Wavelet-based multifractal analysis and fractal Signature of SARS-Cov-2 Coronavirus Variants genomes.

Sid-Ali Ouadfeul (SOuadfeul@ymail.com)  
Algerian Petroleum Institute  https://orcid.org/0000-0002-9501-6819

Keywords: SARS-CoV-2, Variants, Long-Range correlation, Hurst exponent, Indicator matrix, Fractal dimension

DOI: https://doi.org/10.21203/rs.3.rs-788140/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Wavelet-based multifractal analysis and fractal Signature SARS-CoV-2 Coronavirus Variants genomes.

Sid-Ali OUADFEUL

Algerian Petroleum Institute, SONATRACH, Algeria

E-mail: SOuadfeul@ymail.com

Summary:
In this paper, the SARS-CoV-2 coronavirus variants of concern and of interest genomes are analyzed using the wavelet transform modulus maxima lines (WTMM) method. The goal is to track the monofractal behavior of the virus genomes and to investigate the Long-Range-Correlation (LRC) character through the estimation of the Hurst exponent. The obtained results demonstrate the multifractal and the anti-correlated characters in the variants of concern for the Knucleotidic and GC DNA coding. The fractal signatures of SARS-CoV-2 coronavirus variants are investigated through the indicator matrix maps of the genomes, they exhibit the same patterns for the variants (Alpha, Delta) and (Eta, Lota, Kappa) with moving positions, while the variants Beta, Gamma and Epsilon have different indicator matrixes. The fractal dimensions of SARS-CoV-2 variants are oscillating around 1.62, except the Epsilon variant from USA, where the fractal dimension is 1.70.

Keywords: SARS-CoV-2, Variants, Long-Range correlation, Hurst exponent, Indicator matrix, Fractal dimension.

1-Introduction
Severe Acute Respiratory Syndrome Coronavirus SARS-CoV-2 continues to propagate through the world, the first case is observed in Wuhan city located in China. One of the challenges in medicine and biology is the instability of the virus; Ouadfeul (2020a) demonstrated that SARS-CoV-2 is varying form a country to another or in the same country. Mandal et al (2020) studied and analyzed a large number of publicly available SARS-CoV-2 genomes across the world using the multifractal approach. The mutation events in the isolates obey the Markov process and exhibit very high mutational rates. Fakhredini et al (2021) demonstrated that most coronavirus mutations occur intermittently in Iran and other countries and have little effect on the pathogen
city of the virus but have increased its rate of transmission. In mutated viruses, deletion of nucleotide sequences has been observed relatively in some reading frames extensively. Studies have shown that the host protein induced mutagenesis through interaction via viral proteins. The most important mutation in SARS-CoV2 compared to the original Wuhan virus was the spike D614G mutation and the lineage of B.1.1.7, 20I/501Y.V1 become the dominant and exhibit greater virus spread but did not associate with higher viral loads and morbidity. However, it may affect the effectiveness of the vaccines and mortality rate.

In this research, eight variants of SARS-CoV-2 coronavirus are analyzed using the WTMM method, the indicator matrix, and the fractal dimension of these variants are calculated. We begin the paper by describing the Wavelet Transform Modulus Maxima Lines and the indicator matrix methods.

2-The Wavelet Transform Modulus Maxima Lines

The wavelet transform modulus lines (WTMM) is a multifractal formalism revisited by the Continuous Wavelet Transform, it was developed by Mallat and Hwang in 1992 and used for image processing. The Continuous Wavelet Transform (CWT) is a decomposition of a given signal $S(t)$ into a dilated and translated wavelets $\psi^a\left(\frac{t-b}{a}\right)$ obtained from a mother $\psi(t)$ that must have $n$ vanishing moments (Grossmann and Morlet, 1985):

$$\int_{-\infty}^{+\infty} t^n \psi(t) dt = 0 \quad n < +\infty$$

$$CWT(a, b) = \int_{-\infty}^{+\infty} S(t) \psi^a\left(\frac{t-b}{a}\right) dt$$

Where $a \in R^{++}, b \in R$

The first step of WTMM method is to calculate the continuous wavelet transform (CWT) of a given signal and the modulus of CWT, the next step is maxima of the continuous wavelet transform computation. Determination of local maxima is performed using the computation of the first and second derivative of the wavelet coefficients. CWT $(a, b)$ admits a maximum at point $b_0$ if it satisfies the following two conditions:

$$\frac{\partial \text{CWT}}{\partial b} = 0 \quad \text{if} \quad b = b_0$$
\[
\frac{\partial^2 \text{CWT}}{\partial b^2} < 0 \quad \text{if } b = b_0
\]

The function of partition \(Z(q, a)\) is a summation of the modulus of the CWT at local maxima \(b_i\) with a power \(q \in R\):

\[
Z(q, a) = \sum_{b_i} |\text{CWT}(a, b)|^q
\]

