

Analysis of similarities in respiratory tract genomes: an approach using the Hurst exponent and the non-decimated discrete wavelet transform

Análise das semelhanças nos genomas do trato respiratório: uma abordagem usando o expoente de Hurst e a transformação discreta de wavelet não-dizimada

DOI: 10.54033/cadpedv20n9-021

Recebimento dos originais: 08/11/2023
Aceitação para publicação: 11/12/2023

Dulcília Carlos Guezimane Ernesto

Doctoral Student in Agricultural Statistics and Experimentation
Institution: Universidade Púnguê and Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: dcg.ernesto@gmail.com or dcg.franque@gmail.com

Leila Maria Ferreira

PhD in Agricultural Statistics and Experimentation
Institution: Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: leilamaria2003@gmail.com

Thelma Sáfadi

PhD in Statistics
Institution: Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: safadi@ufla.br

ABSTRACT

The implementation of statistical methods that can make the similarity study process more efficient has been very common in current research, because some methods such as wavelets make the processing of data manipulation faster and with the ability to process thousands of nucleotides in a genomic sequencing, and consequently better estimates are obtained. Thus, this work aimed to analyze some genomes of the Coronaviridae family and the Paramyxoviridae family, looking for similarities through signal decomposition considering the non-decimated discrete wavelet transform techniques. The entire genome sequences of the families were initially extracted from the NCBI website and then pattern analysis was performed using a signal processing method, based on the GC

content. For each GC sequence, decomposition was performed using the six-level discrete non-decimated Daubechies wavelet transform. Then, the Hurst exponent was calculated for each level of decomposition and the sequences were verified with similar patterns. After the cluster analysis, it was possible to obtain that the pair Bet1 and MERS proved to be similar in almost all methods, the other pair of sequences that proved to be similar by the absolute moments methods and by the aggregated variance method was the pair Gamma1 and Del, the group composed of Infl1, Infl4, Infl3 and Hendra was similar by the absolute moments methods, aggregated variance, and in the R/S analysis method, there was a substitution of Infl4 by Infl5. At the end of the study, it was possible to conclude that the non-decimated discrete wavelet transform allowed us to decompose each one of the sequences and, consequently, a more detailed study of similarity can be carried out, and it was possible to obtain, by some methods, strains of Coronaviridae that do not resemble the strains of Paramyxoviridae.

Keywords: genetic sequencing, Hurst exponent, similarity, wavelets.

RESUMO

A implementação de métodos estatísticos que podem tornar o processo de estudo de similaridade mais eficiente tem sido muito comum na pesquisa atual, porque alguns métodos, como as ondas, tornam o processamento de manipulação de dados mais rápido e com a capacidade de processar milhares de nucleotídeos em uma sequência genômica, e conseqüentemente, melhores estimativas são obtidas. Assim, este trabalho teve como objetivo analisar alguns genomas da família Coronaviridae e da família Paramyxoviridae, procurando semelhanças através da decomposição de sinal considerando as técnicas de transformação de ondas discretas não-decimadas. Todas as sequências genômicas das famílias foram inicialmente extraídas do site do NCBI e, em seguida, a análise de padrões foi realizada usando um método de processamento de sinais, com base no conteúdo GC. Para cada sequência de GC, a decomposição foi realizada usando a transformação de onda de Daubechies discreta de seis níveis e não-dizimada. Então, o expoente de Hurst foi calculado para cada nível de decomposição e as sequências foram verificadas com padrões semelhantes. Após a análise de cluster, foi possível obter que o par Bet1 e Mers se mostraram semelhantes em quase todos os métodos, o outro par de sequências que se mostraram semelhantes pelos métodos de momentos absolutos e pelo método de variância agregada foi o par Gamma1 e Del, o grupo composto de Infl1, Infl4, Infl3 e Hendra foi semelhante pelos métodos de momentos absolutos, variância agregada, e no método de análise R/S, houve uma substituição de Infl4 por Infl5. No final do estudo, foi possível concluir que a transformação discreta de ondas não-dizimadas permitiu decompor cada uma das sequências e, conseqüentemente, um estudo mais detalhado de similaridade pode ser realizado, e foi possível obter, por alguns métodos, linhagens de Coronaviridae que não se assemelham às linhagens de Paramyxoviridae.

Palavras-chave: sequenciamento genético, expoente de Hurst, similaridade, wavelets.

