

Fractal signature of coronaviruses related to severe acute respiratory syndrome

Thiago de Salazar e Fernandes (✉ thiago.fernandes@ufpe.br)

Universidade Federal de Pernambuco <https://orcid.org/0000-0002-1053-1943>

Jonas Sérgio de Oliveira Filho

Universidade Federal de Pernambuco

Isvania Maria Serafim da Silva Lopes

Universidade Federal de Pernambuco

Short Report

Keywords: Fractal analysis, random walking, deterministic chaos, Coronavirus, COVID-19

Posted Date: April 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-23660/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Coronavirus is the agent which causes the severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), emergent infections such as the present Pandemic of COVID-19. Understanding its genome pattern is important for developing new and faster ways of testing for identifying the genome of the virus, and also for better understanding of its origin and evolution. The aim of this work was to investigate the genome of SARS-CoV, MERS-CoV and SARS-CoV-2 by the paradigm of chaos theory and fractal geometry. For that, it was calculated the alpha coefficient by detrended fluctuation analysis (DFA) for the sequences of these genomes converted to binary numbers in order to determine if it is a chaotic or a random series of data. Also, it was applied the random walking for obtaining a fractal map of the whole genome, and calculated the fractal dimension (FD) by box-counting of this map by two different softwares. With this, it was found that the alpha coefficient of the first SARS viruses was > 0.5 , indicating that the series is chaotic or fractal, and has a persistent long-range memory or self-similarity along its sequence. This is not the case for MERS virus, which showed to have a completely random sequence ($\alpha < 0.5$). For the fractal dimension, SARS viruses presented a FD around 1.5, and for MERS the fractal dimension decreases ($FD < 1.5$). The images generated by random walking of the entire RNA genome are by itself a fractal signature of the virus, which may be applied for studying its origin and for faster diagnostic of COVID-19.

Introduction

A high incidence of flu in adults and children are caused by coronavirus. The SARS-CoV-2 virus is the cause of COVID-19, a severe acute respiratory syndrome (SARS). In 2002, a pathogenic coronavirus caused the first SARS, and ten years later, MERS-CoV caused the Middle East respiratory syndrome (MERS) (Chen J. 2020).

Previous studies have shown that SARS-CoV during the Epidemic of 2002 has passed through mutations to better bind to the human cellular receptor ACE-2 in order to replicate, increasing its virulence (Cui J, Li F, Shi Z. 2019). Also, in other studies related to the mutation, it was identified that SARS-CoV-2 is significantly clustered with the bat coronavirus isolated in 2015, increasing its capacity of human infection (Benvenuto et al., 2020).

Therefore, viruses change its genome from time to time originating emergent infections. But understanding the dynamics and pattern of these mutations and its consequences for the entire RNA genome of these viruses requires also a new paradigm in science, based not on classical methods.

The non-linear mathematical algorithms applied for studying the shape and dynamics of many biological processes is a field so-called fractal geometry and chaos theory (Mandelbrot, 1983; Liebovitch 1998; Stam 2005; Kunicki et al. 2009). Some developments in this field begins with the works of Peng et al.

(1994 and 1995), with the description of the detrended fluctuation analysis (DFA) method for analyzing long-range correlations of genes and intergenic regions, as well as in the interval of heart-beats.

For DFA algorithm, at first, the time series is integrated, and then it is divided into boxes of equal length, n . In each box of length n , a least-square line is fitted to the data (representing the trend in that box). An algorithm is performed to obtain the relationship of the average fluctuation as a function of box size n , in a double log graphic, in which a linear relationship indicates the presence of scaling exponent, or power-law (Stam 2005). The authors have found that sequences rich in genes has a random series of data ($\alpha < 0.5$), and the non coding regions presented a long-range correlation or self-similarity ($\alpha < 0.5$). The method showed to be powerful also for estimating health and disease, in which healthy hearts presented a chaotic or fractal pattern in the interval of heart beats, and pathological hearts presented a random pattern of these intervals. Therefore, self-similarity characteristic of fractal phenomena are presented from molecules such as DNA and proteins to organs such as hearts (Peng et al. 1994; Stanley et al. 1996; Havlin et al. 1999; Wang et al. 2008).

