Frequency representations of COVID-19 and Omicron using Quadrature Phase-Shift Keying (QPSK)

Khalid Mahmood Aamir*, Mahwish Ilyas, Taj Mahmood
Department of Computer Science and IT, University Of Sargodha, Pakistan
Emails: khalid.aamir@uos.edu.pk, mahwishilyas@gmail.com, tajm119629@gmail.com

Abstract

Omicron is a covid family virus of COVID-19 and Delta variant. The Omicron (B.1.1.529.) variant of COVID-19 is an extraordinary flow of infections globally and deadly, affecting the masses. The B.1.1.529 variant was first identified to WHO on November 24, 2021, from South Africa. In South Africa, the epidemiological condition has been determined by three different peaks in reported cases, the most recent of which was dominated by the Delta variant. Infections have risen sharply, corresponding with the discovery of the B.1.1.529 variant. The variant contains many mutations, some of which are potentially harmful. Preliminary research suggests that this variant has a higher risk of reinfection than other variants of concern. Nowadays, many scientists worldwide focus on problems that either improve existing methods used in DNA computing or suggest a new manner with a DNA computing approach. Many researchers are working on analyzing several aspects of Omicron from diverse fields. We have developed a frequency representation of Omicron to visualize its properties in the spectral domain.

Introduction

The WHO declared Omicron (B.1.1.529) as a novel severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) strain of concern on November 26, 2021. This variation has an extremely high number of mutations, 32, on the spike (S) protein, the principal antigenic target of antibodies generated by viruses or immunization. The deadly Delta variation contains only 5 S protein mutations, posing a significant potential global risk and spreading internationally. As a result, the "panic button" has been pressed in multiple cases worldwide, and many nations have instituted travel restrictions to avoid the rapid spread of the Omicron strain.

In signal analysis theory, time-series data is represented in the time domain. When data features are not visible in the time domain, we transform the data to the frequency domain. Certain features that are not visible in the time domain become visible in the frequency domain. The Fourier transform is a popular tool for representing data in the frequency domain.

A DNA is a string of characters composed of alphabets from the set {A, T, C, G} with the only primitive operations of matching and counting performed. This limitation prevents rigorous mathematical operations and transformations from applying to any genomic data.

By mapping strings to complex numbers, we present a computational model of DNA. This model has enabled us to perform any mathematical transformation operations. We presented frequency domain representations of Omicron genomes built with this computational model of DNA.

The rest of the paper is organized as follows. Next section provides an over view of the three types of mappings of DNA to the fields of real/complex numbers. Following this, methodology is described which provides a mapping and a way to use discrete Fourier transform to get frequency representation of a DNA sequence. After this results have been shown and discussed.

Literature Review

The representation of biomolecular sequences as strings of characters makes frequency-domain analysis difficult. If each of these characters is given a numerical value, the resulting numerical sequences are easily processed digital signal processing techniques.

The conversion of a DNA string of characters into numerical form is required for power spectral density (PSD) estimation [1]. To achieve this goal, several researchers have defined modeling methodologies. Existing works are categorized into three types with respect to mappings listed below.

1. Mapping to binary sequences
2. Mapping to field of real numbers
3. Mapping to field of complex numbers

1. Mapping to binary sequences

The Voss[2] representation is a widely used technique that generates $x_a[n]$, $x_t[n]$, $x_c[n]$, and $x_g[n]$ which are four binary indicator sequences, each of which takes a value of 1 or 0 at position n, depending on whether the associated character appears or not.

$$x_c[n] + x_g[n] + x_a[n] + x_t[n] = 1 \text{ for everyone } n.$$

As these indicator sequences are redundant, they demonstrate redundancies.

In [3], a new computational and visual technique for analyzing biomolecular sequences has been demonstrated. D Anastassiou et al. show that providing suitable (complex, in general) numerical values to each character, for digital signal processing of biomolecular sequences, gives a set of innovative and helpful numerical sequences[4].

