Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A fractal scaling analysis of the SARS-CoV-2 genome sequence

M. Meraz\textsuperscript{b}, E.J. Vernon-Carter\textsuperscript{b}, E. Rodriguez\textsuperscript{c}, J. Alvarez-Ramirez\textsuperscript{b,*}

\textsuperscript{a} Departamento de Biotecnología, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, Iztapalapa, CDMX 09340, Mexico
\textsuperscript{b} Departamento de Ingeniería de Process e Hidráulica, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, Iztapalapa, CDMX 09340, Mexico
\textsuperscript{c} Departamento de Ingeniería Eléctrica y Computación, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, Iztapalapa, CDMX 09340, Mexico

\textbf{Keywords:}
Detrended Fluctuation Analysis (DFA)
Sequence Mapping Assignation Rules
Long Range Correlations
SARS-CoV-2 Genome Sequence

\textbf{A B S T R A C T}

An approach based on fractal scaling analysis to characterize the organization of the SARS-CoV-2 genome sequence was used. The method is based on the detrended fluctuation analysis (DFA) implemented on a sliding window scheme to detect variations of long-range correlations over the genome sequence regions. The nucleotides sequence is mapped in a numerical sequence by using four different assignation rules: amino-keto, purine-pyrimidine, hydrogen-bond and hydrophobicity patterns. The originally reported sequence from Wuhan isolates (Wuhan Hu-1) was considered as a reference to contrast the structure of the 2002–2004 SARS-CoV-1 strain. Long-range correlations, quantified in terms of a scaling exponent, depended on both the mapping rule and the sequence region. Deviations from randomness were attributed to serial correlations or anti-correlations, which can be ascribed to ordered regions of the genome sequence. It was found that the Wuhan Hu-1 sequence was more random than the SARS-CoV-1 sequence, which suggests that the SARS-CoV-2 possesses a more efficient genomic structure for replication and infection. In general, the virus isolated in the early 2020 months showed slight correlation differences with the Wuhan Hu-1 sequence. However, early isolates from India and Italy presented visible differences that led to a more ordered sequence organization. It is apparent that the increased sequence order, particularly in the spike region, endowed some early variants with a more efficient mechanism to spreading, replicating and infecting. Overall, the results showed that the DFA provides a suitable framework to assess long-term correlations hidden in the internal organization of the SARS-CoV-2 genome sequence.

\section{1. Introduction}

Several cases of a new severe acute respiratory syndrome (SARS) were detected in Wuhan, China in December 2019. The syndrome showed several similarities with the SARS outbreak also surged in China in 2002, which was caused by a coronavirus. The new outbreak was also caused by a coronavirus that was named SARS-CoV-2 or Covid-19 in short. The new strain propagated outside China and by March 2020 the virus was detected in at least 100 countries with about 100,000 confirmed cases. The infection has persisted with an increasing number of infected cases and deaths and is considered a pandemic threat for the economy, social stability and human health.

The danger of the syndrome was promptly recognized, and so was the necessity of an accurate understanding of the propagation, infection and replication mechanisms and dynamics. Swiftly, researchers and public health administrators were aware that a vaccine was probably the only expedient way to alleviate the social, economic and public health system problems caused by the virus. The first detailed description of the genome sequence was reported on January 20 \cite{1}, unfolding important efforts to characterize its structure and organization to develop an effective vaccine. The monitoring of the virus replication showed the surge of the first variant of concern (Alpha – B.1.1.7) first identified in the United Kingdom in November 2020. New variants of concern were detected, including the nowadays dominant variant Delta (B.1.617.2) documented in May 2021, which spreads much faster and may cause more severe cases than the other variants.

A plethora of studies have been carried out to characterize the structure and mutations of the SARS-CoV-2 genome sequence composed of about 29800 bp. Most of these studies focus on the description of the functional regions and their modifications caused by different mutations aimed to classify the variants into lineages \cite{2}. Mutations, spike structure and immune escape mechanisms have been of particular interest to understand the infection mechanisms and vaccine effectiveness \cite{3}. Commonly, the analysis of the SARS-CoV-2 genome relies on comparing the nucleotide arrangement of two or more sequences to detect modifications linked to mutations of interest. More involved methods have...

\* Corresponding author.
E-mail address: jjar@xanum.uam.mx (J. Alvarez-Ramirez).

https://doi.org/10.1016/j.bspc.2021.103433
Received 18 September 2021; Received in revised form 4 November 2021; Accepted 28 November 2021
Available online 6 December 2021
1746-8094/© 2021 Elsevier Ltd. All rights reserved.
also been used to gain insights into the complexity of the sequence. In particular, methods based on information theory are expected to provide valuable insights into the evolution of genome functionality. Ghanchi et al. [4] found that the earlier genome sequences isolated before June 2020 exhibited a higher mean and site-specific entropy compared with that of posterior isolates. This suggests that the viral genome achieved stability after the initial early period of the pandemic. Nawaz et al. [5] postulated that alignment-free methods can provide valuable insights into the structure and functionality of the SARS-CoV-2 genome sequence. Namazi [6] used sample and approximate entropy methods to conclude that SARS-CoV-2 has a significantly different complex structure than SARS-CoV-1. Fractal dimension methods and scaling analysis based on detrended fluctuation analysis (DFA) were used to find that the genome sequence possesses persistent long-range memory [7].