The spectrum of exponents \(\tau(q)\) is related to the function of partition \(Z(q, a)\) by:

\[
Z(q, a) = a^{\tau(q)}
\]

The spectrum of exponents \(\tau(q)\) is obtained by a simple linear regression of \(\log(Z(q, a))\) versus \(\log(q)\)

The spectrum of singularities is the Legendre transform of the spectrum of singularities:

\[
D(h) = \min_q (qh - \tau(q))
\]

For fractional Brownian motions signals which have monofractal character the spectrum of exponents has a linear behavior with equation (Arneodo et al, 1998):

\[
\tau(q) = qH - 1
\]

So the Hurst exponent is obtained by a simple linear regression of \(\log(\tau(q))\) versus \(\log(q)\).

The WTMM has the ability to decide if a given signal is monofractal or multifractal (Arneodo et al, 1998), if the spectrum of exponents has a linear behavior versus \(q\) the signal is monofractal, otherwise it is multifractal.

### 3-The indicator matrix and the fractal dimension:

The DNA of each organism of a given species is a long sequence of a specific large number of base pairs bp. Each base pair is defined on the 4 elements alphabet of nucleotides (Cattani, 2010):

A: adenine , T: thymine, C: cytosine, G: guanine

\[A = \{A, T, C, G\}\]

A DNA sequence is the finite symbolic sequence, defined by:

\[S = N^* A\]

\(S\) is defined as:

\[S = \{x_h\}_{h=1,2,3,A\ldots N},\]
\[ x_h = (h, x) = x(h) \quad (h = 1, 2, ..., N) \quad x \in A \]

\( x_h \) is the value of \( x \) at the position \( h \)

The 2D indicator function, based on the 1D-definition is the map:

\[
\mathbf{u}(x_h, x_k) = \begin{cases} 
1 & \text{if } x_h = x_k \\
0 & \text{if } x_h \neq x_k 
\end{cases}
\]

The indicator matrix \( A \) is defined as:

\[ A_{h,k} = \mathbf{u}(x_h, x_k) \]

\( A \) is a square matrix with dimension \( N \times N \)

Table 1 shows an example of construction of the indicator matrix.

| \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] | \[ \bullet \] |
|---|---|---|---|---|---|---|---|
| \[ A \] | 0 | 0 | 1 | 0 | 0 | 0 | 1 | . | . |
| \[ T \] | 1 | 0 | 0 | 0 | 1 | 0 | . | . |
| \[ G \] | 0 | 0 | 1 | 1 | 0 | 0 | . | . |
| \[ G \] | 0 | 0 | 0 | 1 | 1 | 0 | 0 | . | . |
| \[ A \] | 0 | 0 | 1 | 0 | 0 | 1 | . | . |
| \[ C \] | 0 | 1 | 0 | 0 | 0 | 0 | . | . |
| \[ T \] | 1 | 0 | 0 | 0 | 1 | 0 | . | . |
| \[ u_{h,k} \] | T | C | A | G | G | T | A | . | . |

Table 1: The indicator matrix components

From the indicator matrix we can have an idea of the “fractal-like” distribution of nucleotides.

The fractal dimension for the graphical representation of the indicator matrix plots can be computed as the average of the number \( p(n) \) of "1" in the randomly taken \( n \times n \) minors of the \( N \times N \) correlation matrix \( u_{h,k} \) (Cattani, 2010; Ouadfeul, 2020):

\[
D = \frac{1}{N} \sum_{n=2}^{N} \frac{\log(p(n))}{\log(n)}
\]
4-SARS-CoV-2 Variants:

Following the Health World Organization (HWO), SARS-CoV-2 variants have two classes: Variants of Concerns (VOCs) and Variants of Interests (VOIs).

4.1-Variants of Concern (VOC):

A SARS-CoV-2 variant that has been demonstrated to be associated with one or more of the following changes at a degree of global public health significance:
- Increase in transmissibility or detrimental change in COVID-19 epidemiology; OR
- Increase in virulence or change in clinical disease presentation; OR
- Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics. Table 02 shows the HWO classification of VOC with Pango lineage, GSAID and Nexstrain clades.

| WHO label | Pango lineages | GISAID clade | Nextstrain clade | Additional amino acid changes monitored | Earliest documented samples | Date of designation |
|-----------|----------------|--------------|-----------------|----------------------------------------|----------------------------|--------------------|
| Alpha     | B.1.1.7        | GRY          | 20I (V1)        | +S:484K +S:452R                        | United Kingdom, Sep-2020   | 18-Dec-2020        |
| Beta      | B.1.351 B.1.351.2 B.1.351.3 | GH/501Y.V2 | 20H (V2)        | +S:L18F                               | South Africa, May-2020    | 18-Dec-2020        |
| Gamma     | P.1, P.1.1, P.1.2 | OR/501Y.V3  | 20J (V3)        | +S:S81H                                | Brazil, Nov-2020          | 11-Jan-2021        |
| Delta     | B.1.617.2 AY.1 AY.2 | G/478K.V1   | 21A             | +S:417N                                | India, Oct-2020           | VOI: 4-Apr-2021 VOC: 11-May-2021 |

*Notable spike (S) amino acid changes under monitoring, which are currently reported in a minority of sequenced samples.