Assuming the letters 'A,' 'T,' 'C,' and 'G' are all mapped to numbers a, t, c and g, in an N-length deoxyribonucleic acid (DNA) sequence. This results into the following numerical sequence.

$$x[n] = a.u_A[n] + t.u_T[n] + c.u_C[n] + g.u_G[n] \qquad n = 0, 1, 2, \ldots N\text{-}1$$

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are binary indicator sequences. This is dependent on whether the relevant character is present at position n, takes the value of 1 or 0[6]. The above equation is subject to the following condition.

$$u_A[n] + u_T[n] + u_C[n] + u_G[n] = 1 \qquad \text{for all } n.$$

The most dominant signal in coding sections of genomic sequences is a three-base periodicity [5]. Our goal is to determine the periodicity using Fourier techniques. Therefore, we are working on developing a method to recognize coding areas in DNA.

A nucleotide sequence of N nucleotides can be represented referred to this as a symbolic string, $\{x_i, i = 1, 2, \ldots, N\}$, here x_i can be any of the four characters G, C, A or T, and indicates that a certain nucleotide is present in location i. A mapping is defined below.

$$U_\alpha(x_i) = \begin{cases} 1 & \text{if } x_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Using the operators U_G, U_C, U_A and U_T, on a strand of DNA in this order generates four binary sequences shown below in Table 1.

Table 1. Application of binary operators on DNA string generates four binary operators.

Sequence	C	C	A	T	A	T	G	A	A	T	C	T
Apply U_A	0	0	1	0	1	0	0	1	1	0	0	0
Apply U_T	0	0	0	1	0	1	0	0	0	1	0	1
Apply U_G	0	0	0	0	0	0	1	0	0	0	0	0
Apply Uc	1	1	0	0	0	0	0	0	0	0	1	0

2. Mapping to the field of real numbers

Fixed mapping and mapping depending based on some sort of optimality criterion are the two types of numerical mapping. Binary integer and complex representations are examples of fixed mappings. In [6]a real-number mapping rule is shown in Figure **1**.

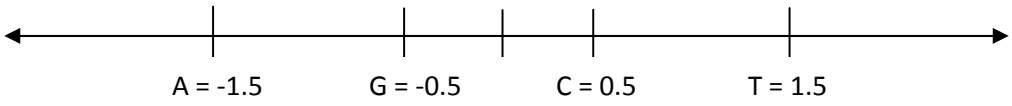

Figure 1. Bijective mapping of nucleotides onto real numbers.

A generalized form of this mapping f for two real numbers α and β is as follows.

f: A → -α

f: G → -β

f: C → β

f: T → α

2. Mapping to the field of complex numbers

A generalized form of this mapping f is as follows.

$$f: \{A, T, G, C\} \rightarrow \{1, -1, i, -i\}$$

Specifications of several forms of mapping f are listed in Table 2.

Table 2. Complex mappings for DNA.

Name of Method	A	T	C	G	Statements /Remarks
K- Quaternary Code-III	+1	+i	-i	-1	Rao and Shepherd
K-Quaternary Code-I	+1	+i	-1	-i	Kwan et al.
Quaternary Code	-1	+i	-i	+1	Manidipa Roy et al.

K-Quaternary Code-I was the best appealing, according to the primary findings of Kwan et. al.[7], however, Rao and Shepherd [8]claimed the "K-Quaternary Code-III" be more acceptable.

To provide projected exons with location accuracy, the Manidipa Roy[9]et al. developed a mapping rule (given in Table 2) in which the Y-axis of the K-Quaternary Code-III has been reversed, assigning numerical values "a = -1", "c = -i", "g = 1", and "t = +i" to the nucleotide sequence. The complex mapping was proven to be one of the most successful mapping rules.

Cheever et al. [10]were the first to map DNA characters into the complex number plane, i.e., A to "1", T to" -1", G to "i" (where $i = \sqrt{-1}$) and C to " $-i$". They made an attempt to locate similarities between two sequences by comparing and contrasting their complex sequences using this mapping.

Fixed mapping and mapping depending based on some sort of optimality criterion are the two types of numerical mapping that are found. Binary integer and complex representations are examples of fixed mappings. Mapping rule is based on the complex mapping's complement attribute.