In the present work, we applied the methods based on non-linear dynamics for understanding the pattern of coronaviruses causing acute severe respiratory syndromes, in order to obtain its fractal signature for better understanding of its origin, mutations and for faster identification of its genome in the diagnostic of COVID-19.

Methods

The RNA sequences of coronaviruses SARS-CoV, MERS-CoV and SARS-CoV-2, were obtained in GenBank (www.ncbi.nlm.nih.gov) with the codes FJ588686.1, PRJNA485481 and MT192773.1, respectively.

The genetic sequences were converted to binary number, in which purines were converted to the number -1, and pyrimidines converted to the number 1. With this conversion, it was generated a series of -1 and 1. This series of data was then submitted to the Detrended fluctuation analysis (DFA). DFA was calculated as described by Peng et al. (1994; 1995), and using the free software GNU Octave (<https://www.gnu.org/software/octave/>). The series was integrated and divided into boxes of equal length (n). In each box of length n , a least-square line was fitted to the data representing the trend in that box. The y coordinate of the straight line segments was denoted by $y_n(k)$. Computing DFA was performed according to equation 1.

[Please see the supplementary files section to view the equation.] (1)

This computation was repeated over all the time scales (box sizes) to provide a relationship between the average root-mean fluctuation function $F(n)$ and the box size n (i.e. the number of length changes in a box which was the size of the window of observation). Typically, $F(n)$ would increase with box size n . A

linear relationship on a double log plot indicated the presence of scaling. The fluctuations could be characterized by a scaling exponent alpha (α), the slope of the line relating $\log F(n)$ to $\log n$: $F(n) \sim n^\alpha$, in order to provide a more accurate estimate of $F(n)$. If $\alpha = 0.5$, the series was the result of a random event; $\alpha > 0.5$ indicated the persistent long-range correlations. The other values were: $\alpha = 1$ corresponded to $1/f$ noise (very rough landscape); the $\alpha \geq 1$ indicates that correlations existed but ceased to be of a power-law form or a random walk-like fluctuation; and finally, $\alpha = 1.5$ indicates Brown noise; the integration of white noise (very smooth landscape) (Peng et al. 1995).

For random walking, the sequences of nucleotides were converted in vectors, in which A is to the left, T is to the right, C is down and G is up, using Python Language Reference, version 2.7, available at <http://www.python.org>. The images were submitted to fractal analysis using box-counting method of the plugin "fracLac" of the software ImageJTM (<https://imagej.nih.gov/ij/plugins/fraclac/fraclac.html>), and using Fractal Dimension Estimator (FDE) (<http://www.fractal-lab.org/Downloads/FDEstimator.html>). This method consists in the superposition of boxes in the image, in which the size of the boxes is continuously reduced. In each of this reductions, it is counted the number of boxes necessary to cover the image.

After this, the fractal dimension (FD) is calculated by equation 2:

[Please see the supplementary files section to view the equation.] (2)

In which N_r is the number of boxes necessary to cover the image in each progressive reduction of the size of the box (r). FD will be the slope of the regression line, generated by the logarithm of the quantity of boxes in function to the logarithm of the size of the boxes.

Results

The values of dimension and value of alpha (α) from DFA method applied to SARS-CoV, MERS-CoV and SARS-CoV-2 are presented in Table 1 below:

Table 1. Values of dimension and alpha coefficients (α) from DFA method applied to the sequences of SARS-CoV, MERS-CoV and SARS-CoV-2.

Virus	Dimension from DFA	Alpha Coefficient (α)
SARS-CoV	2.4020	0.59796
MERS-CoV	2.5061	0.49389
SARS-CoV-2	2.4261	0.57394

In Figure 1, it is shown the double-log graphics of $F(n)$ versus (n) , for all viruses, in which the inclination of the line is the alpha coefficient (α).

