	0.0102038	0.0118802	0.0122184	0.00135919	0.0140567	0.0128567	0.0118531	0.00892627
QGQ59492	0.0184193	0.00269047	0.0120228	0.0118151	0.00300164	0.00237843	0.0109544	0.00931331	0.0113336	0.0159917	0.00191665	0.0124539	0.0112758	0.0128161	0.012186
ANC57577	0.0196172	0.0117989	0.0170355	0.0162032	0.0144541	0.0148427	0.0164381	0.0151121	0.0160551	0.00919266	0.0117391	0.0195513	0.015545	0.0142565	0.0063032
ANC57580	0.0184367	0.0100103	0.0179666	0.0172362	0.0141042	0.0129086	0.0165611	0.0152228	0.0150307	0.00770747	0.0107026	0.0187772	0.0168019	0.0125972	0.00393433
QPI11922	0.0174659	1.00849e-05	0.0136477	0.0137624	0.00343528	0.00328511	0.0121279	0.0104689	0.0120821	0.0120854	0.00123486	0.0142528	0.0129859	0.0122264	0.00875795
QIB99387	0.0145587	0.00951605	0.0114439	0.0119945	0.012394	0.0129994	0.0100521	0.0107345	0.0119287	0.0147728	0.00943115	0.0124646	0.0120274	0.00979965	0.010386
ANC57586	0.0135828	0.0187457	0.0080981	0.00666669	0.0193676	0.021749	0.00851044	0.0067281	0.00869319	0.0220724	0.0181932	0.00951565	0.00711106	0.0121577	0.0194782
QBQ58385	0.00854132	0.011817	0.00203426	0.00291159	0.0125733	0.0141501	0.00154997	0.00227417	0.00450655	0.0183967	0.0111561	0.00481698	0.00248515	0.0101815	0.0148401
ANC57578	0.0174512	0.00860729	0.0148582	0.0154367	0.0123632	0.0112325	0.0136034	0.0124191	0.0149657	0.00737428	0.0094849	0.0167949	0.0139892	0.0126349	0.00399724
BBH51316	0.0169917	0.000936752	0.0125358	0.0119175	0.00332878	0.0033882	0.0113565	0.00982059	0.0107899	0.0121792	0.00102083	0.0135312	0.0119872	0.0109267	0.00836863
QPZ88405	0.0194662	0.00948191	0.0172058	0.0164413	0.013136	0.0115438	0.0160817	0.0145347	0.015362	0.00421174	0.0101449	0.0185859	0.0164692	0.0138289	0.00101834

Figure 36: Distance matrix of D = 1-rw

To rank the effectiveness of each distance metric we define the [31], $\delta(x, y)$

function of two viral sequences x and y as follows:

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

With this function we create a 400x400 distance matrix of the viruses and take the upper triangular matrix as a vector U_δ. Next, we take the upper triangle matrix, U_α, $\alpha \in$ 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 are unrelated. The values of P_α for each distance measure are shown in figure 6. We see that of the distance measures, Kullback-lieber discrimination information, J(x,y) is most closely related with U_δ. Further confirmation of this is shown in the 2D and 3D MDS charts for J(x,y) 38, 41, which show a good separation of the viral sequences into their respective groups. 2mer AAF also shows a linear correlation with U_δ with a P_α of 0.62734. Similarly, the 2D and 3D MDS graphs of 2mer AAF show a good separation and clustering of the viral sequences. It can also be

Method	P_α
J(a,b)	0.640972
2mer AAF	0.62734
D = 1-rw	0.558031
CGR Centroid	0.503566
GFCGR	0.48629
CGR Centroid Bisection	0.48301
S_2	0.309167

Table 6: P_α of Distance metrics

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 and SARS_COV2 show a distance measure of almost 0 as their clusters are overlapping. Other measures such as Shannon entropy and CGR Centroid Bisection which have the lowest correlation with U_δ, $P_\alpha = 0.309167$ and 0.48301 respectively, show a lack of separation between viral groups in their MDS charts.