Methodology

The methodology consists of two steps: a computational model of DNA using quadrature phase-shift keying and discrete Fourier transform (DFT) computation.

Computational Model of DNA

There are four nucleotides, or bases, in DNA: adenine **(A),** cytosine **(C),** guanine **(G),** and thymine **(T)**. These bases form specific pairs (A with T and G with C). From this, we construct the following rules.

1. A is orthogonal to G and C
2. T is orthogonal to G and C
3. G is orthogonal to A and T
4. C is orthogonal to A and T

All the bases have approximately equal masses.

These rules are shown graphically in Figure **2**.

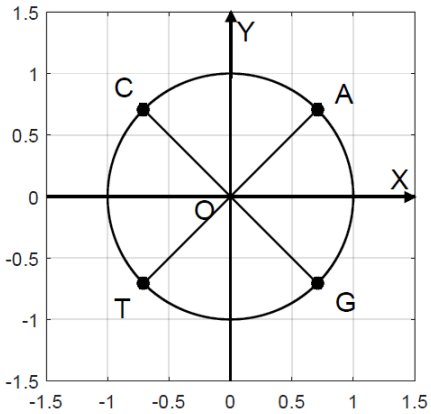

Figure 2. Graphical representation of a computational model of DNA using quadrature phase-shift keying (QPSK).

Anticlockwise angles of several line segments (shown in Figure **2**) are given in Table **3**.

Table 3. Mapping of nucleotides to complex numbers (quadrature phase-shift keying).

Line segment	Angle	Mapping	Nucleotide
XOA	π/4	$e^{j\pi/4}$	A
XOC	3 π/4	$e^{j3\pi/4}$	C
XOT	5 π/4	$e^{j5\pi/4}$	T
XOG	7 π/4	$e^{j7\pi/4}$	G

Frequency transformation

Let g be a genome. Using mapping given in Table **3**, we get a discrete sequence, say $f(n)$. The discrete Fourier transform $F(k)$ of $f(n)$ is defined below.

$$F(k) = \sum_{n=0}^{N-1} f(n).e^{-j2\pi nk/N} \quad (1)$$

Where N is the length of the genome g and $k = 0, 1, 2, \ldots, N\text{-}1$, the $F(k)$ is the discrete samples at frequency index k. The power spectrum $S(k)$ of the $F(k)$ is defined as $|F(k)|$ where $|x|$ represents the absolute value of a complex number x. Therefore, $S(k)$ is computed as:

$$S(k) = \left|\sum_{n=0}^{N-1} f(n).e^{-j2\pi nk/N}\right| \quad (2)$$

For $k = 0, 1, 2, \ldots, N\text{-}1$. The index k is mapped to the normalized frequency by $f_i = k/N$.

Results

Data: We computed power spectra $S(k)$ of Omicron and Corona genome sequences (accession numbers are given in Table 4, respectively).

For the computation of $S(k)$, we have padded signals with zeros so that the total length of the sequence becomes 32768 (2^{15}). This zero-padding enabled us to use the fast Fourier transform algorithm to compute $S(k)$ quickly for the computation of discrete Fourier transform. Zeros padding does not affect the resolution of S(k); however, it produces the effect of averaging.

From frequency representations (shown in Figure 3) of Omicron genomes listed in Table 4, this is obvious that certain dominant modes of frequencies are observed (Figure 5). The threshold was set to 0.2 for extraction of dominant modes.

From frequency representations (shown in Figure 4) of Corona genomes listed in Table 4, this is obvious that certain dominant modes of frequencies are observed (shown in Figure 6 and Figure 7) with thresholds 0.2 and 0.4 (respectively) for extraction of dominant modes.

Table 4. Accession numbers and corresponding lengths of genomes.