The result of random walking of SARS-CoV, MERS-CoV and SARS-CoV-2 sequences is shown in Figure 2 below.

The fractal dimensions (FDs) of the random walking of each virus were calculated by box-counting using two different softwares: imageJ and Fractal Dimension Estimator (FDE). The values of FD are listed in Table 2 below:

Table 2. Values of fractal dimensions (FD) obtained by softwares image J and Fractal Dimension Estimator (FDE).

Virus	FD by image J	FD by FDE
SARS-CoV	1.3079	1.53875
MERS-CoV	1.1689	1.40
SARS-CoV-2	1.2338	1.49827

In Figure 3, it is shown the double log graphic of the box-counting method (log of the number of boxes by the log of the $1/r$ length of the size of the box), in which the inclination of the curve is the fractal dimension (FD). The only software generating this graphic is FDE.

Discussion

Conventional methods concentrate efforts to investigate which mutations are involved in modifications of genomes. In this work, another paradigm was taken based on the fractal signature of the whole genome, and not in specific point mutations.

It is well-known that DNA in eukaryotes presents a fractal pattern with an ideal optimization of space in order to organize a long molecule of DNA into a coiled complex structure of just a few micrometers (Liebovitch 1998; Lieberman-Aiden et al. 2009). Also, the dynamics of gene mutations and organization is not considered to be random, presenting a long-range correlation with a DFA alpha value (α) higher than 0.5 for sequences rich in genes, and a random pattern for intergenic regions ($\alpha=0.5$) (Peng et al., 1994).

When the sequences of viruses known in causing severe acute respiratory syndromes, the SARS-CoV (2002), MERS-CoV (2012) and SARS-CoV-2 (2019) were investigated using DFA, it was noticed that the SARS-CoV and SARS-CoV-2 presented a long-range memory characteristic of a self-similar sequence, but for MERS-CoV the sequence is random ($\alpha\approx 0.5$). When the fractal dimension (FD) was estimated by two different softwares, it was evidenced also that SARS-CoV and SARS-CoV-2 are fractal, and this fractality decreases in MERS-CoV. This reinforces the result with DFA that indicates that SARS-CoV and SARS-CoV-2 have more self-similarity or long-range memory along the sequence than MERS-CoV.

The softwares applied for estimating fractal dimensions (imageJ and Fractal Dimension Estimator) showed a variation in the obtained values. Despite that, both showed the same pattern of increase and decrease of FD, with just a difference of around 0.2, what confirms the results obtained.

When the random walking of the investigated virus is analyzed alone, even with no other data of fractal dimension or detrended fluctuation analysis, it is clear that the pattern generated by it, is its own specific signature, a 'fractal signature' of the virus. This could be of great help for identifying the virus by only

observing the generated pattern, for a quicker answer. This together with faster and easier methods for DNA and RNA sequencing would be helpful for studying the origin of SARS-CoV-2, its mutations and for diagnostic of COVID-19.

Conclusion

This work presented an approach for obtaining a genetic signature for coronavirus based on detrended fluctuation analysis (DFA), random walking and fractal dimension of random walking, allowing the specific identification, whole genome analysis, what may be helpful as a faster tool for the diagnostic of COVID-19.

Declarations

Acknowledgement

The authors would like to thank the Federal University of Pernambuco.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J Med Virol.* 2020; 92: 455–459.
- Chen J. Pathogenicity and transmissibility of 2019-nCoVdA quick overview and comparison with other emerging viruses. *Microbes and Infection.* 2020; 22: 69-71.
- Cui J, Li F, Shi Z. Origin and evolution of pathogenic coronaviruses. *Nature Reviews - Microbiology.* 2019; 17: 181-192.
- Havlin S, Buldyrev SV, Bunde A, Goldberger AL, Ivanov PCh, Peng CK, et al. Scaling in nature: from DNA through heartbeats to weather. *Phys A.* 1999; 273: 46-49.
- Kunicki ACB, Oliveira AJ, Mendonça MBM, Barbosa CTF, Nogueira RA. Can the fractal dimension be applied for the early diagnosis of non-proliferative diabetic retinopathy? *Braz J Med Biol Res.* 2009; 42: 930-934.
- Lieberman-Aiden E, Berkum NL, William L, Imakaev M, Ragooczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009; 326: 289-293.