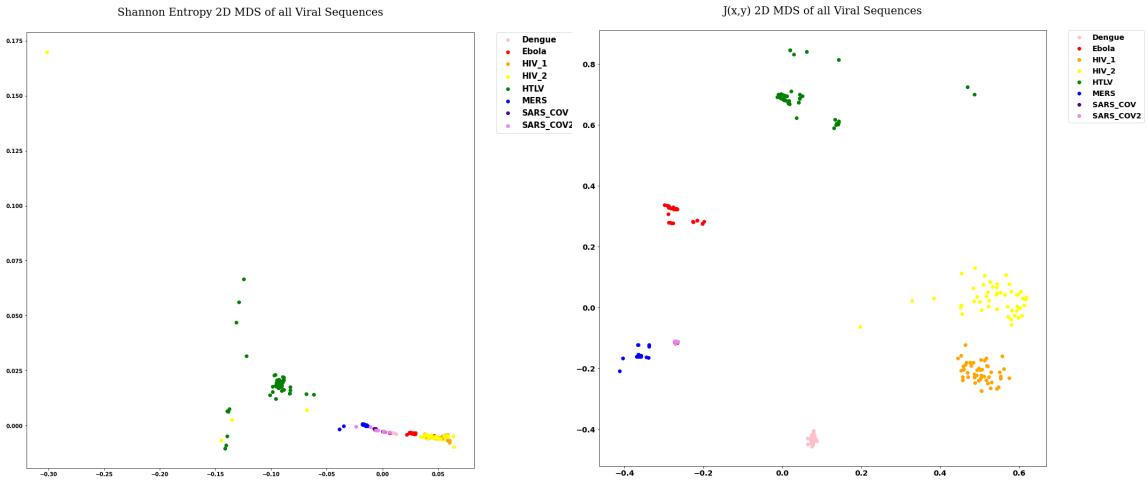


Figure 38: 2D MDS of S_2 and J(x,y)