Omicron		Covid19	
Accession No.	Length	Accession No.	Length
OM424289.1	29752	MN970004.1	290
OM424290.1	29752	OL535405.1	29900
OM424291.1	29752	OU967438.1	29900
OM319696.1	29779	OL477009.1	675
OM341396.1	3813	OL671531.1	676
OM311576.1	29731	MT503048.1	676

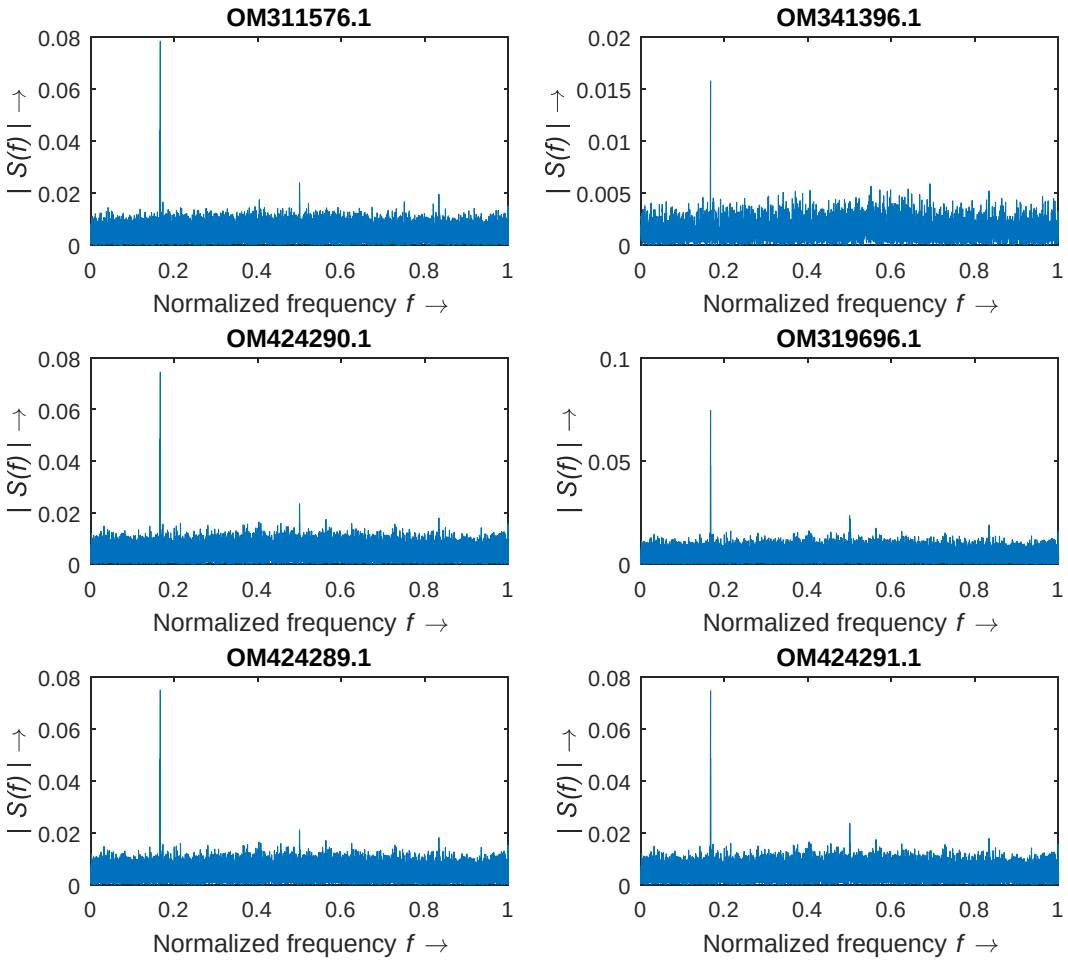

Figure 3. Frequency representations of genomes are listed in Table 1.