The reduced number of studies on the informational structure (e.g., long-range correlations) of the SARS-CoV-2 genome sequence has a limited impact on the characterization of the mechanisms and mutation effects in the sequence structure. Results in this line should provide valuable insights that might complement the studies based on evolutionary genetics. While most studies have relied on comparing the locations of nucleotides in different regions, the effect of mutations (substitutions and displacements) on the intrinsic informational organization, order and complexity of the SARS-CoV-2 genome sequence has not been explored. The reduced number of studies only provided an average characterization of the genome structure [4,5,7], without focusing on the local structure. Mutations are located in small regions, such that the modifications of the genome structure are also expressed in local terms. In this regard, the present study aims to explore the local organization of the SARS-CoV-2 genome sequence using long-term correlation analysis. The approach is based on a sliding window implementation of the detrended fluctuation analysis (DFA) to characterize the local fluctuations of the fractal scaling properties [8]. The main results showed that the SARS-CoV-2 genome sequence is more structured (i.e., less random) than its SARS-CoV-1 counterpart, a feature that would endow the virus with improved capacities for replication and infection. Mutations were linked to local variations of the scaling properties, particularly in the surface glycoprotein (spike) region. Overall, the results showed that the DFA is a suitable framework to characterize the internal organization of the genome sequence in terms of long-range correlations.

2. Detrended fluctuation analysis

For a given numerical mapping rule of the genome sequence, the aim is to estimate local autocorrelations that can be of interest in terms of mutations reported in the literature. A diverse set of methods has been proposed in recent decades, including spectral analysis and rescaled range (R/S) analysis. However, real sequences can be affected by non-stationarities and long-term trends, which might induce biased estimates of fractal scaling properties. The so-called detrended fluctuation analysis (DFA) was proposed in statistical physics to overcome the drawbacks of traditional methods. The method is an improvement of the classical fluctuation analysis by removing long-term trends [9]. The DFA allows the detection of intrinsic self-similarity embedded in a seemingly nonstationary time series, and also avoids the spurious detection of apparent self-similarity, which may be an artifact of extrinsic trends. These features have made the DFA the most widely used approach for the fractal analysis of complex time series in a large diversity of fields, from geophysics [10] and financial systems [11] to biology [12], physiology [13] and chemistry [14,15], and provides an easy interpretation of fluctuation patterns in terms of scaling exponents.

For the sake of completeness, the DFA can be briefly described as follows. Given a bounded sequence \( y_k, \ k = 1, \ldots, N \), the summation first converts the sequence in an unbounded sequence

\[
y_k = \sum_{j=1}^{N} y_j
\]

where \( \langle x \rangle \) is the mean value of the sequence, and \( y_k \) corresponds to the profile. Next, the sequence \( y_k \) is divided into subsequences of length \( n \) samples each. The size \( n \) of the subsequences represents the scale at which the fluctuations of the unbounded sequence \( y_k \) are characterized. A local least-squares straight-line fit is computed for each subsequence. Let \( y_{x,n} \) be the resulting piecewise sequence of straight-line fits, which represents the long-term trend of the profile sequence for the scale \( n \). Then, the fluctuations are quantified as the root-mean-square deviation from the trend:

\[
F(n) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (y_j - y_{x,n}(n))^2}
\]

In this way, the fluctuation function \( F(n) \) is dependent on the scale \( n \). For small values of \( n \), the least-squares fitting \( y_{x,n} \) provides a better description of the fluctuations \( y_k \). In contrast, for large values of \( n \), a coarser fitting of the fluctuations is obtained and the difference \( y_k - y_{x,n}(n) \) is higher than for small values of the scale \( n \). The procedure of detrending and fluctuation quantification is repeated over a range of different subsequence lengths \( n \). If the process \( x_k \) is self-affine, the fluctuation function \( F(n) \) can be described as a power-law function of the form

\[
F(n) = Kn^a
\]

The exponent \( a \) is a generalization of the Hurst exponent and reflects the scaling behavior of the detrended fluctuations. For sufficiently long sequences (i.e., \( N \) very large), one has that \( a = 0.5 \) for uncorrelated random walks (i.e., the walk displacement grows like \( \sqrt{n} \)). Deviations from \( a = 0.5 \) indicate the presence of self-correlations. In particular, \( a \in (0.0, 0.5) \) reflects anti-persistent behavior (i.e., a positive move is likely followed by a negative move and vice versa), whereas \( a \in (0.5, 1.0) \) corresponds to persistent behavior (i.e., a positive/negative move is likely followed by a positive/negative move). The usual range of scales for the computation of the DFA scaling exponent is \( [5, N/5] \) [9], and the scaling exponent is estimated as the slope of a log-log plot of the fluctuation function \( F(n) \) versus the scale \( n \).

2.1. Randomness test

The genome sequence analysis involves the detection of regions containing serial correlations. In this way, one should test whether the DFA predicts deviations from randomness. In principle, serial correlations are present when the scaling exponent \( a \) is not equal to 0.5. However, the value \( a = 0.5 \) indicating the absence of serial correlations, holds for asymptotic statistics of the fluctuation function \( F(n) \) (i.e., for very long values of the scale \( n \)) [16]. This is a drawback since commonly one should deal with sequences of relatively small length. If the probability distribution \( P(X) \) that generated the values of the sequence \( x_k \) is available, an approach is to generate many random sequences of finite size \( N \) and to compute the statistics of the DFA scaling exponent to obtain the confidence intervals (CI). However, the exact distribution is hardly available in practice for a given process. Bootstrapping estimates can be used by considering an approximate (i.e., empirical) distribution. The iso-distributional surrogate data test [17] was used to test randomness. Briefly, the following procedure is proposed to estimate the CI for randomness: a) Compute \( X_k \) shuffled sequences \( x_{sh,k} \) from the original sequence \( x_k \). Shuffling destroys serial correlations while retaining the statistical distribution of values. That is, the sequences \( x_{sh,k} \) and \( x_k \) were generated from a common distribution \( P(X) \). b) Compute the DFA scaling exponent for the shuffled sequences \( x_{sh,k} \) which reflects the scaling behavior of a random sequence. c) Carry out the statistical
analysis of the $N_{sh}$ scaling exponent values to obtain the corresponding CI for randomness. In the sequel, $N_{sh} = 5000$ randomized sequences were used to estimate the confidence intervals.

