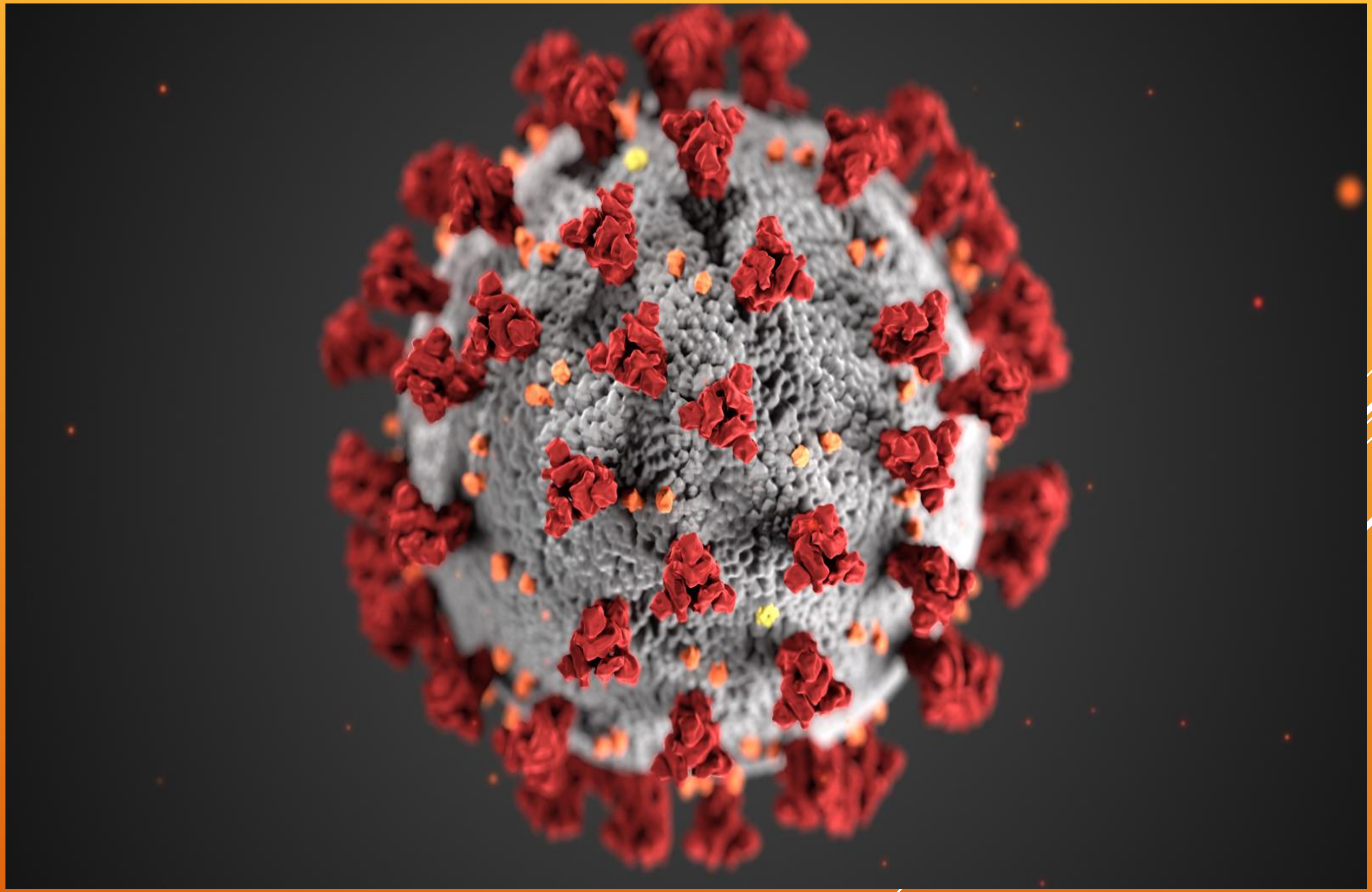
GENOMIC ANALYSIS AND CHAOS GAME REPRESENTATION OF RATG13 AND SARS-COV2

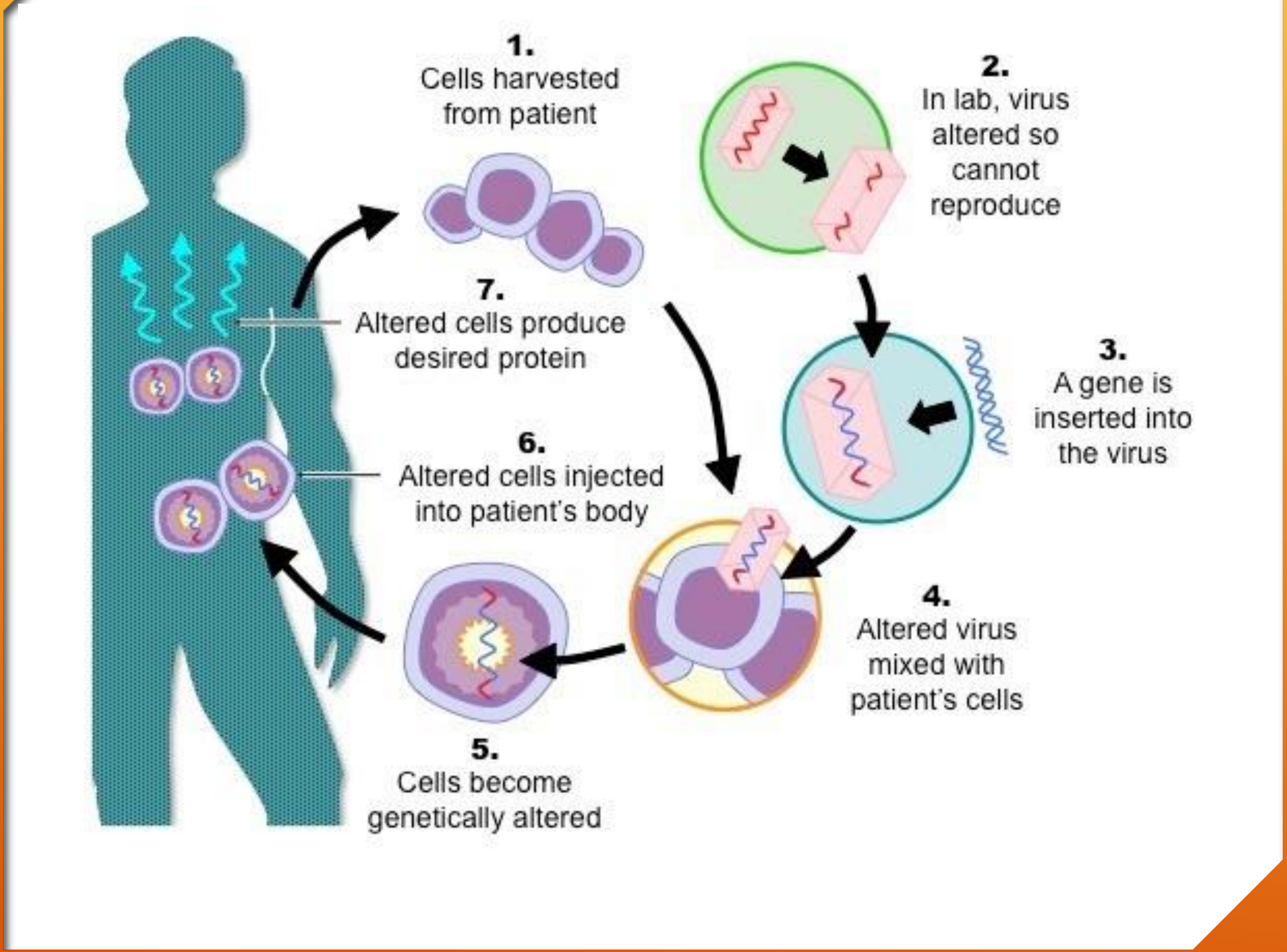
VENKATAKRISHNAN. R

Introduction

- Coronaviruses cause respiratory illnesses in both animals and humans.
- We propose in this research to evaluate the SARS-CoV-2 genomic signature using a combination of nucleotide representations to determine its genetic origin.
- SARS-CoV-2 sequences were compared to Bat coronavirus (Betacoronavirus RaTG)
- The Factors used for the comparion are
    - Most Frequent Kmer
    - Clump
    - GC skew
    - CGR representation



ORIGIN OF REPLICATION (ORIC)


1.
Cells harvested from patient
2.
In lab, virus altered so cannot reproduce
3.
A gene is inserted into the virus
4.
Altered virus mixed with patient's cells
5.
Cells become genetically altered
6.
Altered cells injected into patient's body
7.
Altered cells produce desired protein

- Replication starts in a genomic region termed the replication origin (denoted oriC) and is carried out by DNA polymerases, which are molecular copy machines.