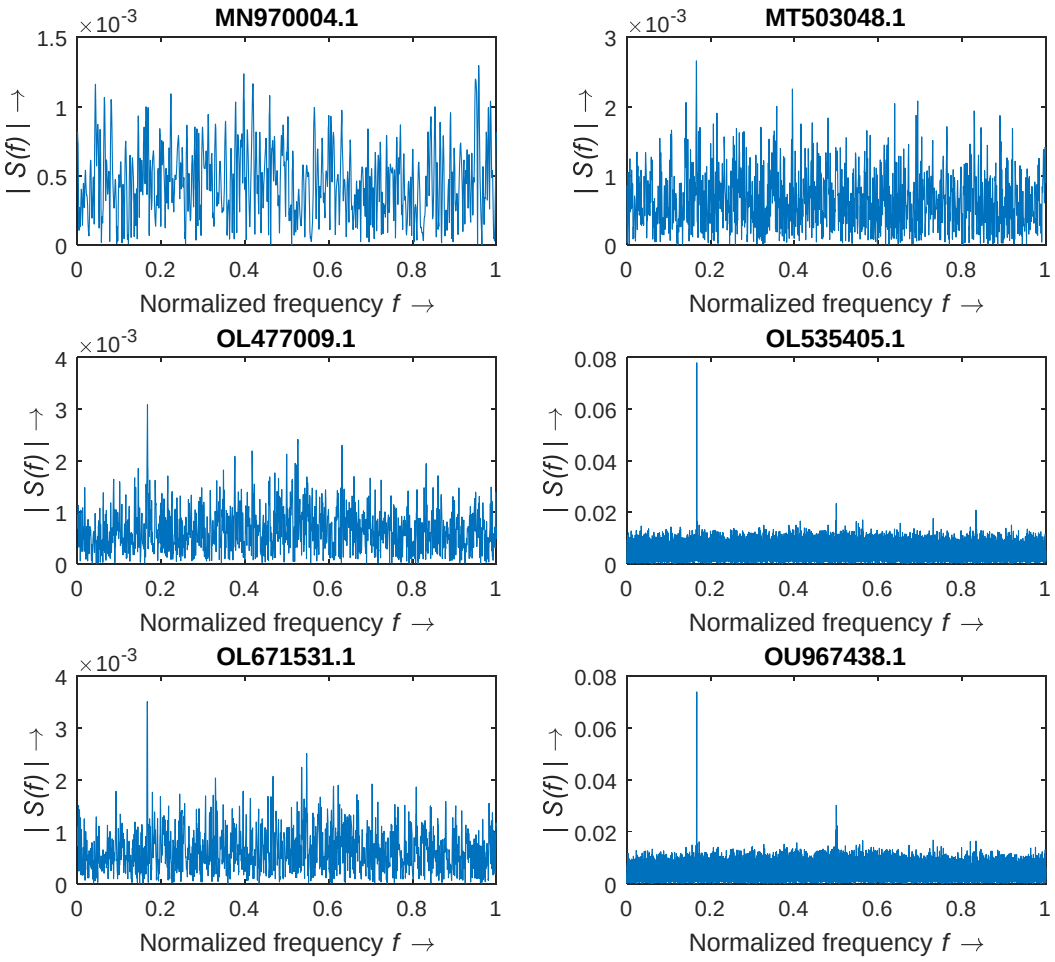

Figure 4. Frequency representations of genomes are listed in Table 2.