3. Numerical mapping of the genome sequence

A genome is a text sequence of the nucleotides ($n_k$) adenine (A), cytosine (C), guanine (G) and tyrosine (T). To apply the DFA to a sequence of nucleotides, the text sequence should be mapped into a one-dimensional numerical sequence. In this regard, four different mappings were considered [18]:

a) Amino-Keto Rule (AK Rule). $x_k = +1$ if $n_k$ is an amino (A or C) and $x_k = -1$ if $n_k$ is a keto (G or T).

b) Purine-Pyrimidine Rule (PP Rule). $x_k = +1$ if $n_k$ is a purine (A or G) and $x_k = -1$ if $n_k$ is a pyrimidine (C or T).

c) Hydrogen Bond Energy Rule (HB Rule). $x_k = +1$ for strongly bonded pairs (C or G) and $x_k = -1$ for weakly bonded pairs (A or T).

d) Hydrophobicity Rule (HP Rule). The numerical value corresponds to the dipolar moment of the nucleotide according to the following assignment [19]: 2.46 for A, 7.04 for C, 6.75 for G and 4.56 for T.

4. SARS-CoV-2 genome sequences

The genome sequences were downloaded from the NCBI GenBank database at the publicly accessible site www.ncbi.nlm.nih.gov. The genome sequence for SARS-CoV-1 (NC_004718.3) is composed by 29,751 nucleotides [20]. The firstly reported genome sequence for SARS-CoV-2 (Wuhan Hu-1, MN908947, 29903 bp) collected from Wuhan, China [1] was taken as a reference. Genome sequences for SARS-CoV-2 from samples collected in the early months of 2020 were used to contrast the results of the Wuhan Hu-1 isolate: MT019529 (China, December 23, 2019, 29,989 nucleotides, isolated), MT012098 (India, Collection Date January 27, 2020, 29854 bp), MT039890 (South Korea, Collection Date January 2020, 29903 bp), MT066156 (Italy, January 30, 2020, 29867 bp), MN985325 (USA, Collection Date January 19, 2020, 29882 bp) and MT126808 (Brazil, Collection Date February 28, 2020, 29876 bp). It is noted that the above sequences correspond to

![Fig. 1](image-url)

Fig. 1. (a) Behavior of the fluctuation function for the four different mapping rules. The DFA was carried out for the full sequence (about 29800 bp). (b) Local scaling exponent obtained as the derivative of the log-log plot of the fluctuation function. (c) Fluctuation function for the AK Rule compared with that obtained from 200 shuffled (i.e., randomized) sequences.
isolates made before the variants of concerns by WHO. For instance, the Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617.2) earliest samples were obtained not before May 2020. The aim of considering the above isolates was to evaluate earlier variations of the genome sequence in terms of the fractal scaling properties.

5. Results and discussion

5.1. Global scaling pattern

Fig. 1.a shows the results of applying the DFA to the Wuhan Hu-1 sequence (MN908947) for the four different numerical mapping rules. The dotted line denotes the scaling behavior for an uncorrelated control noise, corresponding to the scaling exponent $\alpha = 0.5$. The four mapping rules exhibited a scaling behavior similar to that of uncorrelated noise for scales of up to 800 bp, which is in line with results reported by Buldyrev et al. [18] for genome sequences from different sources from eukaryotes to vertebrates. A marked deviation from the uncorrelated behavior was observed for scales higher than 1000 bp. In such a case, the fluctuation function points are located above the line of the uncorrelated noise, suggesting the presence of long-term correlations. A more detailed view of the scaling behavior can be assessed by computing the local scaling exponent $\alpha(n)$ computed as the local derivative of the log–log plot [21] and the results are exhibited by Fig. 1.b. These results confirm that the scaling behavior of the Covid-19 genome is scale-dependent, with a pattern close to that of uncorrelated dynamics for relatively small scales, up to about 700–800 bp. On the other hand, the local scaling exponent $\alpha(n)$ is visibly higher than 0.5 for long scales, a behavior that in principle may be ascribed to the presence of long-term serial correlations. Buldyrev et al. [22] postulated that the detected correlations for scales of hundreds and small thousands of bp could be attributed to simple-repeat expansions. Repeated patterns of two or more nucleotides commonly exhibit fat tails, which may be fit by a power-law function [22]. In this way, the deviations of the scaling exponent from $\alpha = 0.5$ for high scales do not necessarily reflect long-term correlations, but instead the presence of fat-tailed distributions.

---

**Fig. 2.** Behavior of the AK Rule scaling exponent (red line) along the genome domain for three different sliding window sizes. (a) 300, (b) 400 and (c) 500 bp. The gray band and the blue line denote respectively the mean value and the 90% CI of the scaling exponent computed from 10,000 shuffled sequences. The vertical lines denote the boundaries of the different functional regions of the genome sequence.
that can be described as power-law functions. This feature is illustrated by Fig. 1.c where the scaling behavior of the AK Rule is compared with that from 200 sequences obtained by shuffling (i.e., randomizing) the original genome sequence. It is noted that the fluctuation function of the original sequence is located into the envelope of fluctuations functions of the shuffled sequences, such that the scaling behavior of the original sequence may not be distinguished from that of random sequences having the same distribution. In this way, one concludes that the analysis of the genome sequence is hampered by spurious effects induced by, e.g., fat-tailed distributions. Carpena et al. [23] reported that no single scaling exists in genome sequences. In this way, the scaling analysis should consider variations of the scaling behavior along the genome sequence.