- Finding oriC is crucial not only for understanding how cells replicate

K-Mer

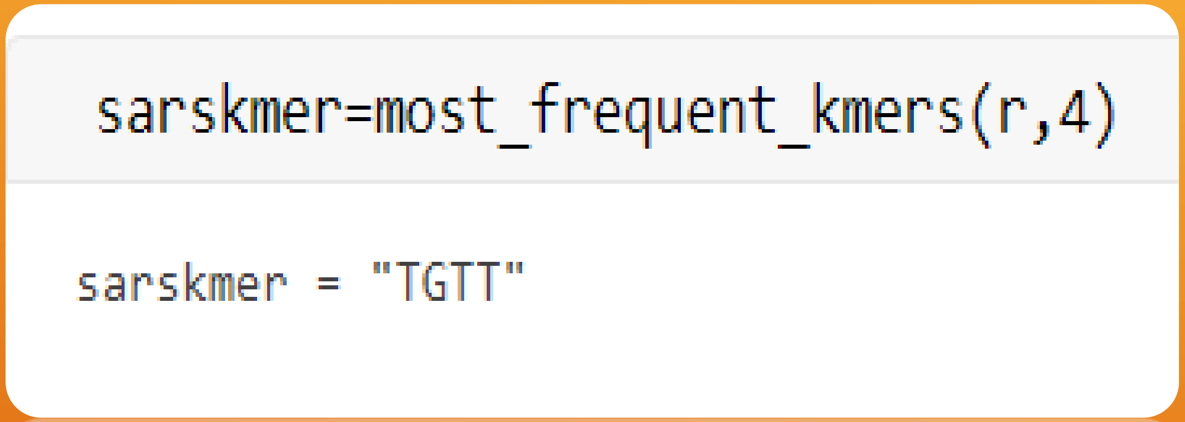
- K-mers are substrings of length k in a given string in simple terms (can be DNA, RNA, protein, or any string sequence).
- For example, the 11-nucleotide DNA sequence "ACGAGGTACGA." Let's see if we can find all of the 4-mers (four-letter substrings) in this DNA sequence.
- If length of DNA sequence is ‘N’ then we get (N - k+1) k-mers
- In our case N=11, k=4, so total k-mers we have is 11-4+1=8 k-mers

N=11, k=4;
11-4+1=8.
8 k-mers

Most frequent k-mer

- If Pattern maximises COUNT(Text, Pattern) among all k-mers, it is said to be the most frequent k-mer in the Text.
- The most common 5-mer for Text = ACAACTATGCATACTATCGGGAACTATCCT is ACTAT.
- Whereas the most common 3-mer for Text = CGATATATCCATAG is ATA.

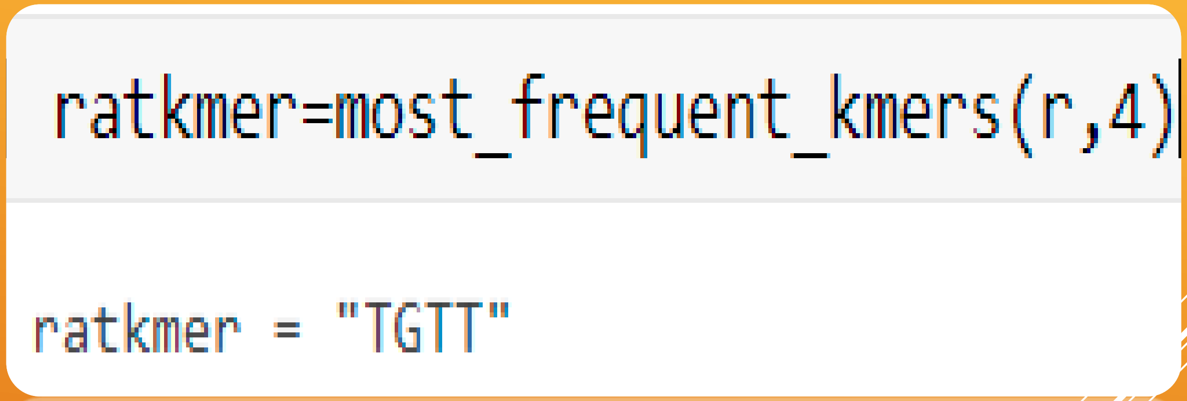
MOST FREQUENT K-MER (4MER) COMPARISON



```
sarskmer=most_frequent_kmers(r,4)
```

```
sarskmer = "TGTT"
```