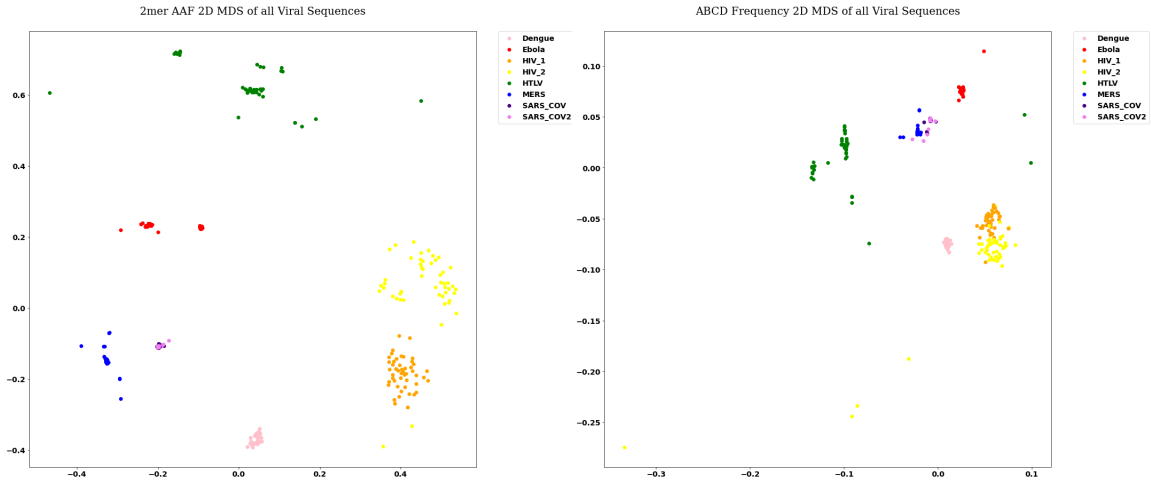


Figure 37: 2D MDS of 2mer AAF and GFCGR

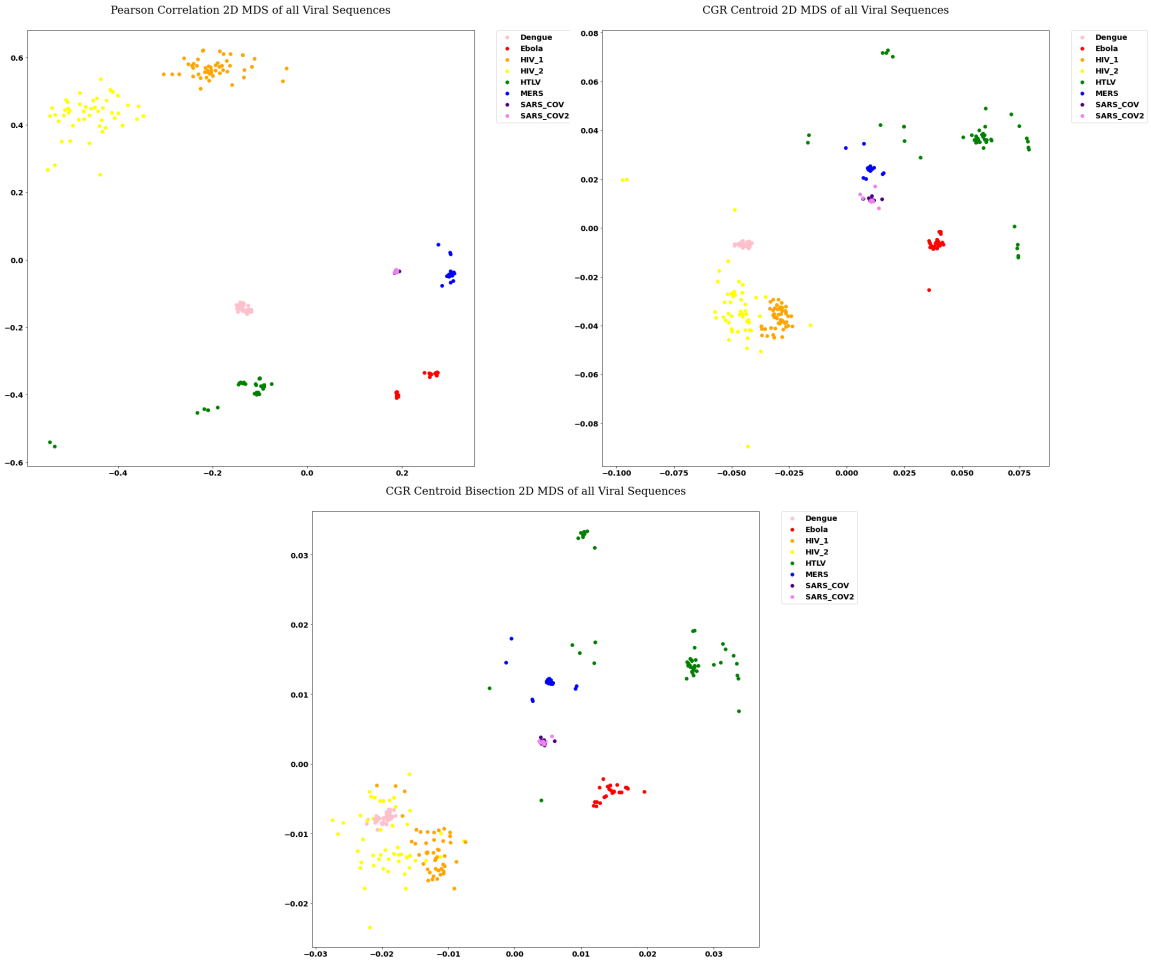


Figure 39: 2D MDS of Pearson Correlation, CGR Centroid, and CGR Centroid Bisection