5.2. Sliding window scaling analysis

The results in Fig. 1 showed that the Covid-19 genome is a complex sequence that is affected by long-term non-stationarities and as such detection of spurious correlations for relatively high scales. In this regard, the analysis is focused on sequences of smaller size via the DFA implementation in a sliding window. This approach offers the advantage that variations of the scaling properties over different genome regions may be detected [9]. Given the uncertainty of the scaling exponent estimates for high scales, Fig. 1.c suggests carrying out the DFA analysis for scales non-higher than 500 bp. To illustrate the effect of the sliding window size, Fig. 2 presents the variations of the scaling exponent for the AK Rule along with the bp position for three different values of the sliding window size (300, 400 and 500 bps) and sliding length of 100 bps. The gray band denotes the 90% randomness confidence interval (CI) band obtained from 10,000 shuffled sequences of the original sliding window sequence, whereas the green line denotes the corresponding mean value. The vertical lines denote the boundaries of the functional domains of the SARS-Cov-2 genome sequence. The scaling exponent exhibits alternating regions of random and correlated behavior, which can be linked to the different functional regions of the genome sequence. In general, the structure of the scaling exponent is preserved when the sliding window size increases from 300 (Fig. 2.a) to 500 bps (Fig. 2.c). One of the effects of the sliding window size was to smooth the scaling exponent fluctuations. The scaling exponent for 300 observations offered more details, although with some high-frequency fluctuations. In this way, the sliding window size of 400 offers a balance between the details and smoothing of the scaling exponent variations. In the sequel, the window size of 400 bp will be used to analyze the Covid-19 genome sequence.

Fig. 3 presents the variation of the scaling exponent for the four different mapping rules. As already shown in Fig. 2, the scaling exponent exhibits important variations along with the genome position. However, the variations and the correlation structure depend on the mapping rule. The scaling exponent for the AK Rule (Fig. 3.a) fluctuates between the 90% CI randomness band and the region corresponding to long-range correlations (i.e., above the randomness band). The large peak at about 5620–6350 bp reflects significant correlations of the region ascribed to the non-structured protein nsp5. It is also remarkable the peak at about 23700–24700 bp, which is in the region responsible for the spike glycoprotein. The PP Rule exhibited a different picture (Fig. 3. b), with most peaks that are outside the 90% CI randomness band.
anti-correlation region (i.e., below the randomness band). A prominent anti-correlation peak was displayed at about 8890–9620 bp, which is linked to the non-structured protein nsp4. The pattern exhibited by the HB Rule is quite interesting (Fig. 3c). In the first third part of the genome sequence, this scaling exponent indicates the presence of long-range anti-correlated patterns. This region is linked to several non-structured proteins (nsp2-nsp9) as well as to the 3C-like proteinase [24]. The behavior of the scaling exponent for higher bp showed a pattern of alternate randomness and anti-correlation, which can be ascribed to the different functional regions. In particular, the anti-correlation peak at 23500–25000 bp indicates that hydrogen bonding structure plays an important role in the formation of the surface glycoprotein. The results for the HP Rule are presented in Fig. 3d, showing that randomness is dominant in the fluctuations of the scaling exponent. The wide peak at about 23150–24300 bp denotes anti-correlations in the surface glycoprotein region. Overall, the results presented in Fig. 3 show that the long-range correlation pattern of the genome sequence depends on the rule used to map the nucleotides sequence. In this way, every mapping rule reveals the presence of correlations/anti-correlations that can be interpreted in terms of the characteristics (e.g., energy bounding) of the molecules involved in the genome structure. Table 1 summarizes the findings of Fig. 3 by denoting the type of scaling pattern (randomness, correlation and anti-correlation) linked to every functional region of the SARS-CoV-2 genome sequence. Except for the initial part linked to the leading protein, all regions can be ascribed to at least one type of scaling pattern. For instance, the helicase region (16237–18039 bp) exhibited correlations with the AK Rule and anti-correlation for the other three rules. Interestingly, the surface glycoprotein region (21553–25384 bp) exhibited the same pattern.

5.3. SARS-CoV-1 versus SARS-CoV-2

The SARS-CoV-1 was detected in November 2002 in Yunnan, China. An interesting question is the similarity of genome sequences between SARS-CoV-1 and SARS-CoV-2. Fig. 4 comparesthe variation of the genome organization of SARS-Cov-2 according to the detrended fluctuation analysis. (DFA).

Table 1

| Region | Function | AK Rule | PP Rule | HB Rule | HP Rule |
|--------|----------|---------|---------|---------|---------|
| 1–265  | 5' UTR   | R       | R       | R       | R       |
| 266–805| Leader Protein | R    | R       | A       | A       |
| 806–2719| nsp2    | R       | R       | A       | R       |
| 2270–8554| nsp3   | C       | R       | A       | C       |
| 8555–10054| nsp4  | C       | C,A     | A       | R       |
| 10055–10972| 3C-like Proteinase | R    | A       | A       | A       |
| 10973–11842| nsp6   | R       | A       | A       | R       |
| 11843–12091| nsp7   | R       | A       | A       | R       |
| 12092–12685| nsp8   | R       | C       | A       | R       |
| 12686–13024| nsp9   | R       | R       | A       | A       |
| 13025–13441| nsp10  | C       | A       | R       | R       |
| 13442–16236| RNA Polymerase | C    | A       | A       | R       |
| 16237–18039| Helicase | C       | A       | A       | A       |
| 18040–19620| EndoRNase| R       | A       | A       | R       |
| 19621–20658| 3'-to-5' Exonuclease | C   | A       | A       | A       |
| 20659–21552| 2'-O-Ribose | R       | C       | R       | A       |
| 21553–25384| Methyltransferase | C   | A       | A       | A       |