SARS-CoV-2



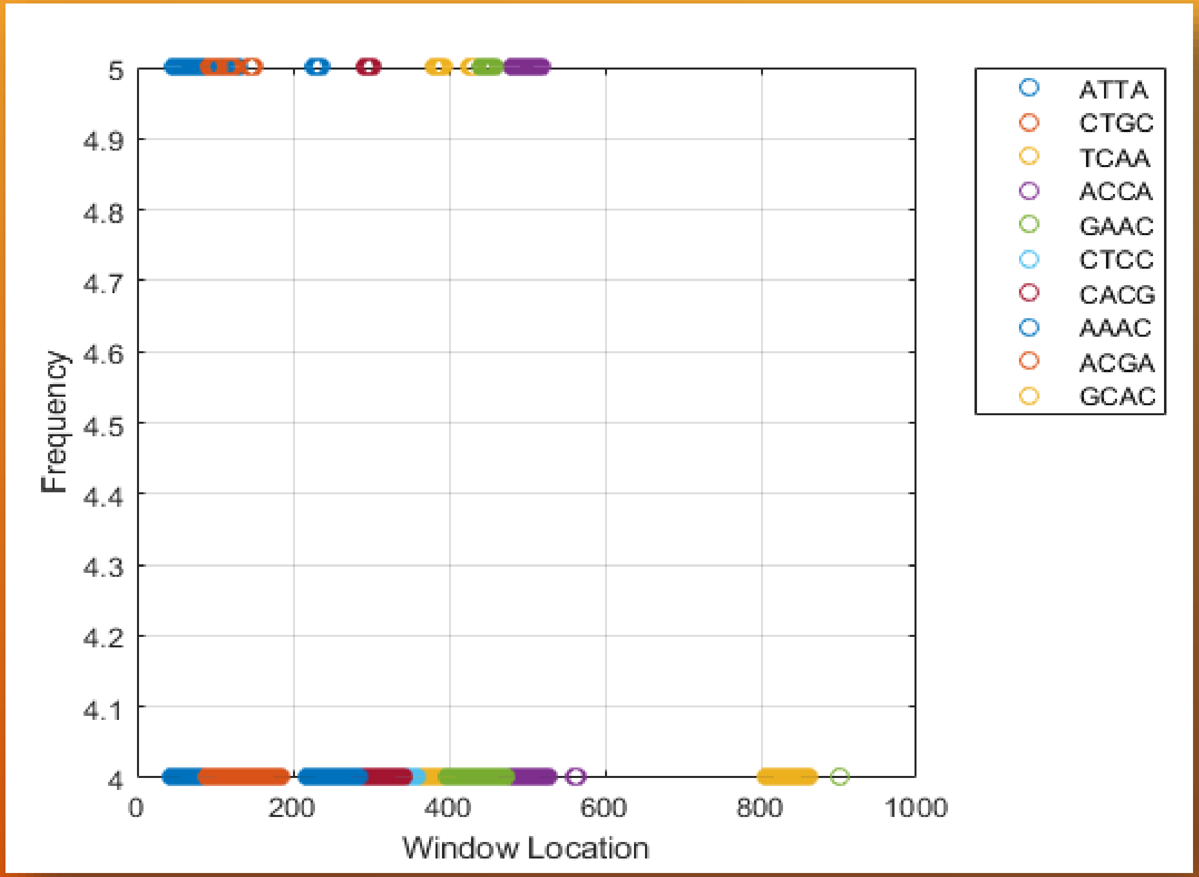
```
ratkmer=most_frequent_kmers(r,4)
```

```
ratkmer = "TGTT"
```