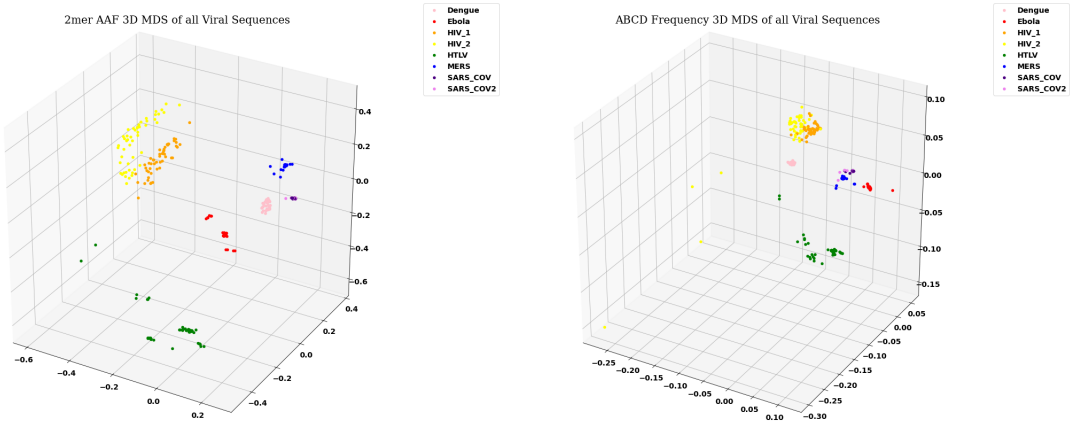


Figure 40: 3D MDS of 2mer AAF and GFCGR

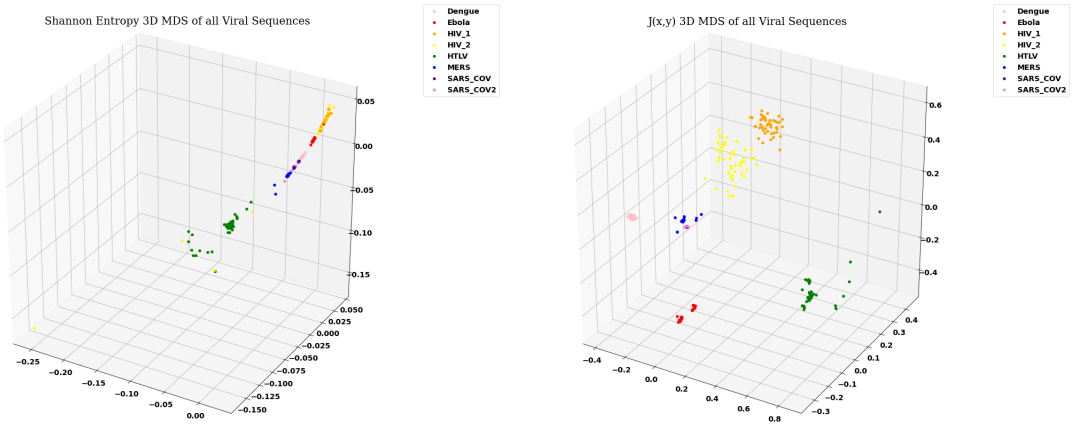


Figure 41: 3D MDS of S_2 and J(x,y)

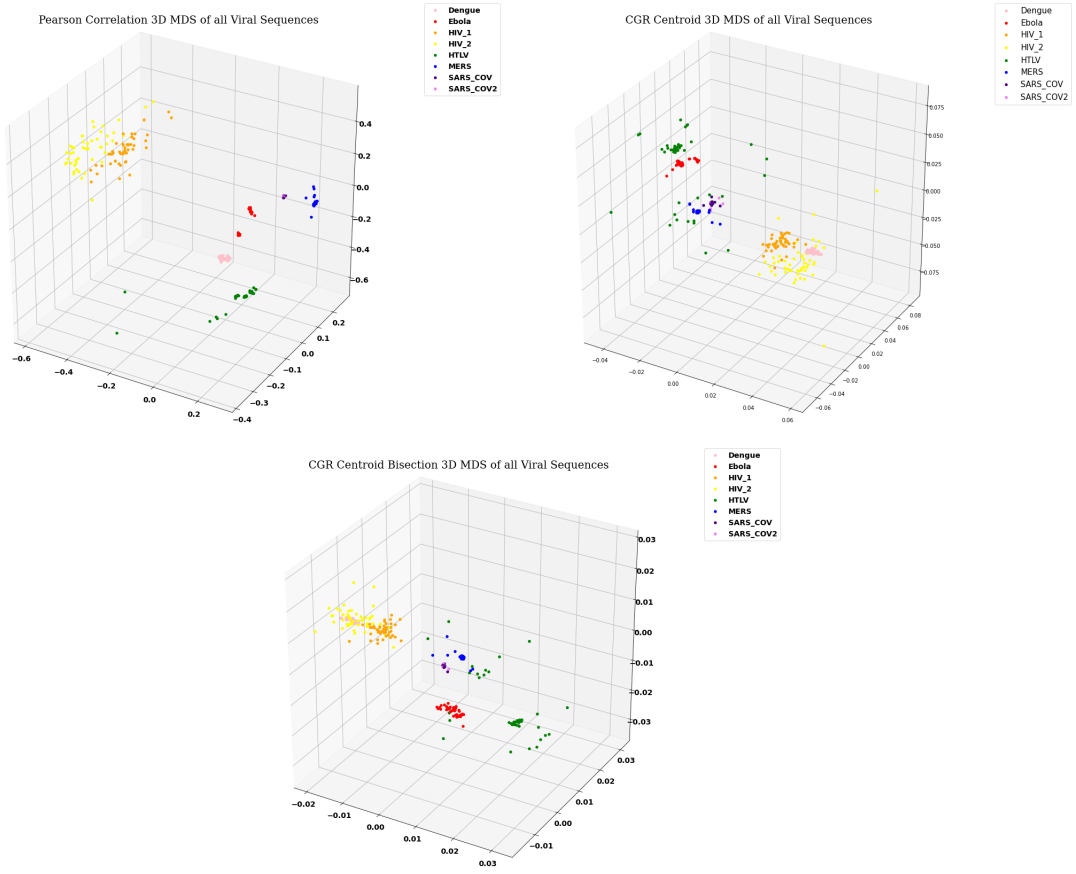


Figure 42: 3D MDS of Pearson Correlation, CGR Centroid, and CGR Centroid Bisection