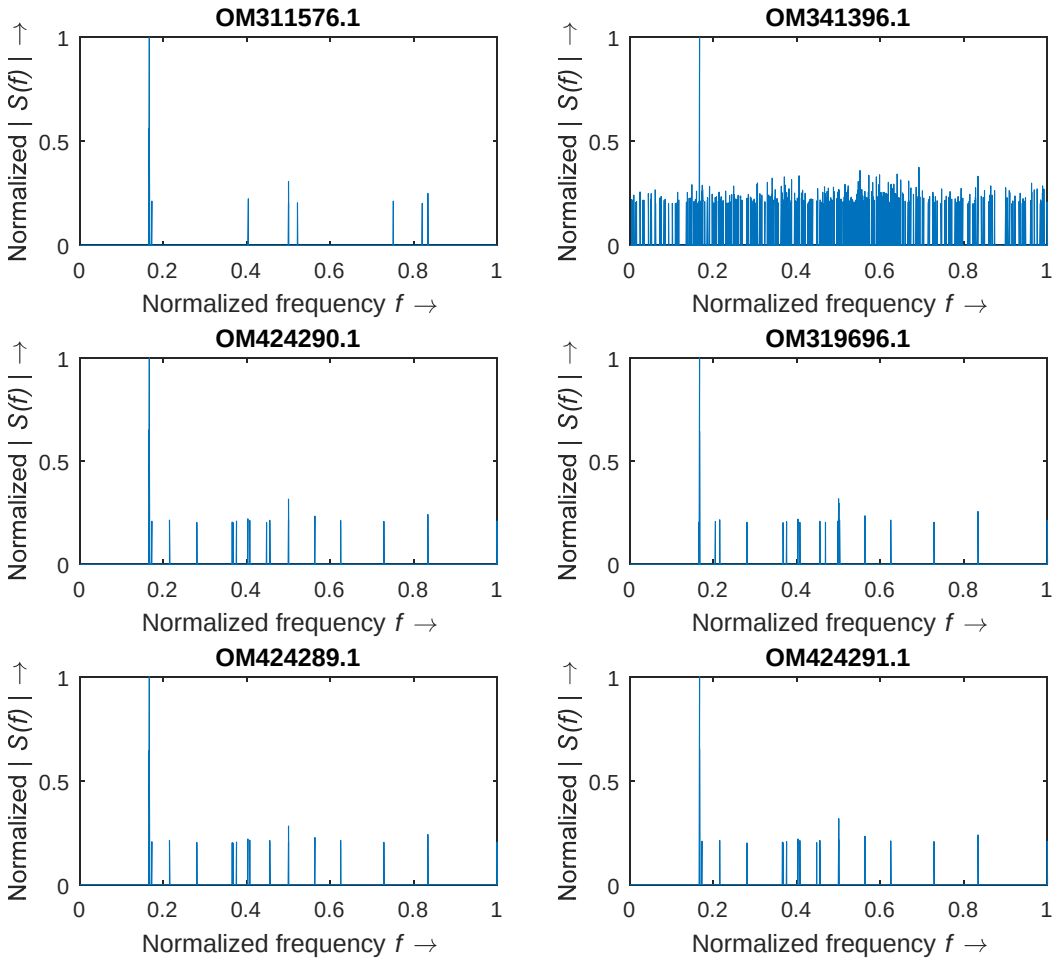

Figure 5. Dominant modes in frequency plots are shown in Fig. 1, with a threshold 0.2.