| Region | Glycoprotein | Genes: S (Surface Glycoprotein) |
|--------|--------------|---------------------------------|
| 25385–26620| Gene: ORF3a | C       | R       | A       | C       |
| 26621–26472| Gene: E   | R       | R       | A       | R       |
| 26523–27191| Gene: M   | R       | R       | A       | R       |
| 27202–27387| Gene: ORF6 | R       | R       | A       | R       |
| 27394–27759| Gene: ORF7a | R     | C       | A       | R       |
| 27756–27887| Gene: ORF7b | R       | R       | A       | R       |
| 27994–28285| Gene: ORF8 | C       | R       | R       | A       |
| 28274–29533| Gene: N   | R       | R       | A       | A       |
| 29558–29674| Gene: ORF10| R       | R       | A       | R       |

R: random, A: anti-correlated, C: correlated.

5.4. Randomness and site-specific entropy

To assess the level of randomness in the genome sequences, in terms of the two genome sequences, the results in Table 2 indicated that, except for the HP Rule, the SARS-CoV-2 shows a reduced level of randomness relative to the SARS-CoV-1. That is, the genome sequence of SARS-CoV-2 has a less random organization than the genome sequence of SARS-CoV-1 in terms of the amine-keto, purine-pyrimidine and hydrogen bond sequencing. In contrast, the SARS-CoV-2 exhibits a higher degree of randomness relative to the hydrophobicity sequencing. Overall, the results in Fig. 4 and Table 2 indicate that the genome sequences of SARS-CoV-1 and SARS-CoV-2 exhibit important differences in the long-range order (i.e., scaling structure) of the nucleotide base. Results in this line were reported by Ghanchi et al. [4], who found that the earlier genome sequences isolated before June 2020 exhibit a higher mean and site-specific entropy compared with that of posterior isolates. This suggests that the viral genome achieved stability after the initial early period of the pandemic.

The results in Fig. 4 showed important differences between the genome sequences of SARS-CoV-1 and SARS-CoV-2. In particular, the genome sequence of SARS-CoV-2 appeared to be more locally correlated than the genome sequence of SARS-CoV-1. The entropy of the genome sequences was computed to contrast the findings obtained with the DFA [25]. For such computations, the mapped genome sequence is considered as a sequence of \( N_L \) letters. For instance, the PP mapping rule produced a sequence of the two \( (N_L = 2) \) letters +1 and \(-1\). These letters can form \( N_w = N_L^{N_L} \) words of length \( M_w \). Let \( \{i_1, i_2, \ldots, i_{N_w}\} \) the set of words of length \( M_w \). For a given sequence, one can compute the normalized frequency \( f_j \) of the occurrence of the \( j \)-th word \( L_j \), such that \( \sum_j f_j = 1 \). The entropy for the set of words of length \( M_w \) is computed with the Shannon entropy formula
High (resp., low) entropy values indicate a homogeneous (resp., heterogeneous) occurrence of the words of the set \( \{L_1, L_2, \ldots, L_{N_w}\} \). That is, a given word is equally likely to occur than the other words of the set.

The maximum entropy is obtained when \( f_j = 1/N_w \) and is equal to \( \ln(N_w) \). In this way, the above formula can be normalized to obtain the normalized entropy

\[
E_{M_w} = \frac{1}{\ln(N_w)} \sum_{j=1}^{N_w} f_j \ln(f_j)
\]

In this way, the range of \( E_{M_w} \) is \([0,1]\). The entropy computations were carried for words of length \( N_w = 3 \) over a sliding window of 100 nucleotides. Triplets play an important role in the structure of genome sequences since these words have been linked to important diseases [26] and advantageous mutations in evolution [27].

Ad done for the DFA, the randomness of the sequence in the sliding window was tested using the iso-distributional surrogate data test [17] described in Section 2.1. Fig. 5 presents the variation of the normalized entropy \( E_{M_w} \) through the SARS-CoV-1 and SARS-CoV-2 genome sequences. The entropy is contained within the randomness band for most sequences, indicating that the genomic mosaic is highly disorganized in most regions. However, important deviations from randomness can be observed in the nsp4 and surface glycoprotein regions, mainly for the AK and HB mapping rules (Fig. 5a and c). These important regions exhibited domains with important nucleotide organization. The HP Rule was not sufficiently sensitive to variations of the entropy (Fig. 5c) whereas the PP Rule showed marked deviations from randomness in the region of about 3000 bp.
bps. In general, the entropy results in Fig. 5 agree well with the results obtained with the DFA (Fig. 4) in the sense that the genome sequence exhibits alternating regions of disordered and ordered nucleotide arrangements. The information obtained with one method should be complemented with that obtained by the other method. The DFA provides insights on the serial organization while the entropy reveals the diversity of patterns involved in such serial organization [28].

The lower section of Table 2 presents the degree of randomness computed with Eq. (4) for the entropy computations shown in Fig. 5. The SARS-CoV-2 exhibited an increased DR relative to the SARS-CoV-1 for the PP and HB rules. In contrast, the DR decreased for the AK rule. The DR for the HP Rule was practically unchanged. The variation of the DR in Table 2 suggests that the transit from the SARS-CoV-1 to the SARS-CoV-2 was realized by marked increases in the organization of the purine-pyrimidine and hydrogen bond energy mosaic organizations and an increase of the diversity of the amino-keto patterns, mainly in the nsp4 and surface glycoprotein regions. In this way, the hydrogen bonding capacity of the spike protein has been the main objective of drug developments to reduce the negative impact of SARS-CoV-2 infections [29].