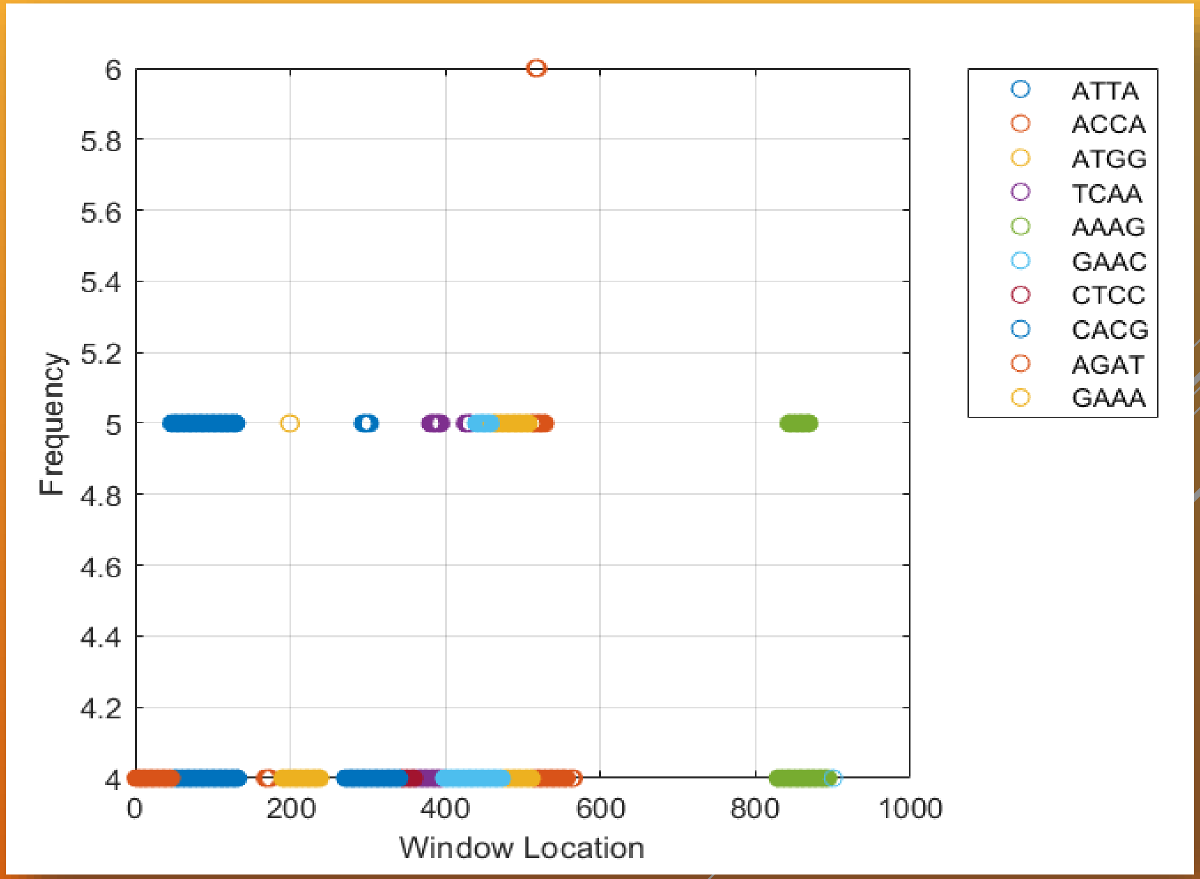
Betacoronavirus RaTG

Comparison of frequency of 4 mers in

SARS-CoV-2 and Betacoronavirus RaTG

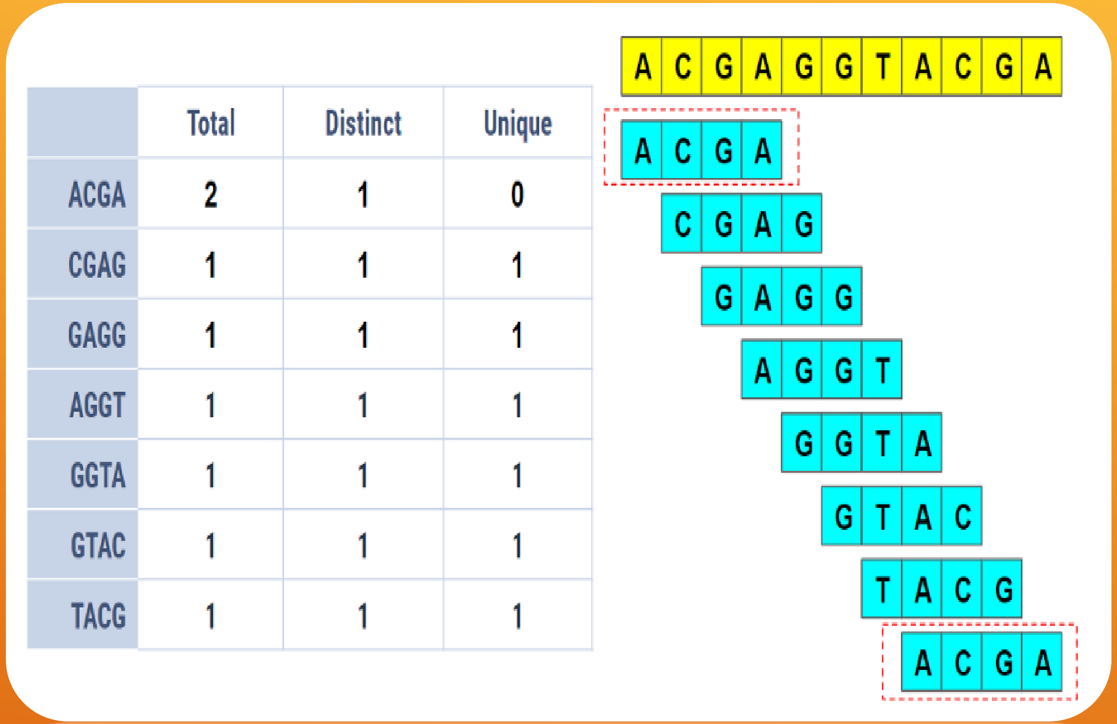


SARS-CoV-2



Betacoronavirus RaTG

k-mer counting problem



	Total	Distinct	Unique
ACGA	2	1	0
CGAG	1	1	1
GAGG	1	1	1
AGGT	1	1	1
GGTA	1	1	1
GTAC	1	1	1
TACG	1	1	1

1. Total count

2. Distinct count

3. Unique count

Many bioinformatics applications that evaluate sequencing data include k-mer counting as a crucial step.

Counting the number of substrings with length k in a string S, or a group of strings, where k is a positive integer, is known as k-mer counting.

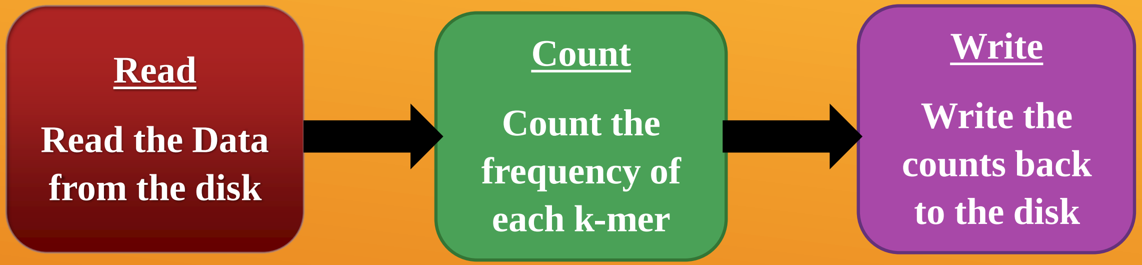
Application Of k-mer counting:

K-mer counting is utilised in a variety of applications:

- **Genome assembly**
- **Sequence alignment**
- **Sequence clustering**
- **Genome size estimation**
- **Repeat identification**
- **Error correction of sequencing reads**

k-mer counting in Computer Science

K-mer counting has only three main phases at the most basic level



Read

Read the Data from the disk

Count

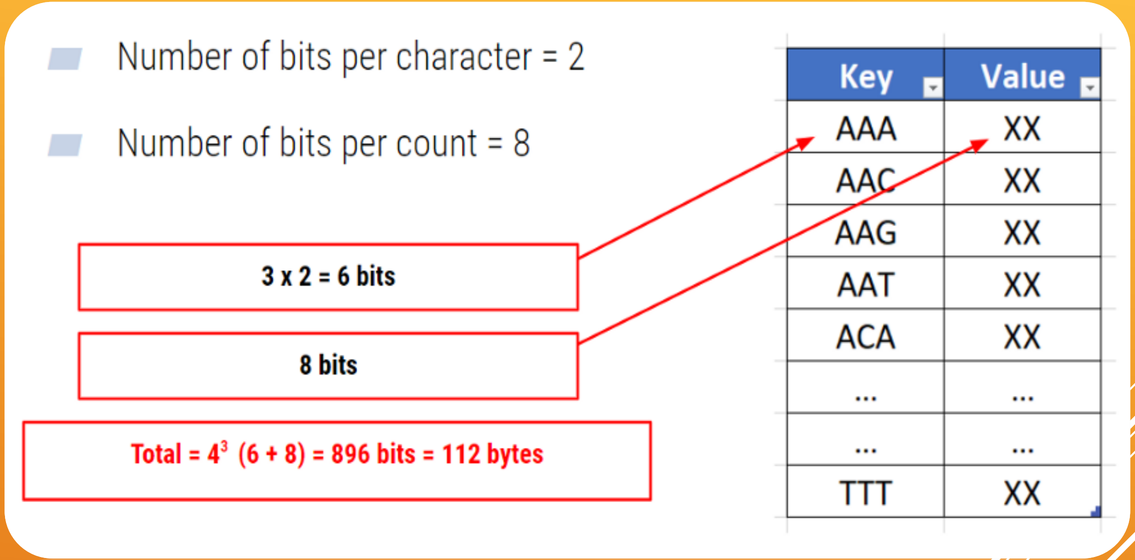
Count the frequency of each k-mer

Write

Write the counts back to the disk

- **Because there are only four nucleotides (A, C, G, and T), each character may be represented with only two bits. For the count, we'll use 1byte / 8bits.**
- **You'll need about 6000GB of RAM for k=20, which is enormous!!! This memory barrier makes k-mer counting a difficult yet fascinating task for computer scientists.**

We have k-mers and the counts that go with them. Assume k=3.