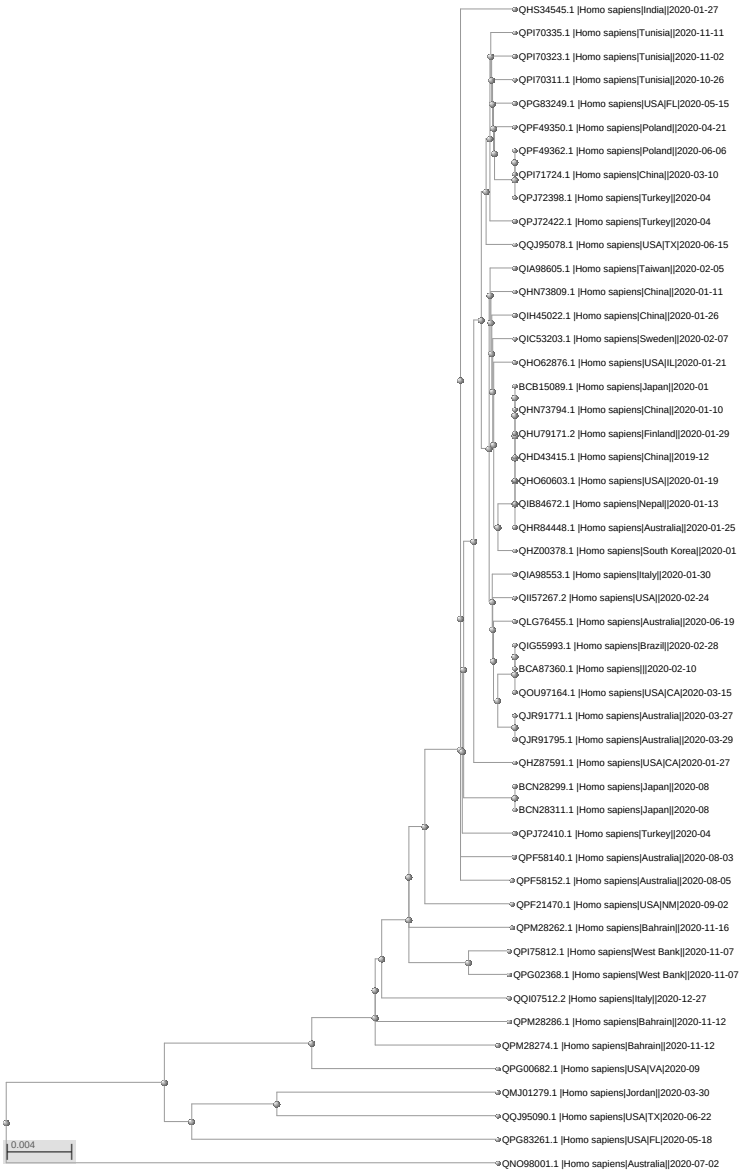


Figure 43: Phylogenetic Tree of SARS_COV2 from NCBI website

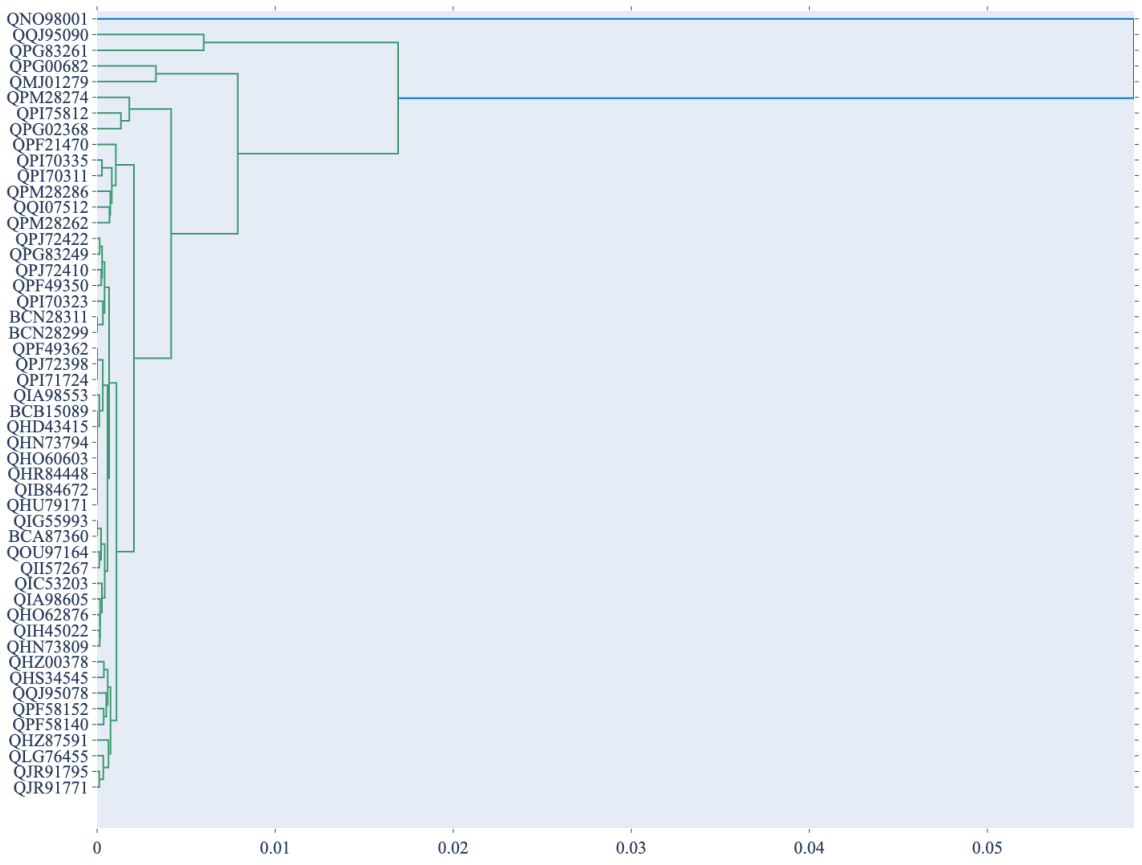


Figure 44: Phylogenetic Tree made using J(x,y)

6 Conclusion

We set forth to accurately identify viral sequences and place them in their respective groups. Several distance metrics were introduced for comparison as well as a method of ranking these metrics. Our findings suggest that the Kullback-Lieber Discrimination Information as well as the manhattan distance of 2mer AAF are best in clustering viruses into their respective groups. This shows the importance of the frequency of 2mers in correctly identifying viral sequences. Additional evidence of this is shown by the closeness in the phylogenetic tree of SARS_COV2 from NCBI and from J(x,y) figures 43, 44. Overall, we were able to distinguish between the viral groups as well as cluster the sequences appropriately using our methods.

7 Future Works

Given our results, the next steps would be to further increase the size of the datasets as well as look into other kmers lengths for amino acids. This would allow for further testing into the impact of amino acid frequency. Another aspect to look into is the grouping of viral sequences by geographical location and also protein types. This would allow us to get an idea on how the virus transfers from country to country in the world. Lastly, we would look at other ways to verify the accuracy of the distance metrics and additional metrics that could be used such as cosine distance.