Table 02: Health World Organization Nomenclature of SARS-CoV-2 Variants of Concern (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/)
4.2- Variants of Interest (VOI)
It is SARS-CoV-2 variant with:
- With genetic changes that are predicted or known to affect virus characteristics such as transmissibility, disease severity, immune escape, diagnostic or therapeutic escape; AND
- Identified to cause significant community transmission or multiple COVID-19 clusters, in multiple countries with increasing relative prevalence alongside increasing number of cases over time, or other apparent epidemiological impacts to suggest an emerging risk to global public health. For more information about SARS-CoV-2 Variants, we invite readers to visit the SIB Swiss Institute of Bioinformatics web site https://viralzone.expasy.org/9556.

5-Wavelet-based multifractal analysis of SARS-CoV-2 variants:
A sample of DNA sequence of each variant of the SARS-CoV-2 virus is analyzed using the WTMM method, the goal is check the monofractal behavior and to investigate the LRC character.
Table 03 shows the GenBank of each variant (GenBanks are downloaded from the NCBI database), the country where the sample is taken and the estimated Hurst exponent of each DNA coding method. For more information about the DNA coding methods, see the paper of Messaoudi et al (2012); MF denotes the multifractal behavior. We observe that the LRC is dominant in purine and pyrimidine DNA coding of the eight SARS-CoV-2 variants, the anti-correlated and multifractal behaviors are observed in the variants of concern for the Knucleotidic DNA coding. This character is also observed in Ameno coding of the Gamma and Delta variants and in the GC coding of the Gamma variant. These results are compared with those of genomes analysis using the WTMM method shown in table 04 illustrated in the paper of Ouadfeul (2020a).

We can see that the virus genomes lose the long range correlation character and their structures become more complex and less self-organized with time.

6-Fractal Signature of some SARS-CoV-2 variants:
The same eight sequences of Alpha, Beta, Gamma, Delta, Epsilon, Eta, Lota and Kappa variants are analyzed using the indicator matrix and fractal dimension. Figure 01 shows the indicator matrix of the GenBanks: MW598419.1, MZ169911.1, MZ359841.1, MW453103.1,
MW941945.1, MW943279.1 and MW943960.1. They exhibit the same patterns for the variants (Eta, Lota and Kappa) and (Alpha, Delta) with moving positions, while the variants Beta, Gamma and Epsilon have different indicator matrixes. Table 05 shows the fractal dimensions of SARS-CoV-2 variants, they are oscillating around 1.62, except the Epsilon variant from USA, where the fractal dimension is 1.70. This variant is observed first time in USA, it is characterized by a lower transmissibility of 20% and a decreased neutralization by convalescent plasma and vaccines. Alpha variant is characterized by an increased transmissibility of 29%, a potential increase in risk of hospitalization (40-64%) and risk of death (30-70%) and no adverse effects on vaccine efficacy. Beta variant is characterized by an increased transmissibility of 25% a potential increase in risk of death during hospitalization (20%) and an impacts immune escape after infection and after vaccination. Gamma variant is characterized by an increased transmissibility of 38%, no impact reported on disease severity and impacts immune escape after infection and after vaccination. Delta variant is characterized by an increased transmissibility of 97% (SIB Swiss Institute of Bioinformatics, 2021).

7- Conclusions:
SARS-CoV-2 coronavirus has many variants circulating in the world; four of them are of concerns, which are: Alpha, Beta, Gamma and Delta. In this paper, the monofractal character and the Long-Range correlation of the virus genomes are investigated through the Hurst exponent estimation using the wavelet transform modulus maxima lines method. The obtained results show the anti-correlated and the multi fractal behavior for the Knucleotidic and the GC coding of the variants of concern, they demonstrate that the virus genomes structure become more complex and less self-organized compared to the earlier variant.

The indicator matrix and the fractal dimension are used to analyze these variants, the obtained results demonstrate that the variants (Eta, Lota and Kappa) and (Alpha, Delta) exhibit the same patterns with moving positions, while the variants Beta, Gamma and Epsilon have different indicator matrixes. The fractal dimensions of SARS-CoV-2 variants are oscillating around 1.62, except the Epsilon variant form USA, where the fractal dimension is 1.70.