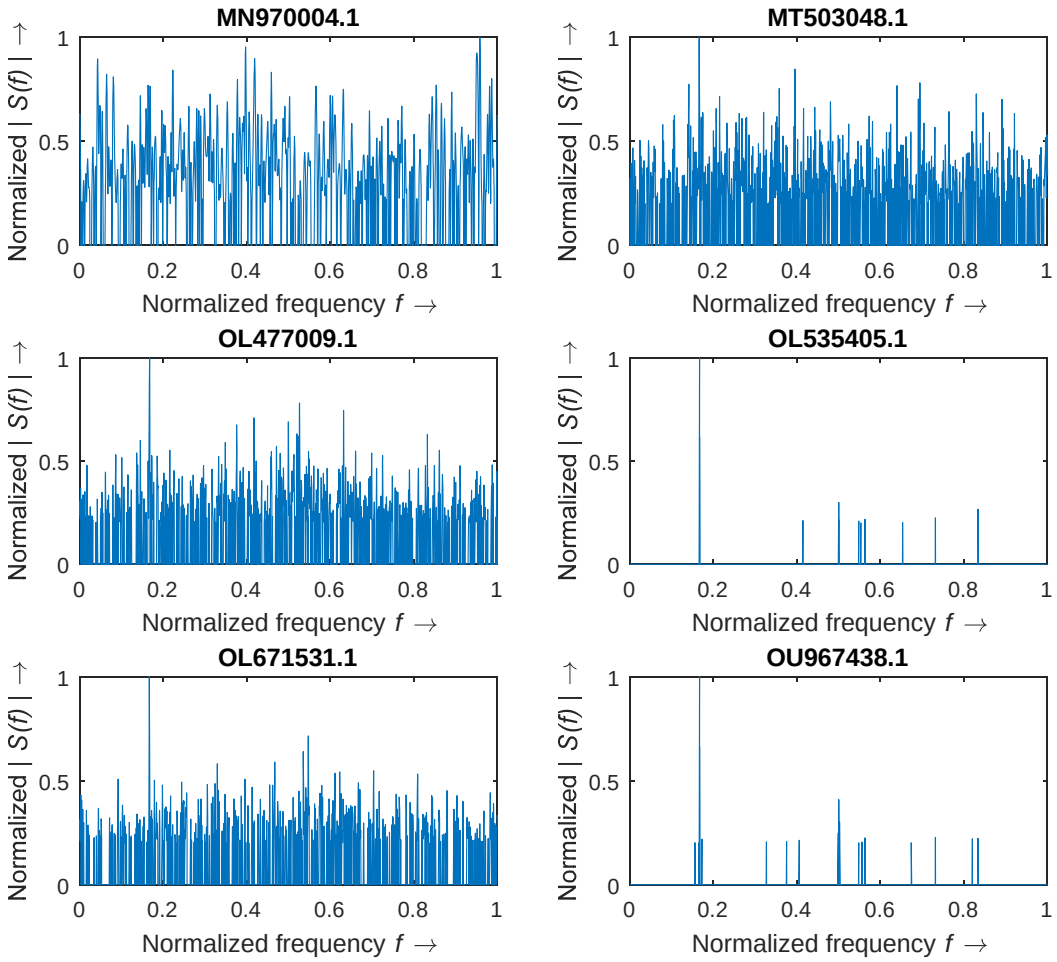

Figure 6. Dominant modes in Frequency plots are shown in Fig. 2, with a threshold 0.2