8 Pseudo-Code

Main Code
Loop through FASTA files of viral sequences
Get accession number and amino acid sequence of each strain from file
Call function to perform CGR and graph CGR
Call function to calculate CGR Centroid and CGR Centroid Bisection
Call function to calculate amino acid frequency
Call function to calculate GFCGR
Call function to calculate Shannon Entropy and J(x,y)

Table 7: Main Code

CGR Function
For each amino acid in a viral sequence
$(\mathrm{x}_{i+1}, y_{i+1}) = \frac{x_i + T_x(i)}{2}, \frac{y_i + T_y(i)}{2}$ where $(T_x(i), T_y(i)) \in \{(0,1),(0,0),(1,1),(1,0)\}$ and $0 \leq i \leq$ length of sequence
Append x and y values to a list
Set figure size to 18" x 18"
Graph x and y coordinates from CGR using scatter plot
Save CGR plot
Stop

Table 8: CGR Function

CGR Centroid and CGR Centroid Bisection Function
For x,y coordinates in CGR
Cluster points into four quadrants of unit square
Calculate centroid for each quadrant as follows $\mathrm{C}_k = \frac{\sum_{i=1}^{n}(a_i(x),a_i(y))}{\text{length of sequence}}$ where $(a_i(x), a_i(y))$ are the x and y coordinates respectively in a cell
Draw rectangle using four vertices as well as draw diagonals of rectangle
Obtain $\mathrm{B}_C(x) = \frac{C_1+C_4}{2} asCGRCentroidBisection$
Stop