5.4. 

Early SARS-CoV-2 variants

The first case of SARS-CoV-2 was detected in December 2019 in Wuhan, China. The genome sequence of the virus was firstly reported on January 5, 2020 [1]. The uncontained worldwide spreading of the virus provoked its intensive replication. As part of the strategy to track the dynamics of the SARS-CoV-2, the evolution of the genome sequence was monitored and the results were reported in publicly available sites (e.g., www.ncbi.nlm.nih.gov). Promptly, the WHO was aware of the different variations of the genome sequence, which would threaten the control of the infection. Variants are characterized by a set of mutations in the genome sequence aimed to enhance viral fitness. Mutations involve different substitutions of nucleoid bases, which in principle should modify the scaling characteristics of the genome sequence. The degree of randomness for sequences isolated in five different countries (see Section 4) isolated in the first months of 2020 is shown in Table 2. Isolates from USA, South Korea and Brazil showed only slight differences with the reference Wuhan Hu-1. In contrast, isolates MT066156 (India) and MT012098 (Italy) exhibited important deviations from Wuhan Hu-1. In such cases, the DR of the amino-keto structure decreased, whereas the DR of the purine-pyrimidine structure was unchanged. In contrast, the DR of the hydrogen-bond and hydrophobicity structures shows a slight increase. It should be recalled that the infection was particularly strong in the Lombardian Region of Italy where the isolate was obtained. The stronger changes of the scaling structure equipped the Italy variant with an improved fitting strategy to prevail over the original and other variants of concern. A more detailed analysis of the scaling differences between the Wuhan Hu-1 and the isolate MT066156 can be carried out by looking at the fluctuations of the scaling exponent. Similar results
were obtained for the isolate MT012098 (India). Fig. 6 presents the results for a sliding window size of 300 observations, which provides a more detailed view of the scaling exponent fluctuations. In general, the scaling exponents of the Wuhan Hu-1 (red line) and the isolate MT066156 (blue line) coincide over the whole range. However, some pointwise differences can be observed, and some of them are indicated by arrows. Such differences are linked to increases in the distance from randomness, and so increases in the local ordering of the genome sequence. The large peak at 6115 bp for the AK Rule (Fig. 6.a) is located in the nsp4 region. Increased correlations about 23,450 and 24,870 bp are linked to mutations of the surface glycoprotein, which can be ascribed to the mutations P681R and D950N, respectively [30]. It has been postulated that the P681R mutation enhanced the cleavage of the full-length spike to the spike 1 and S2 subunits, leading to increased infection via cell surface entry [30]. Besides, mutations in the surface glycoprotein have been linked to increases in hospitalizations with severe outcomes [31]. The peak at about 3000 bp for the PP Rule (Fig. 6.b) reflects a change of the correlation structure in the nsp4 region. On the other hand, the large peak at about 21500 bp can be attributed to the mutation T19R15 [30], which is linked to the formation of the surface glycoprotein. The large long-range anti-correlation peak at about 23650 bp for the HB Rule (Fig. 6.c) is also linked to the surface glycoprotein and can also be related to the P681R mutation. The increase of anti-correlated behavior was also reflected at about 23800 bp for the HP Rule (Fig. 6.d).

5.5. Discussion

The results aforementioned showed that the SARS-CoV-2 genome sequence has a complex structure, with the correlation structure depending on the rule used for mapping the sequence into a numeric series. A correlation structure can be ascribed to every functional region, which reflects certain ordering in the mosaic organization of the nucleotide arrangement. Of particular interest is the region 21553–25384 bp responsible for the surface glycoprotein. The spike glycoprotein is located on the outside of the virus and gives coronavirus viruses their crown-like appearance. This glycoprotein plays an important role in the attachment mechanisms of the virus particle and entry into the host cell, and as such is an important target for vaccine development, antibody therapies and diagnostic antigen-based tests [32]. The results in Table 1 showed that the region 21553–25384 bp exhibits a marked scaling structure over the four different mapping rules, such that the correlation structure is linked to the amino-keto, purine-pyrimidine, hydrogen-bond and hydrophobicity sequencing. Besides, early variants, like those isolated in India and Italy, exhibit increased levels of correlation and ordering in the surface glycoprotein region. In this way, the results described before pointed out that an accurate characterization of the structure and function of the surface glycoprotein region should provide valuable insights to develop efficient treatments and vaccines [33,34]. In particular, mutations in the surface glycoprotein region have been explored as the drivers of enhanced transmissibility and
infectibility of the Delta variant [35], isolated as earliest as October 2020, which suggests that variants with mutations that are distinctive of the Delta variant were already present in the early months of 2020. The issue is of prime importance since deeper mutations in the surface glycoprotein region might lead to variants (e.g., Delta plus) with greater binding capacity and escape from antibodies action [36], besides increased pathogenicity and mortality [37]. In this way, several studies have focused on characterizing and understanding the structure of the surface glycoprotein region and the effects of different mutations in such structure [35].