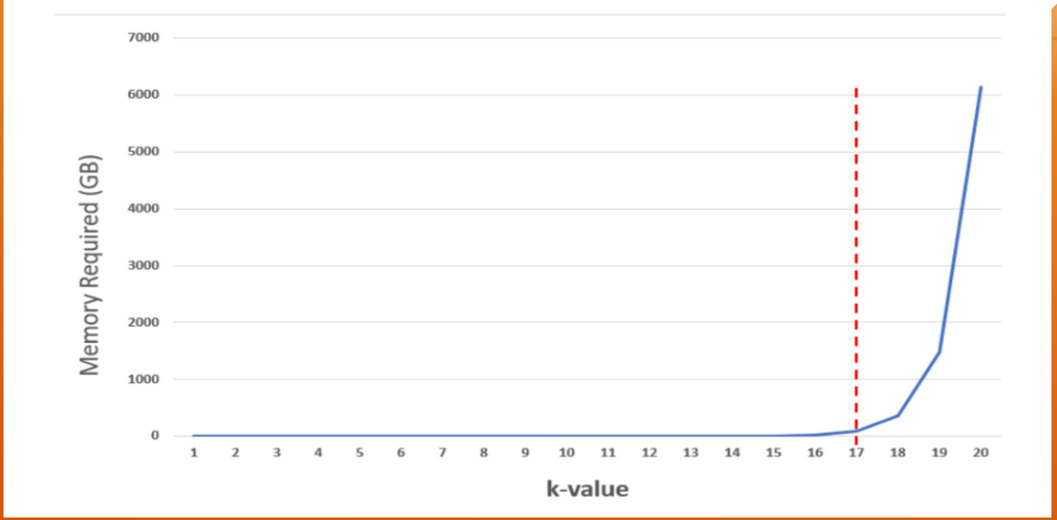
- Number of bits per character = 2
- Number of bits per count = 8

3 x 2 = 6 bits

8 bits

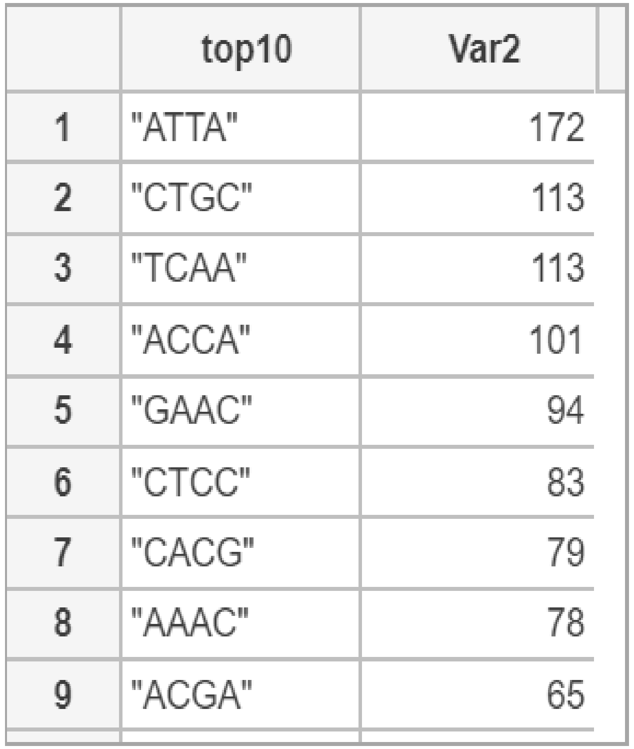
Total = 4^3 (6 + 8) = 896 bits = 112 bytes

Key	Value
AAA	XX
AAC	XX
AAG	XX
AAT	XX
ACA	XX
...	...
...	...
TTT	XX



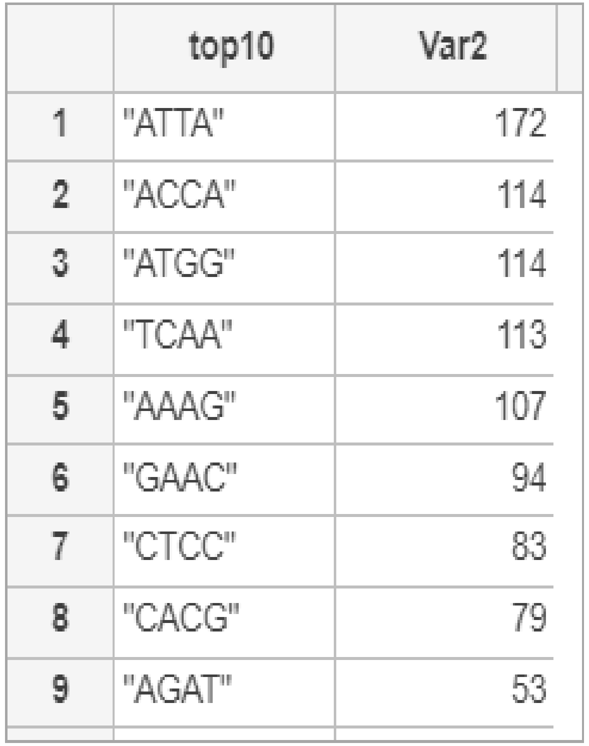
Clumps

- If a k-mer appears multiple times within a brief interval of the genome, it is considered a "clump."
- In more technical terms, if there is an interval of Genome of length L in which this k-mer appears at least t times, a k-mer Pattern creates a (L, t)-clump inside a (longer) string Genome.



	top10	Var2
1	"ATTA"	172
2	"CTGC"	113
3	"TCAA"	113
4	"ACCA"	101
5	"GAAC"	94
6	"CTCC"	83
7	"CACG"	79
8	"AAAC"	78
9	"ACGA"	65

SARS-CoV-2



	top10	Var2
1	"ATTA"	172
2	"ACCA"	114
3	"ATGG"	114
4	"TCAA"	113
5	"AAAG"	107
6	"GAAC"	94
7	"CTCC"	83
8	"CACG"	79
9	"AGAT"	53

Betacoronavirus RaTG

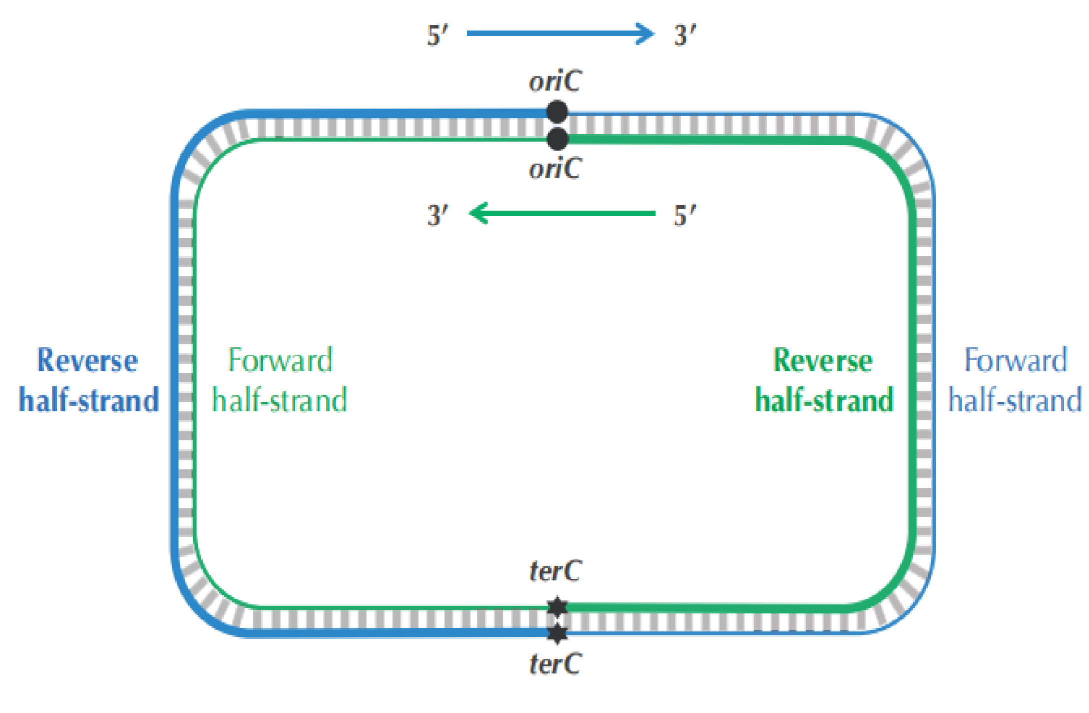
Peculiar Statistics of the Forward and Reverse Half-Strands