Table 9: CGR Centroid and CGR Centroid Bisection Function

Amino Acid Frequency Function
Create Dictionary of all possible 2mers
For every 2 letters in viral sequence
Add 1 to count of 2mer present
Divide total for each 2mer by length of sequence
Stop

Table 10: AAF Function

GFCGR Function
For each quadrant of CGR
Count points in quadrant and divide by length of sequence
Stop

Table 11: GFCGR Function

Shannon Entropy and J(x,y) Function
For each GFCGR
$S_2 = \sum_{i=1}^{k} p_i * log_2(\frac{1}{p_i})$ where p_i are the frequencies of GFCGR
Stop
For each 2mer amino acid frequency
$I(x,y) = \sum_{i=1}^{n} x(a_i) log \frac{x(a_i)}{y(a_i)}$ where x and y are 2mer AAF
J(x,y) = I(x,y) + I(y,x)
Stop

Table 12: Shannon Entropy and J(x,y) Function

References

[1] Jeffrey HJ, Chaos game representation of gene structure. Nucleic Acids Res 1990, 18(8):2163–2170. Retrieved from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-243

[2] Bhoumik, P., Hughes, A. L. (2018). Chaos game representation: an alignment-free technique for exploring evolutionary relationships of protein sequences. BioRxiv, 276915. Retrieved from: https://scholar.google.com/scholar?hl=enas_sdt=0

[3] Nick Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC309551/pdf/nar00059-0196.pdf

[4] Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A., Fletcher, M. (2001). Analysis of genomic sequences by Chaos Game Representation. Bioinformatics, 17(5), 429-437. Retrieved from: https://scholar.google.com/scholar?hl=enas_sdt=0

[5] He, L., Dong, R., He, R. L., Yau, S. S. T. (2020). Positional Correlation Natural Vector: A Novel Method for Genome Comparison. Inter-

national Journal of Molecular Sciences, 21(11), 3859. Retrieved from: https://scholar.google.com/scholar?hl=enas_sdt=0

[6] Abd Elwahaab, M. A., Abo-Elkhier, M. M., Abo el Maaty, M. I. (2019). A Statistical Similarity/Dissimilarity Analysis of Protein Sequences Based on a Novel Group Representative Vector. BioMed research international, 2019. Retrieved from: https://www.hindawi.com/journals/aaa/2013/926519/

[7] Seladi-Schulman, J. (2019) *What is a Retrovirus* Retrieved from: https://www.healthline.com/health/what-is-a-retrovirus

[8] Li, N. N., Shi, F., Niu, X. H., Xia, J. B. (2009). A novel method to reconstruct phylogeny tree based on the chaos game representation. J. Biomed. Sci. Eng, 2, 582-586. Retrieved from: https://www.scirp.org/journal/paperinformation.aspx?paperid=973

[9] Rigden DJ. From protein structure to function in bioinformatics. New York: Springer-verlag; 2009. Retrieved from: https://www.worldcat.org/title/from-protein-structure-to-function-with-bioinformatics/oclc/762183194

[10] MedlinePlus *What are proteins and what do they do?* U.S. National Library of Medicine Retrieved from: https://medlineplus.gov/genetics/understanding/howgeneswork/protein/

[11] Lynch M. Intron evolution as a population-genetic process. Proc Natl Acad Sci USA 2002; 99: 6118-23 Retrieved from: https://www.pnas.org/content/99/9/6118

[12] Gotoh O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, *162*(3), 705–708. Retrieved from: https://doi.org/10.1016/0022-2836(82)90398-9

[13] Moret, B. M., Wang, L. S., Warnow, T., Wyman, S. K. (2001). New approaches for reconstructing phylogenies from gene order data. *Bioinformatics (Oxford, England)*, *17 Suppl 1*, S165–S173. Retrieved from: https://doi.org/10.1093/bioinformatics/17.suppl_1.s165

[14] Qi Z, Li K, Ma J, Yao Y, Liu L. Novel method of 3-dimensional graphical representation for proteins and its application. Evol Bioinforma. 2018;14:1–8.

[15] Li C, Zhao J, Wang C, Yao Y. Protein sequence comparison and DNA-binding protein identification with generalized PseAAC and graphical representation. Comb Chem High Throughput Screen. 2018;21:100–10.

[16] Mehri M, Fatemeh A, Vahid Z. A novel graphical representation and similarity analysis of protein sequences based on physiochemical properties. Physica A. 2018;510:477–85.