In terms of the approach used in our work, an interesting question is how the mutations of the early-2020 isolates MT012098 (India) and MT066156 (Italy) modified the organization of the surface glycoprotein region [38]. To address this point, the scaling exponent was computed over the surface glycoprotein region 21553–25384 bp. The sliding window size was set as 150 bp to dispose of a detailed view of the fluctuations of the scaling exponent. The results for the four different mapping rules are exhibited in Fig. 7, where the scaling exponents of the Wuhan Hu-1 and the MT066156 isolate genome sequences were compared. Similar results were obtained for the isolate MT012098. In general, the scaling exponent of the two genome sequences has a similar pattern, with several peak deviations from the random behavior. Small differences are noted, which might be induced by the different mutations. Interestingly, the scaling exponent of the MT066156 isolate (blue line) exhibits increased deviations from randomness at certain peak locations indicated by arrows. Such increase of the distance from randomness indicates that the isolates MT066156 and MT012098 have a more ordered genome sequence than the reference sequence Wuhan Hu-1, and the increased ordering is manifested in terms of amino-keto, purine-pyrimidine, weak-strong hydrogen bonding and hydrophobicity. It is apparent that the increased transmissibility documented in the Lombardian Region in the early months of 2020 was linked to modifications of the surface glycoprotein region to obtain a more ordered organization, which would lead to fewer replication errors. Some peaks of the scaling exponent of the isolates MT066156 and MT012098 coincided with reported mutations in the spike region of the Delta variant [39,40]. For instance, the spike L452R mutation has been linked to increased infectivity and potential cellular immunity evasion [41]. Also, it has been postulated that the spike mutation P681R is responsible of the advantages of the Delta variant to bind human lung epithelial cells and primary human airway tissues [42]. The possible mutation L452R was reflected in variations of the scaling structure of the amino-keto (Fig. 7.a) and purine-pyrimidine (Fig. 7.b) structures, whereas the mutation P681R was detected in the hydrogen-bond structure (Fig. 7.c). In this way, the variations of the scaling exponent between the different variations relative to the reference Wuhan Hu-1 sequence can be used as a guideline to explore the presence of mutations of interest [43].

6. Conclusions

This work used detrended fluctuation analysis (DFA) to study the mosaic organization of the SARS-CoV-2 genome sequence. The DFA was

---

**Fig. 7.** Comparison of the scaling exponent fluctuations between the Wuhan Hu-1 (red line) and the isolate MT066156 (Italy, blue line) genome sequences in the surface glycoprotein region for (a) AK Rule, (b) PP Rule, (c) HB Rule, and (d) HP Rule. The arrows indicate the location of some reported mutations. The gray band denote the 90% CI of the scaling exponent computed from 10,000 shuffled sequences. The vertical lines denote the boundaries of the different functional regions of the genome sequence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
implemented on a sliding window scheme to visualize the fluctuations of the scaling exponent over the different genome sequence functional regions. The method was endowed with a bootstrap scheme to compute a randomness confidence band computed from the statistics of many shuffled sequences. The sliding window scaling exponent showed that some regions of the genome sequence are likely random, whereas other regions exhibit serial correlated/anti-correlated behavior. The behavior of the scaling exponent depended on the rule used to map the nucleotide genome sequence into a numerical sequence. In this way, the scaling exponent provides valuable information on the amino-keto, purine-pyrimidine, hydrogen-bond and hydrophobic organization of the genome sequence. The Wuhan Hu-1 was used as the reference sequence to contrast the results of mutations and variants. The different SARS-CoV-2 spikes variants exhibit a markedly more ordered sequence than the preceding SARS-CoV-1, which would imply that the genome sequence of SARS-CoV-2 has more algorithmically efficient surveillance mechanisms. That is, the more ordered genome sequence of SARS-CoV-2 would be more efficient for replication by reducing the replication errors. Most SARS-CoV-1 variants exhibited only slight modifications of the scaling exponent behavior as compared to the reference Wuhan Hu-1. However, some samples isolated in the early months of 2020 contain modifications that increase the sequence order. In particular, the surface glycoprotein (spike) region of the isolates MT012098 (India) and MT066156, 29,867 (Italy) showed increased deviations from randomness at specific locations that can be linked to reported important mutations. It is apparent that the increase in the order in the spike region equipped some variants with a more informationally efficient procedure for replication, and hence for binding with the host cells. Overall, the results reported in this work show that the scaling analysis based on a sliding window DFA can provide valuable insights on the organization of the SARS-CoV-2 genome sequence, and the drawn information could be used to help the detection and monitoring of regions of interest in terms of mutations of concern.

CRediT authorship contribution statement

**M. Meraz:** Conceptualization, Investigation, E.J. Vernon-Carter: Conceptualization, Investigation. E. Rodriguez: Visualization, Investigation, Writing – review & editing. J. Alvarez-Ramirez: Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y.-J. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y.-L. Qiu, M.-Q. Wang, J.-Z. Zheng, L. Xu, E.C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature 579 (7798) (2020) 265–269, https://doi.org/10.1038/s41586-020-2008-3.

[2] R.A. Khailany, M. Safdar, M. Ozaslan, Genomic characterization of a novel SARS-CoV-2, Gene Rep. 19 (2020) 100682, https://doi.org/10.1016/j.genrep.2020.100682.

[3] W.T. Harvey, A.M. Carabelli, B. Jackson, R.K. Gupta, E.C. Thomson, E.M. Harrison, J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J.D. Farmer, Testing for nonlinearity in time series: the method of surrogate data, Physica D 58 (1992) 77–94, https://doi.org/10.1016/0167-2789(92)90102-S.

[4] W.T. Harvey, A.M. Carabelli, B. Jackson, R.K. Gupta, E.C. Thomson, E.M. Harrison, J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J.D. Farmer, Testing for nonlinearity in time series: the method of surrogate data, Physica D 58 (1992) 77–94, https://doi.org/10.1016/0167-2789(92)90102-S.

[5] M. Meraz et al., M.S. Nawaz, P. Fournier-Viger, X. Niu, Y. Wu, J.C.W. Lin, July). COVID-19 genome - genrep.2020.100682 .

[6] H. Namazi, Complexity-based classification of the coronavirus genome versus genomes of the human immunodeficiency virus (HIV) and dengue virus, Fractals 28 (2020) 2050012, https://doi.org/10.1142/S0218348X20500127.