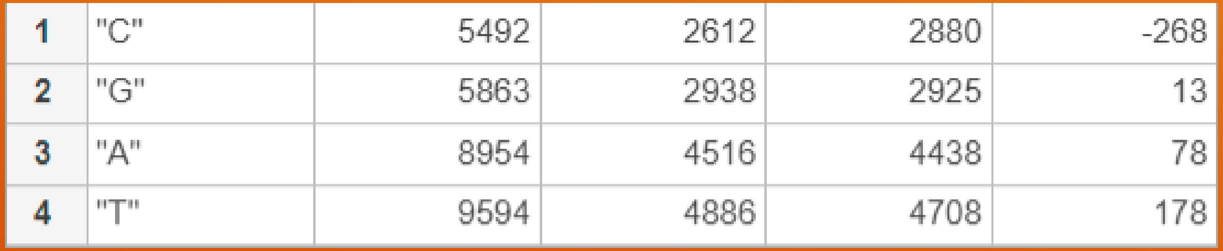
Two of these half-strands are traversed from *oriC* to *terC* in the 5 to 3direction and are thus called forward half-strands

The other two half-strands are traversed from *oriC* to *terC* in the 3 to 5 direction and are thus called reverse half-strands

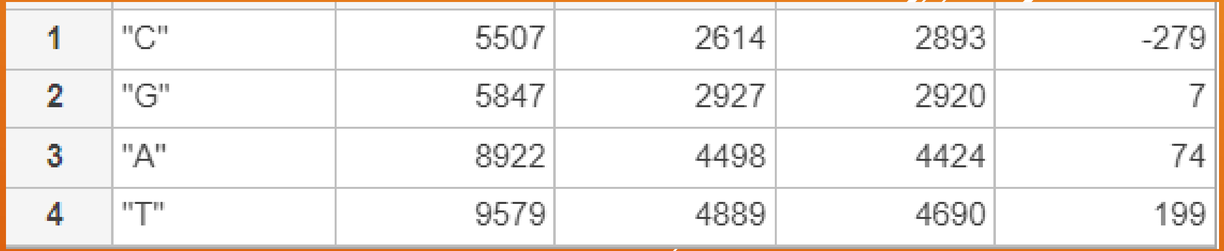
Leading half-strand

Lagging half-strand





1	"C"	5492	2612	2880	-268
2	"G"	5863	2938	2925	13
3	"A"	8954	4516	4438	78
4	"T"	9594	4886	4708	178



1	"C"	5507	2614	2893	-279
2	"G"	5847	2927	2920	7
3	"A"	8922	4498	4424	74
4	"T"	9579	4889	4690	199

SARS-CoV-2

Betacoronavirus RaTG

GC Skew:

GC skew is a measure of the strand asymmetry in the distribution of Guanines and Cytosines.
Positive GC skew represents richness of G over C and the negative GC Skew represents richness of C over G.

The GC skew is proven to be useful as the indicator of the DNA leading strand, lagging strand, replication origin, and replication terminal.

Most prokaryotes and archaea contain only one DNA replication origin.
The GC skew is positive in the leading strand and negative in the lagging strand respectively.

The peaks in the cumulative GC skew diagram correlate to the switch points (terminus or origin).

The terminal corresponds to the largest value of the cumulative skew, whereas the origin of replication corresponds to the least value.

GC Skew:

The skew diagram is defined by plotting SKEW_i (Genome) as i ranges from 0 to |Genome|, where SKEW_0 (Genome) is set equal to zero.

The complete genome is commonly plotted 5' to 3' using an arbitrary start and strand in this approach.

Algorithm:

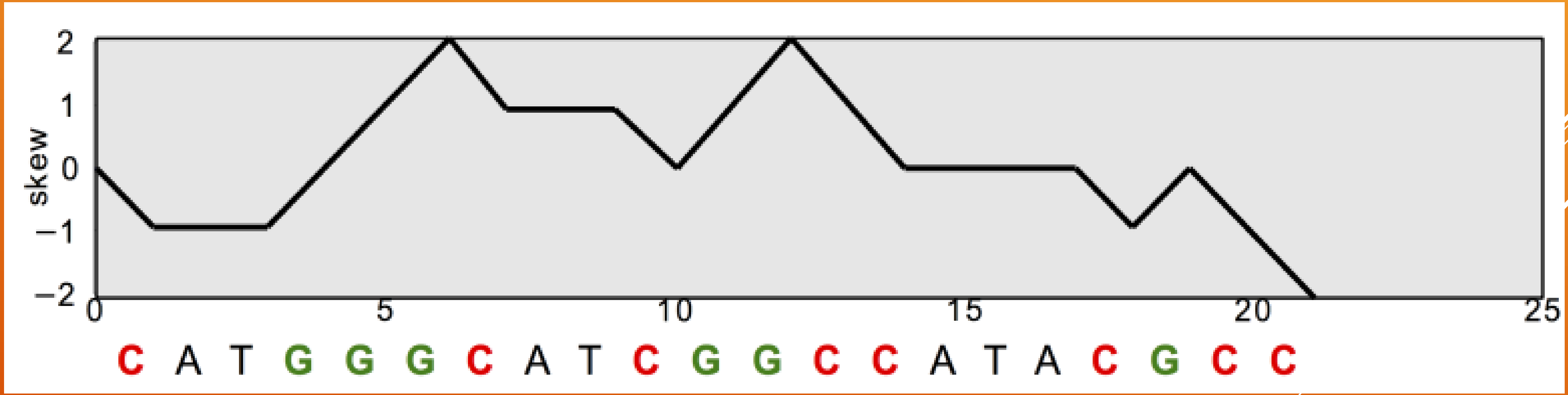
if nucleotide = G

SKEWi+1(Genome) =SKEWi(Genome)+1

if nucleotide = C

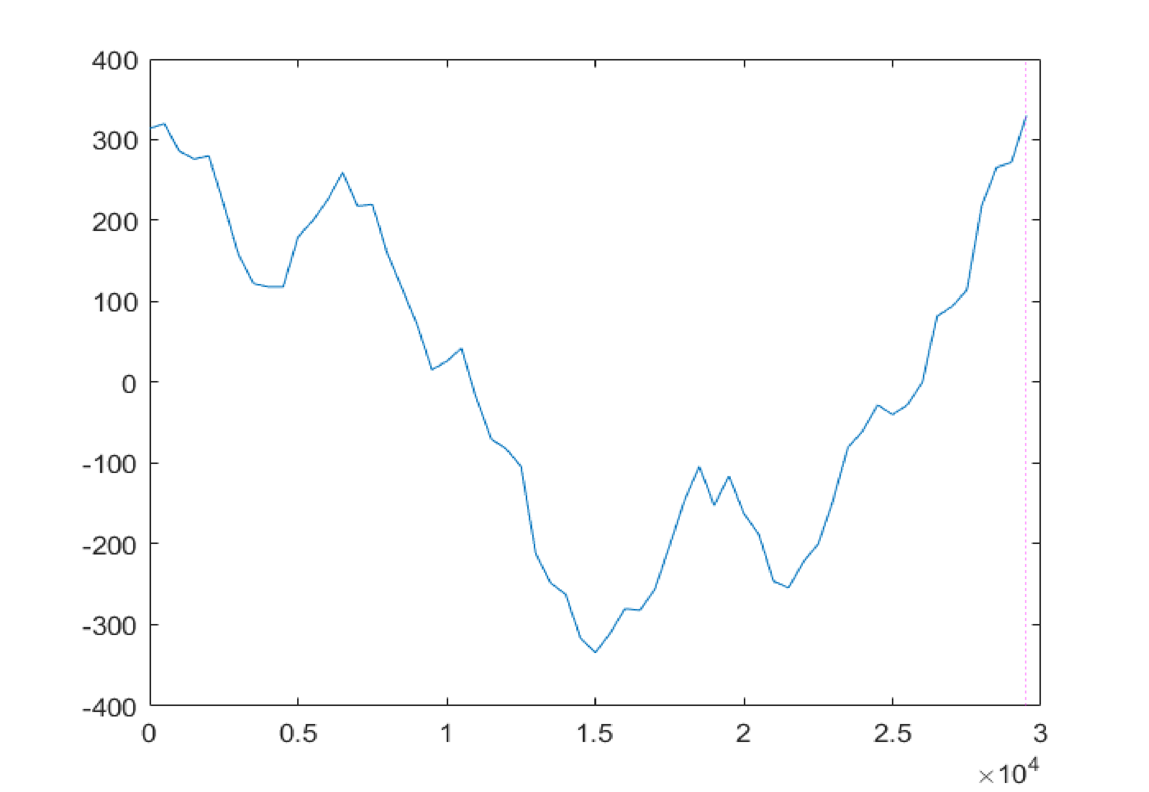
SKEWi+1(Genome) =SKEWi(Genome)-1

$$\text{GC skew} = (G - C)/(G + C)$$

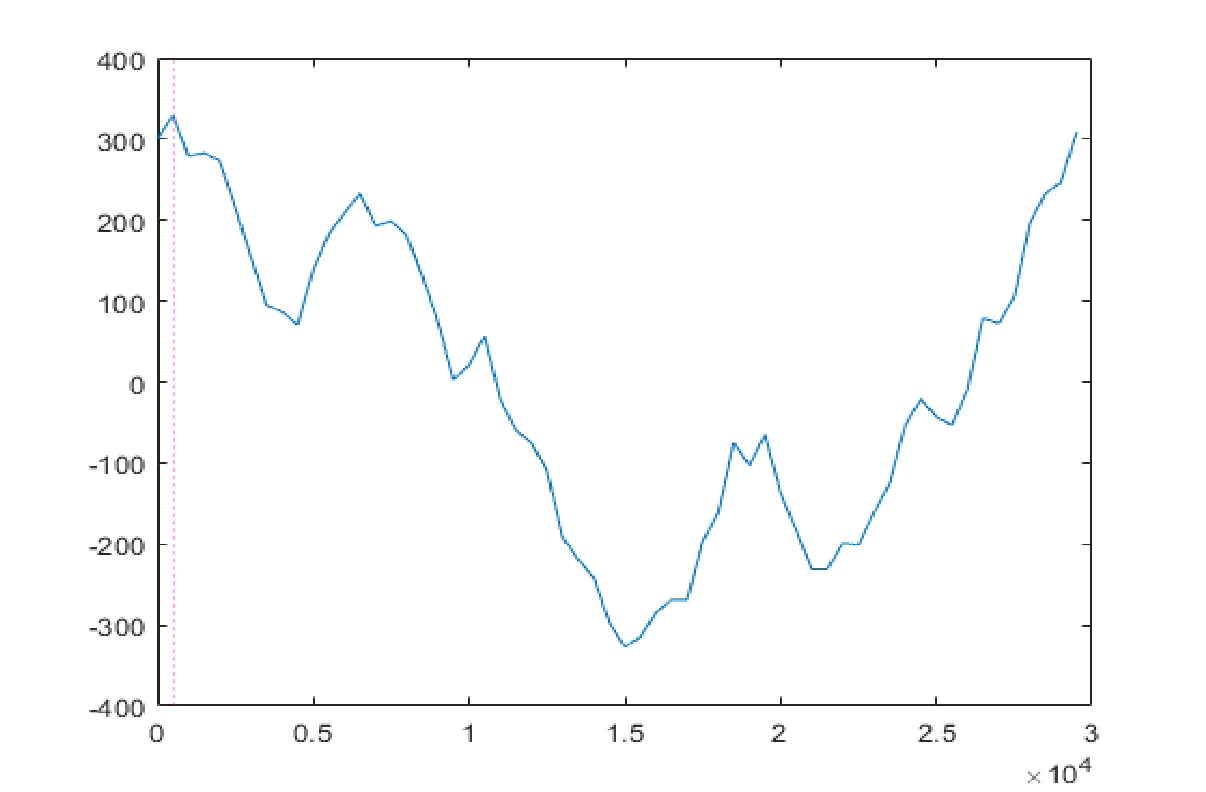


The skew diagram for Genome = CATGGGCATCGGCCATACGCC.

Comparison of GC SKEW of 4-mers in SARS-CoV-2 and Betacoronavirus RaTG



SARS-CoV-2



Betacoronavirus RaTG

CGR Representation

- For DNA sequences, the CGR approach was presented as a unique and scale-independent representation. Fractal landscapes are created by mapping the genomic sequence using the frequency chaos game representation (FCGR).
- This iterative mapping technique assigns a unique coordinate in a 2-dimensional space to each nucleotide in DNA or amino acid in a protein (x, y).
- This two-dimensional graphic depicts the distribution of dots as a 0–1 square matrix, with 0 representing an empty coordinate and 1 representing a dot. As a result, a point represents an element in the nth position of the DNA sequence (Seq = S1, S2,..., SL) made of L nucleotides (A, T, C, or G).
- This point is continually plotted midway between the previous plotted point and the segment connecting the read letter Sn's vertex.