Prediction of SARS-CoV-2 genome structure, organization and mutation is a challenge for the scientific community; with the instability, the non-equilibrium and the less self-organization of the new variants of the virus.
Table 03: Estimated Hurst exponent using the WTMM method of SARS-CoV-2 Variants

| Variant | GenBank    | Location | Knucleotidic | Purine A=G=1 | Pyrimidine C=T=1 | Ameno A=C=1 | Keto G=T=1 | GC coding G=C=1 |
|---------|------------|----------|--------------|---------------|------------------|-------------|-------------|-----------------|
| Alpha   | MZ344997   | England  | 0.47         | 0.64          | 0.60             | 0.57        | 0.60        | 0.60            |
| Beta    | MW598419.1 | Ghana    | MF           | 0.61          | 0.60             | 0.56        | 0.60        | 0.68            |
| Gamma   | MZ169911.1 | Brazil   | 0.47         | 0.64          | 0.63             | MF          | 0.58        | MF              |
| Delta   | MZ359841.1 | India    | 0.46         | 0.62          | 0.59             | MF          | 0.61        | 0.67            |
| Epsilon | MW453103.1 | USA      | MF           | 0.59          | 0.68             | 0.68        | 0.51        | 0.66            |
| Eta     | MW941945.1 | USA      | 0.59         | 0.64          | 0.68             | 0.69        | 0.58        |                 |
| Lota    | MW943279.1 | USA      | 0.59         | 0.64          | 0.68             | 0.69        | 0.58        |                 |
| Kappa   | MW943960.1 | USA      | 0.59         | 0.63          | 0.68             | 0.69        | 0.56        |                 |

MF: denotes the mutifractal behavior
| GenBank     | Variant | Location | Fractal Dimension |
|------------|---------|----------|------------------|
| MZ344997   | Alpha   | England  | 1.61             |
| MW598419.1 | Beta    | Ghana    | 1.64             |
| MZ169911.1 | Gamma   | Brazil   | 1.64             |
| MW453103.1 | Epsilon | USA      | 1.70             |
| MZ359841.1 | Delta   | India    | 1.64             |
| MW941945.1 | Eta     | USA      | 1.63             |
| MW943279.1 | Lota    | USA      | 1.63             |
| MW943960.1 | Kappa   | USA      | 1.63             |

Table 05: GenBanks, their variant, location and the fractal dimension
MW941945.1 SARS-CoV-2 Eta Variant

MW943279.1 SARS-CoV-2 Lota Variant
Figure 01 Indicator Matrix of Alpha, Beta, Gamma, Delta, Epsilon, Eta, Lota and Kappa SARS-CoV-2 Variants

8-References:

Arneodo, A., Bacry, E., & Muzy, J.-F., 1995, The thermodynamics of fractals revisited with wavelets. Physica A 213, 232-275.

Cattani, C., 2010, Fractals and Hidden Symmetries in DNA, Hindawi Publishing Corporation Mathematical Problems in Engineering, Volume 2010, Article ID 507056, 31 pages, doi:10.1155/2010/507056.

Grossmann, A., and Morlet, J.-F., 1985, Decomposition of functions into wavelets of constant shape, and related transforms. In: “Mathematics and Physics, Lecture on Recent Results” L. Streit, Editor. World Scientific Publishing (Singapore) 1985.

Fakhredini, K., Soleimanjahi, H., Banijamali, R S., Bamdad, T., 2021, Evaluation of Mutation Spread in The SARS-CoV2 Genome. mjms. 2021; 23 (1) :41-49.

Mallat, S.; Hwang, W.L., 1992, Singularity detection and processing with wavelets, IEEE Transactions on Information Theory, volume 38, number 2, pages 617–643, Mar 1992 doi:10.1109/18.119727

Messaoudi, I., Elloumi, A., and Lachiri, Z., 2012, Genomic Data Visualisation, SETIT 2012, Sousse, Tunisia.
National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/genbank/

Ouadfeul, S.A., 2020a, Multifractal Analysis of SARS-CoV-2 Coronavirus genomes using the wavelet transform, doi: https://doi.org/10.1101/2020.08.15.252411

Ouadfeul, S.A., 2020b, Fractal signatures of SARS-CoV2 coronavirus, the indicator matrix, the fractal dimension and the 2D directional wavelet transform: A comparative study with SARS-CoV, MERS-CoV and SARS-like coronavirus, doi: https://doi.org/10.1101/2020.08.26.269118

SIB Swiss Institute of Bioinformatics, Viral Zone, 2021, Sars-CoV-2 circulating variants, https://viralzone.expasy.org/9556

Vipr Variants of Concern, https://bv-brc.org/view/VariantLineage/#view%5Ftab=overview