[7] T. de Salazar e Femandes, J.S. de Oliveira Filho, I.M.S. da Silva Lopes, Fractal signature of coronaviruses related to severe acute respiratory syndrome, Res. Biomol. Eng. (2020), https://doi.org/10.1016/j.sbiom.2020.0609-5.

[8] C.K. Peng, S.V. Buldyrev, H.E. Stanley, F. Liljeros, Y. Aharony, M. Simons, H. E. Stanley, Statistical properties of DNA sequences, Physica A 221 (1–3) (1995) 180–192, https://doi.org/10.1016/0378-4371(95)00247-5.
X. Deng, M.A. García-Knight, M.M. Khalid, V. Servellita, C. Wang, M.K. Morris, B. Schr, R. Challen, E. Brooks-Pollock, J.M. Read, L. Dyson, K. Tsaneva-Atanasova, D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M.M. Rajah, S.R. Kannan, A.N. Spratt, A.R. Cohen, S.H. Naqvi, H.S. Chand, T.P. Quinn, C. H.G. Toor, D.I. Banerjee, S. Lipsa Rath, S.A. Darji, Computational drug repurposing targeting the spike glycoprotein of SARS-CoV-2 as an effective strategy to neutralize COVID-19, Eur. J. Pharmacol. 890 (2021) 173720, https://doi.org/10.1016/j.ejphar.2020.173720.

S. Ogountomi, O.E. Akindiyi, B.O. Olamoyegun, V.O. Aremu, A.O. Arowosaye, D. O. Abodertin, H.B. Bello, T.Y. Sentadejo, E.H. Awoyela, A.A. Oladipo, B.B. Oladipo, I.O. Ajayi, O.N. Majolaębę, O.M. Gawaye, J.R. Okole, Exploration of surface glycoprotein to design multi-epitope vaccine for the prevention of Covid-19, Inf. Med. Unlocked 21 (2020) 100438, https://doi.org/10.1016/j.imu.2020.100438.

U.F. Chowdhury, M.U.S. Shohan, K.I. Hoque, M.A. Beg, M.K.S. Siam, M.A. Moni, M. Meraz et al., L. Courtellemont, H. Pad, A. Velay, J. Le Guen, N. Kassis-Chikhani, D. Edriss, L. Belec, A. Seve, –276, 202012/1: matched cohort study, BMJ 372 (2021) n579, https://doi.org/10.1136/bmj.n579.

L. Danon, Risk of mortality in patients infected with SARS-CoV-2 variant of concern Delta to antibody neutralization, Nature 596 (7871) (2021) 102715, https://doi.org/10.1016/j.ejphar.2020.173720.

C. Liu, H.M. Ginn, W. Dejnirattisai, P. Supasa, B. Wang, A. Tuekprakhon, R. Nultalai, D. Zhou, A.J. Mentzer, Y. Zhao, H.M.E. Dayvesteyn, Cesar Lopez-Camacho, J. Sloan-Campon, T.S. Walter, D. Skelly, S.A. Johnson, T.G. Ritter, C. Mason, S.A. Costa Clemens, F. Gomes Naveca, V. Nascimento, P. Nascimento, F. Fernandes da Costa, P.C. Resende, A. Pauvolid-Correa, M.M. Siqueira, C. Dold, N. Tempertton, T. Dong, A.J. Pollard, J.C. Knight, D. Crook, T. Lambe, E. Chuterback, S. Bibi, A. Flaxman, M. Bittaye, S. Belij-Rammerstorfer, S.C. Gilbert, T. Malik, M.W. Carroll, P. Klenerman, E. Barnes, S.J. Dunachie, V. Bailie, N. Serafin, Z. Ditto, K. Da Silva, N.G. Paterson, M.A. Williams, D.B. Hall, S. Madhi, M.C. Nunes, P. Boulder, E.E. Fry, J. Mangokolapya, J. Ren, D.I. Stuart, G. R. Sereaton, Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum, Cell 184 (16) (2021) 4220–4236.e13.

C. Motozono, M. Toyoda, J. Zahradnik, A. Saito, H. Nasser, T.S. Tan, I. Ngare, I. Kimura, K. Uru, Y. Konogi, Y. Yue, R. Shimizu, J. Ito, S. Torii, A. Yonekawa, N. Shimono, Y. Nagasaki, R. Minami, T. Toya, N. Sekiya, T. Fukuhara, Y. Matsuzura, G. Schreiber, T. Ikeda, S. Nakagawa, T. Ueno, K. Sato, SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity, Cell Host Microbe 29 (7) (2021) 1124–1136.e11.

B. Li, A. Deng, K. Li, Y. Hu, Z. Li, Q. Xiong, Z. Liu, Q. Guo, L. Zou, H. Zhang, M. Zhang, F. Quyang, J. Su, W. Su, J. Xu, H. Lin, J. Sun, J. Peng, H. Jiang, P. Zhou, T. Hu, M. Luo, Z. Zhang, H. Zheng, J. Xiao, T. Liu, B. Che, H. Zeng, Z. Zheng, Y. Huang, J. Yu, Y. Ji, W. Ji, J. Chen, H. Zhong, X. Deng, M. Kang, O. G. Pybus, M. Hall, K. A. Lythgoe, Y. Li, J. Yuan, J. He, J. Lu, Viral infection and transmission in a large well-traced outbreak caused by the Delta SARS-CoV-2 variant, medRxiv (2021) doi:10.1101/2021.07.07.21260122.

L. von Dory, C.J. Holdcroft, D. Richard, F. Baloux, COVID-19, the first pandemic in the post-genomic era, Curr. Opin. Virol. 50 (2021) 40–48, https://doi.org/10.1016/j.coviro.2021.07.002.