[17] Yin, Changchuan, and Jiasong Wang. "Periodic power spectrum with applications in detection of latent periodicities in DNA sequences." *Journal of mathematical biology* 73.5 (2016): 1053-1079. Retrieved from: https://arxiv.org/pdf/1504.02367.pdf

[18] Hoang, Tung, Changchuan Yin, and Stephen S-T. Yau. "Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison." *Genomics* 108.3-4 (2016): 134-142. Retrieved from: https://www.sciencedirect.com/science/article/pii/S0888754316300854

[19] Vinga, Susana. "Information theory applications for biological sequence analysis." *Briefings in bioinformatics* 15.3 (2014): 376-389. Retrieved from: https://www.deepdyve.com/lp/oxford-university-press/information-theory-applications-for-biological-sequence-analysis-sTBwI8Bi02

[20] Deng M, Yu C, Liang Q, He RL, Yau SS-T (2011) Correction: A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. PLoS ONE 6(3): 10.1371/annotation/22351496-73dc-4205-9d9a-95a821ae74ca. Retrieved from: https://doi.org/10.1371/annotation/22351496-73dc-4205-9d9a-95a821ae74ca

[21] Li, Y., He, L., Lucy He, R. *et al.* A novel fast vector method for genetic sequence comparison. *Sci Rep* **7, **12226 (2017). Retrieved from: https://doi.org/10.1038/s41598-017-12493-2

[22] L. L. Cavalli Sforza and A. W. Edwards. (1967) Phy- logenetic analysis: Models and estimation procedures [J], Genetics, **19(3)**, 233–257. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706274/

[23] J. Felesenstein. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach [J], J Mol Evol, **17(6)**, 368–376. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/7288891/

[24] Mu, Z., Yu, T., Qi, E., Liu, J., Li, G. (2019). DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information. *BMC bioinformatics*, *20*(1), 1-10. Retrieved from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2943-x

[25] Basu, S., Pan, A., Dutta, C., Das, J. (1997). Chaos game representation of proteins. *Journal of molecular graphics modelling*, *15*(5), 279–289. Retrieved from: https://doi.org/10.1016/s1093-3263(97)00106-x

[26] P. Tino, "Spatial representation of symbolic sequences through iterative function systems," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 29, no. 4, pp. 386-393, July 1999, doi: 10.1109/3468.769757.

[27] Fiser A, Tusnády GE, Simon I. Chaos game representation of protein structures. J Mol Graph. 1994 Dec;12(4):302-4, 295. doi: 10.1016/0263-7855(94)80109-6. PMID: 7696222.

[28] Randic, Milan Butina, Darko Zupan, Jure. (2006). Novel 2-D graphical representation of proteins. Chemical Physics Letters. 419. 10.1016/j.cplett.2005.11.091.

[29] Sengupta, D. C., Hill, M. D., Benton, K. R., Banerjee, H. N. (2020). Similarity Studies of Corona Viruses through Chaos Game Representation. Computational molecular bioscience, 10(3), 61–72. Retrieved from: https://doi.org/10.4236/cmb.2020.103004

[30] Tanchotsrinon, W., Lursinsap, C. and Poovorawan, Y. (2015) A High Performance Prediction of HPV Genotypes by Chaos Game Represen-

tation and Singular Value Decomposition. BMC Bioinformatics, 16, 71. Retrieved from: https://doi.org/10.1186/s12859-015-0493-4

[31] Karamichalis, R., Kari, L., Konstantinidis, S., et al. (2015) An Investigation into In- ter- and Intra-Genomic Variations of Graphic Genomic Signatures. BMC Bioin- formatics, 16, Article No. 246. Retrieved from: https://doi.org/10.1186/s12859-015-0655-4

[32] Solis-Reyes, S., Avino, M. and Poon, A. (2018) An Open-Source k-mer Based Ma- chine Learning Tool for Fast and Accurate Subtyping of HIV-1 Genomes. PLoS ONE, 13, e0206409. Retrieved from: https://doi.org/10.1371/journal.pone.0206409

[33] Hu, B., Ge, X., Wang, L., et al. (2015) Bat Origin of Human Coronaviruses. Virology Journal, 12, 221. https://doi.org/10.1186/s12985-015-0422-1

[34] Akhter, S., Bailey, B., Salamon, P., et al. (2013) Applying Shannon's Information Theory to Bacterial and Phage Genomes and Metagenomes. Scientific Reports, 3, Article No. 1033. https://doi.org/10.1038/srep01033

[35] Fractal Foundation (2009) Fractal Pack 1 Educator's Guide Retrieved from: https://fractalfoundation.org/fractivities/FractalPacks-EducatorsGuide.pdf

[36] Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27