Algorithm:

1.Input: a genomic sequence with length N

2.Intialize step: creating a square with each corner:

Adenine (A) with coordinates (x_A =0,y_A=0)

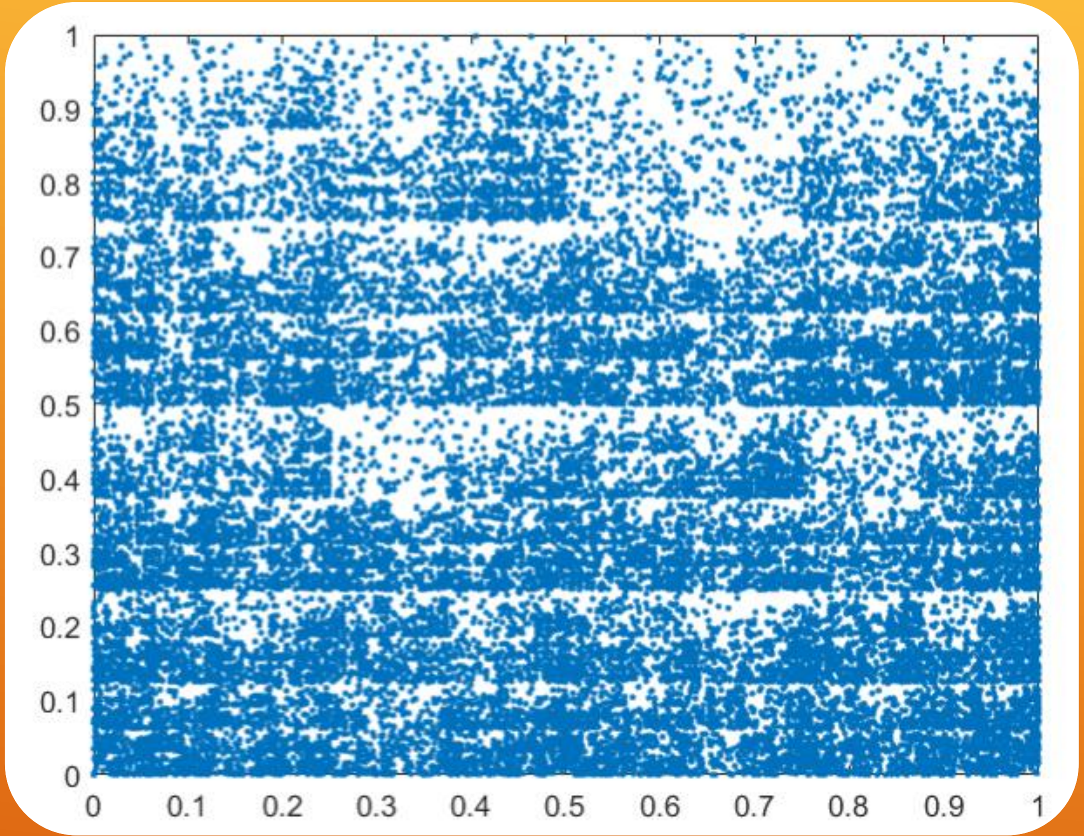
Thymine(T) with coordinates (x_T=1, y_T=0)

Cytosine(T) with coordinates (x_C=0, y_T=1)

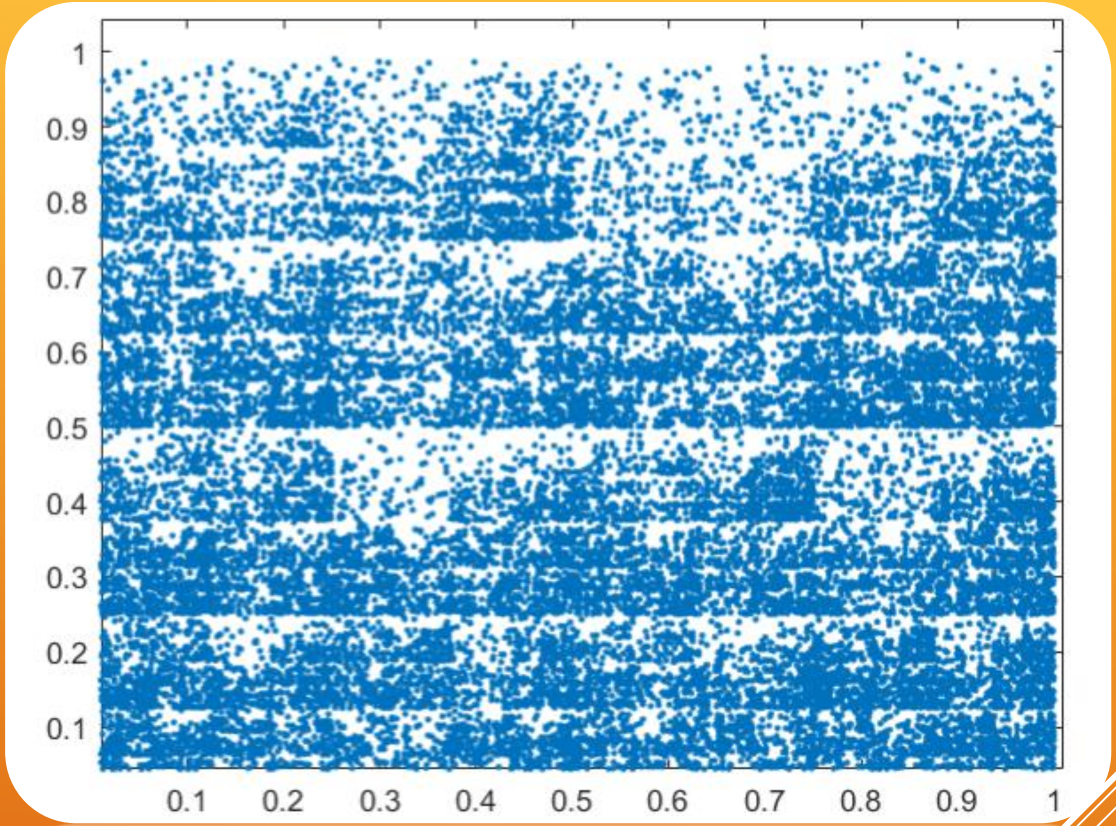
Guanine(T) with coordinates (x_G=1, y_G=1)

3.starting point :X_0 (x_0=0.5 ,y_0=0.5)

```
4) case 1
A: place the dot at $X1=0.5(xA+x0)$ ;$Y1=0.5(yA+y0)$ ;
T: place the dot at $X1=0.5(xT+x0)$ ;$Y1=0.5(yT+y0)$ ;
C: place the dot at $X1=0.5(xC+x0)$ ;$Y1=0.5(yC+y0)$ ;
G: place the dot at $X1=0.5(xG+x0)$ ;$Y1=0.5(yG+y0)$ ;
End Case
For the other nucleotides : from 2: N
Case is
A: place the dot at $Xi=0.5(xA+xi−1)$ ;$Yi=0.5(yA+yi−1)$ ;
T: place the dot at $Xi=0.5(xT+xi−1)$ ;$Yi=0.5(yT+yi−1)$ ;
C: place the dot at $Xi=0.5(xC+xi−1)$ ;$Yi=0.5(yC+yi−1)$ ;
G: place the dot at $Xi=0.5(xG+xi−1)$ ;$Yi=0.5(yG+yi−1)$ ;
EndFor
```



SARS-CoV-2



Betacoronavirus RaTG

Comparison of CGR of 4-mers in SARS-CoV-2 and Betacoronavirus RaTG

Conclusion

The findings demonstrate that the Bat coronaviruses are the most closely related to SARS-CoV-2, with 96 percent identity across the genome,

THANK YOU

VENKATAKRISHNAN. R