Liebovitch LS. Fractals and chaos simplified for life sciences. New York: Oxford University Press; 1998.

Mandelbrot BB. The Fractal Geometry of Nature. 2nd. ed. New York: Ed. Freeman; 1983.

Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. Phys Rev E. 1994; 49: 1685-1689.

Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in non-stationary heartbeat time series. Am Inst Phys. 1995; 5: 82-87.

Stam CJ. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. Clin Neurophysiol. 2005; 116: 2266-2301.

Stanley HE, Afanasyev V, Amaral LAN, Buldyrev SV, Goldberger AL, Havlin S, et al. Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology econophysics. Phys A. 1996; 224: 302-321.

Wang SC, Li PC, Tseng HC. Long range correlation and possible electron conduction through DNA sequences. Phys A. 2008; 387: 5159-5168.

Figures

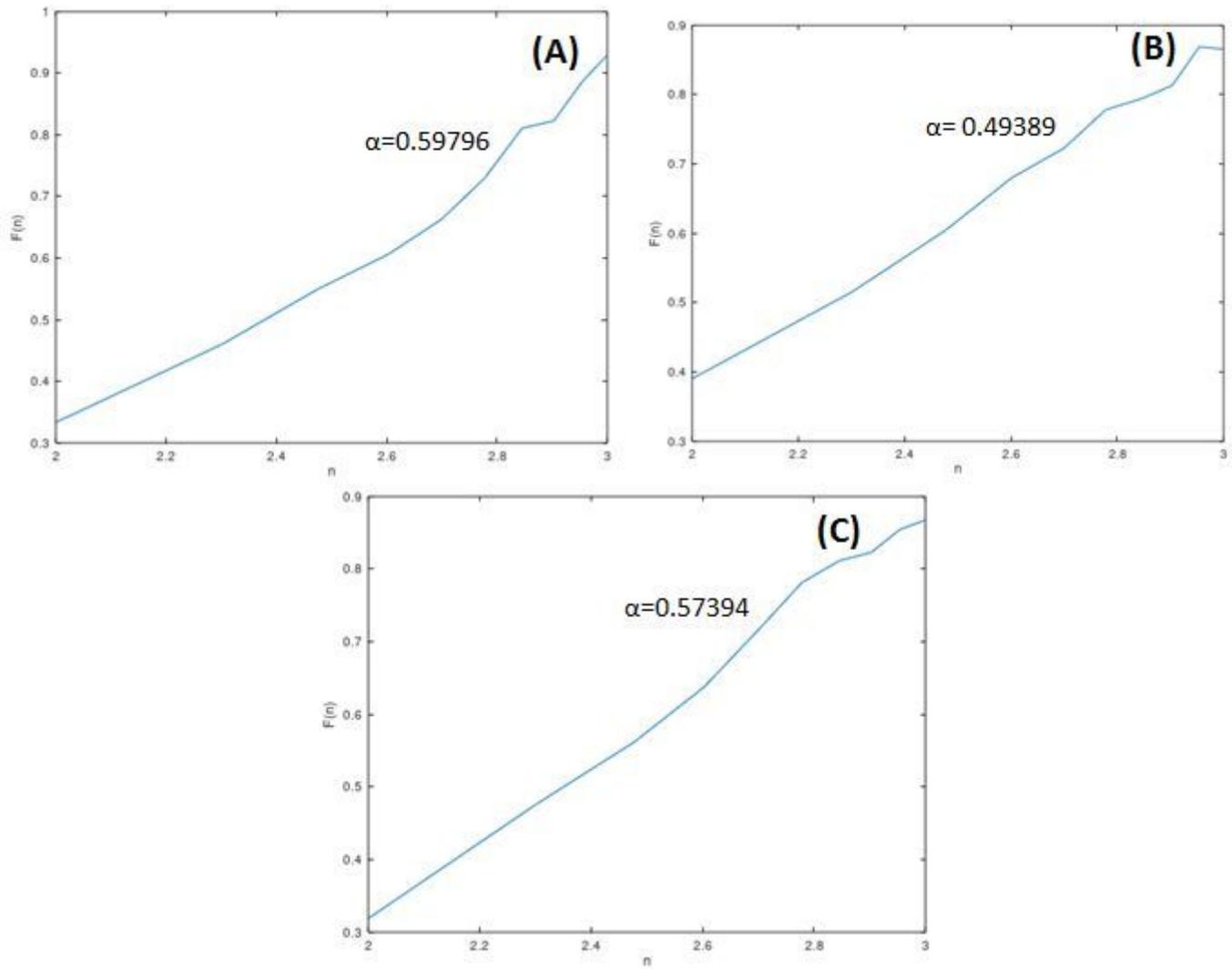


Figure 1

Double log graphic of $F(n)$ versus n . (A) is the result of SARS-CoV, (B) MERS-CoV, (C) SARS-CoV-2.

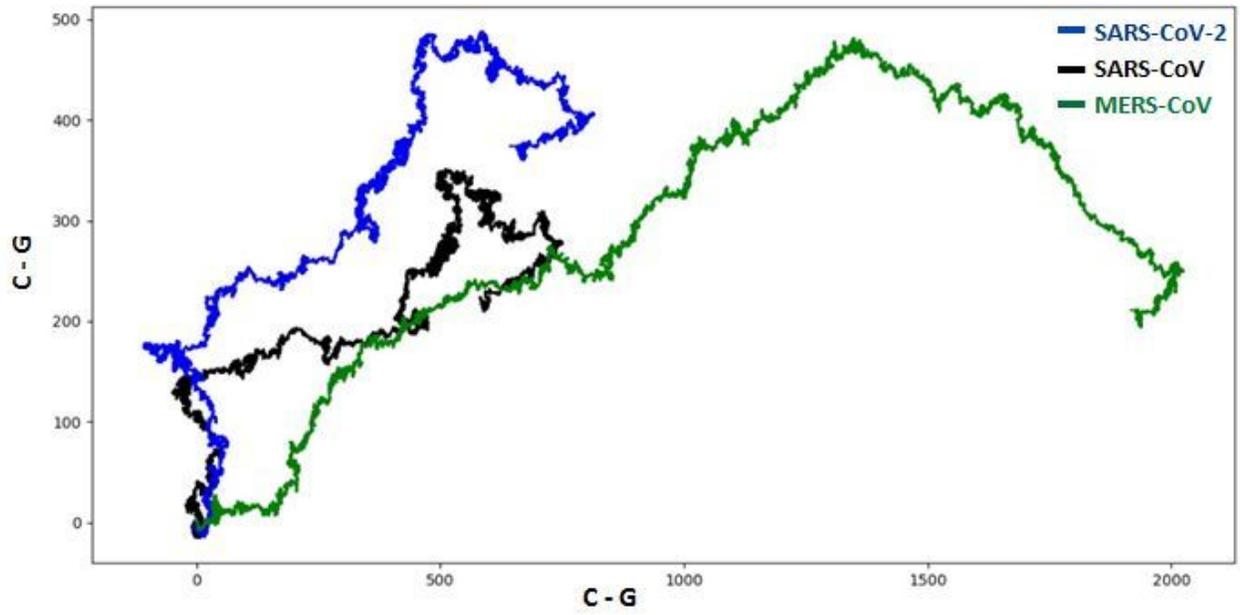


Figure 2

Random walking for the sequences of SARS-CoV, MERS-CoV and SARS-CoV-2.