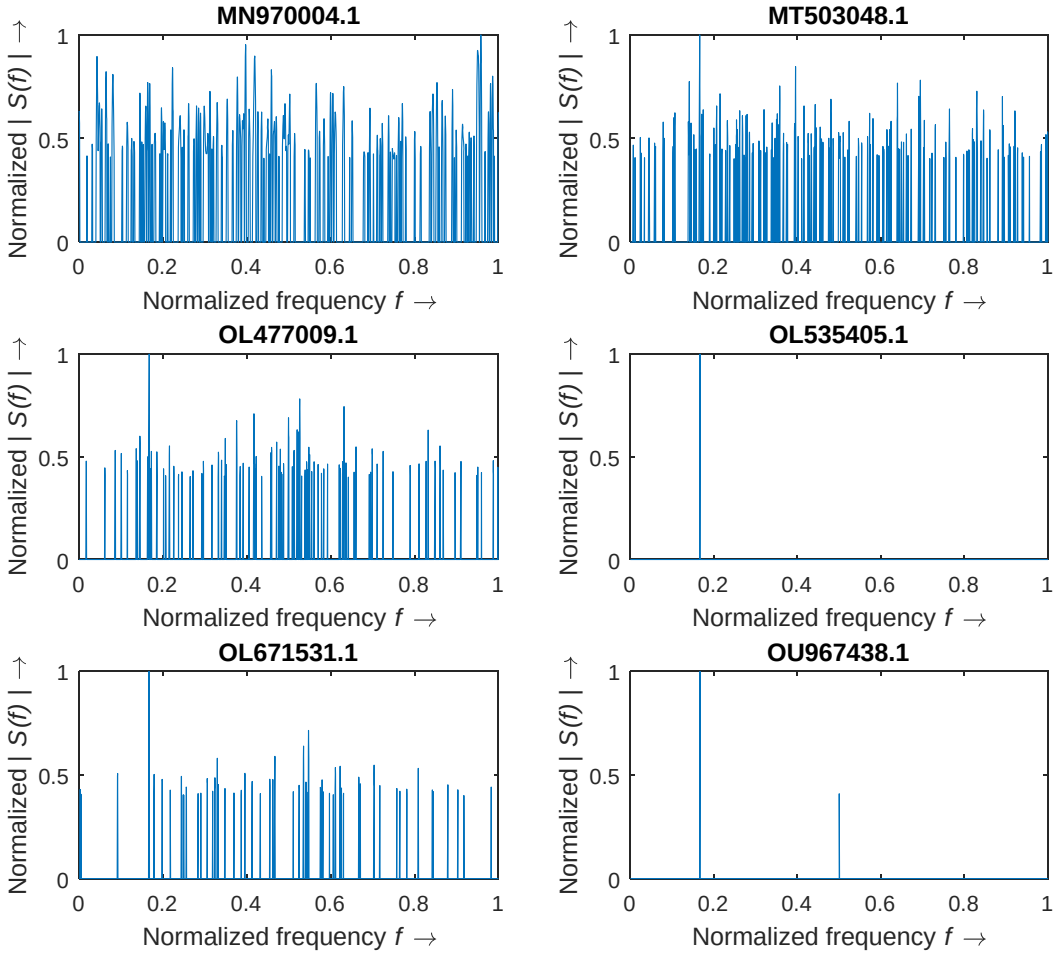

Figure 7. Dominant modes in Frequency plots are shown in Fig. 2 with a threshold 0.4.

Conclusions

We have constructed a computational model for DNA based on complex mapping of nucleotides. It is also demonstrated how this computational model is helpful for frequency representation of genome using digital signal processing-based analysis.

The genome of Omicron is found to have a highly dominant frequency band with a central frequency of 0.1667 (normalized) and bandwidth of 0.0002.It is observed that corona genome has more dominant modes in number than Omicron has, and both the genomes have some modes in common.

References:

[1] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, “Autoregressive Modeling and Feature Analysis of DNA Sequences,” *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 13–28, Jan. 2004, doi: 10.1155/S111086570430925X.

[2] R. F. Voss, “Evolution of long-range fractal correlations and 1/f noise in DNA base sequences,” *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992, doi: 10.1103/PHYSREVLETT.68.3805.

[3] D. A.- Bioinformatics and undefined 2000, “Frequency-domain analysis of biomolecular sequences,” *academic.oup.com*, vol. 16, no. 12, pp. 1073–1081, 2000, Accessed: Jun. 22, 2022. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/16/12/1073/214538

[4] D. Sussillo, A. Kundaje, and D. Anastassiou, “Spectrogram Analysis of Genomes,” *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 29–42, Jan. 2004, doi: 10.1155/S1110865704310048.

[5] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, “Prediction of probable genes by Fourier analysis of genomic sequences,” *academic.oup.com*, vol. 13, no. 3, pp. 263–270, 1997, Accessed: Jun. 22, 2022. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/13/3/263/423173

[6] D. A.-I. signal processing magazine and undefined 2001, “Genomic signal processing,” *ieeexplore.ieee.org*, Accessed: Jun. 23, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/939833/

[7] H. K. Kwan and S. B. Arniker, “Numerical representation of DNA sequences,” *Proceedings of 2009 IEEE International Conference on Electro/Information Technology, EIT 2009*, pp. 307–310, 2009, doi: 10.1109/EIT.2009.5189632.

[8] N. Rao, S. S. -, C. and S. (IEEE Cat. No, and undefined 2004, “Detection of 3-periodicity for small genomic sequences based on AR technique,” *ieeexplore.ieee.org*, Accessed: Jun. 22, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1346354/?casa_token=E85zXq6K2_8AAAAA:drUjuwr1y4fxfpw5w6bCmFZ2fs18PUILQFHDAgZuVAUdu9echOzeP67xdsPMX1juZdyOC_2joDU

[9] M. Roy and S. Barman, “Effective gene prediction by high resolution frequency estimator based on least-norm solution technique,” *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2014, no. 1, 2014, doi: 10.1186/1687-4153-2014-2.

[10] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, “Using signal processing techniques for DNA sequence comparison,” *Bioengineering, Proceedings of the Northeast Conference*, pp. 173–174, 1989, doi: 10.1109/NEBC.1989.36756.