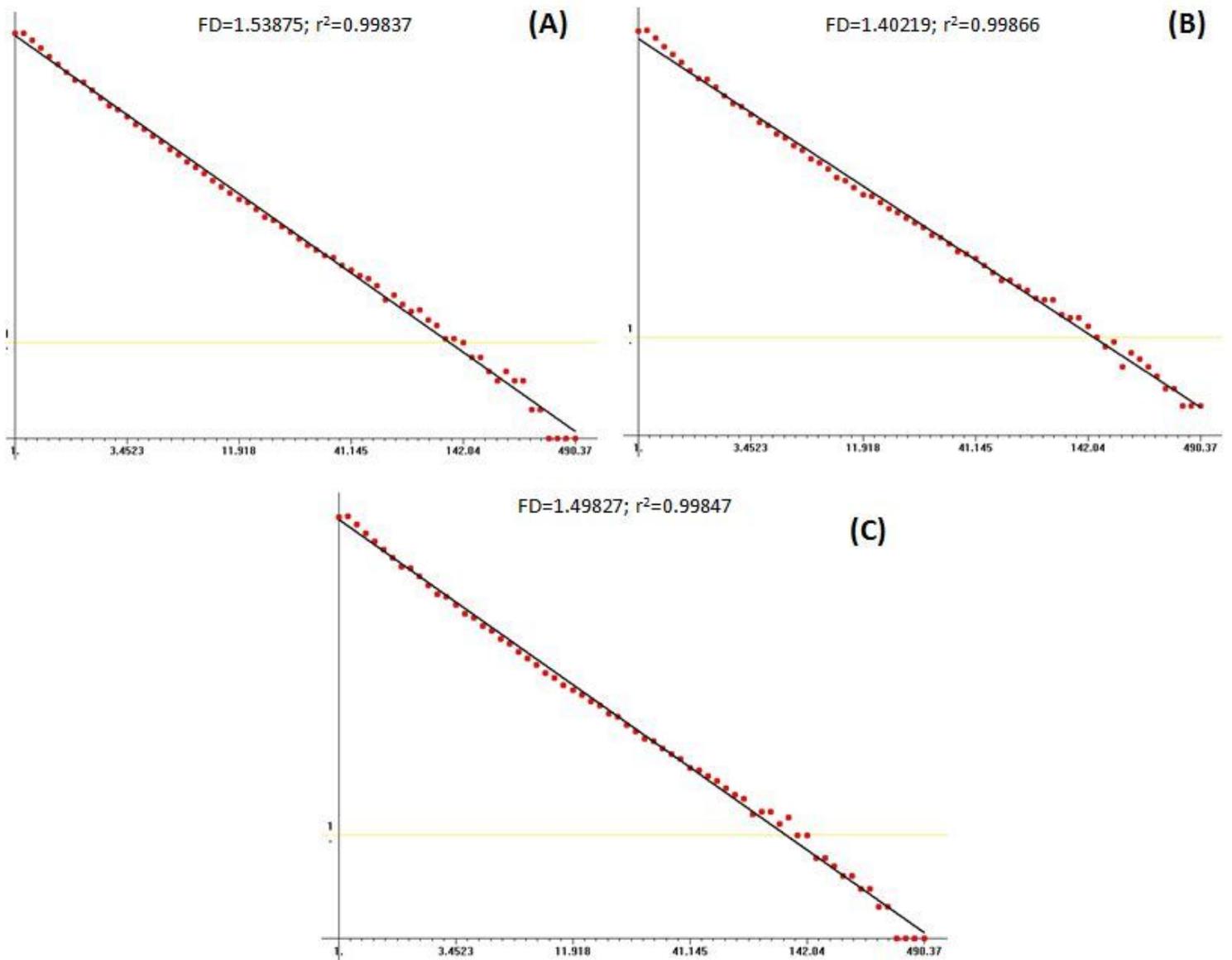


Figure 3

Double-log graphic of box-counting algorithm, for SARS-CoV (A), MERS-CoV (B) and SARS-CoV-2 (C) random walking images.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Equations.docx](#)