1 INTRODUCTION

Finding similarities in the sequencing of some evolutionarily similar species of viruses could be a way for scientists to find the formula for a preventive vaccine, or for the development of a cure for a disease. Therefore, the implementation of statistical methods that can make the process of studying similarities more efficient is welcome in the sense that the faster and with the capacity to process thousands of nucleotides in a genomic sequence, the better estimates will be obtained. Furthermore, if the researcher intends to identify emerging viruses and bacteria and work with a very large genomic database, it is essential to implement a clustering methodology that is reliable for genomic sequences in real time.

Several methodologies have been used by researchers to study genetic sequencing; however, wavelets have become an effective and efficient alternative for processing genetic sequences in real time, to study similarities, but specifically the method of estimating the exponent of Hurst in this type of study involving respiratory tract virus genomes has been very little used to date. In this work, 5 different methods were applied to estimate the Hurst exponent to study similarities in the genomes of respiratory tract viruses of the families Coronaviridae and Paramyxoviridae. This methodology has already been applied in the work by Ferreira et al. (2020) finding similarities in tuberculosis strains.

There are many approaches so far involving the Hurst exponent and wavelets. Das and Kumar (2021) carried out research in the financial area with the objective of optimizing a portfolio and showing that by combining the Hurst exponent and wavelet analysis they could help in increasing portfolio returns. In this study the authors use the Hurst exponent and wavelets to analyze the long-term dependencies between sovereign bonds and sector indices in India.

Furthermore, Arouxet et al. (2022) analyzed the impact of covid19 on cryptocurrencies using the Hurst exponent and wavelets. Their research examines long-term memory of return and volatility using seven-currency high-frequency time series. This analysis mainly focused on two periods, namely: the pre-covid19 period and the subsequent pandemic period, implemented the wavelet transform as it allowed to have more robust estimators of the Hurst exponent.

Considering all above, this study in general aims to analyse some genomes of the Coronaviridae family and the Paramyxoviridae family looking for similarities through signal decomposition with the support of the non-decimated discrete transform techniques of the Daubechies wavelet. Furthermore, the study seeks to apply the R/S, Peng methods, aggregate variation, differentiated aggregate variation, and the method of absolute moments to analyze similarities between the genomes of the Coronaviridae and Paramyxoviridae family.

2 MATERIAL AND METHODS

Complete genome sequences of the Coronaviridae and Paramyxoviridae virus families obtained from the National Center for Biotechnology Information (NCBI) database were considered. Therefore, it is essential to know the virus that circulates in the population infected by both Sars-Cov-2 and parainfluenza so that researchers can track its mutations and map the circulating strains and later develop more efficient methods of combating diseases from these viruses. Table 1 shows information related to each of the sequences in the NCBI, the first column is the identification of the access number on the platform, the second column is the nomenclature of each sequence, within the third column is the abbreviation created to facilitate the elaboration of the dendrograms, the fourth column shows the GC coefficient rate per sequence, and the last column shows the number of nucleotides for each sequence.

Table 1. Genomes Description of the Paramyxoviridae and Coronaviridae families

NCBI Access	Name of the Virus	Sequence	Total GC Content Rate	Nucleotide number
NC_038294	Betacoronavirus England 1	Bet1	0.3205908	14676
NC_019843	Middle East respiratory syndrome-related coronavirus	MERS	0.4117764	14676
NC_006213	Human coronavirus OC43	Bet2	0.36788	14676
NC_003045	Bovine coronavirus	Bet4	0.36788	14676
NC_004718	SARS coronavirus TOR2	Bet5	0.4076166	14676
NC_028752	Camel alpha coronavirus	Alf1	0.3446086	14676
NC_002645	Human coronarirus 229E	Alf2	0.3841323	14676
NC_048214	Coronavirus duck	Gama1	0.3826189	14676
NC_010800	Turkey coronavirus	Gama2	0.4123643	14676
NC_005831	Coronavirus humano NL63	Alfa	0.3927362	14676

NCBI Access	Name of the Virus	Sequence	Total GC Content Rate	Nucleotide number
NC_006577	Human Coronarivirus HKU1	Beta	0.382521	14676
KJ481931.1	Deltacoronavirus	Del	0.4317878	14676
NC_003461.1	Human parainfluenza virus 1	Influ1	0.3724359	14701
NC_001796.2	Human parainfluenza virus 3	Influ3	0.3452335	14676
NC_021928.1	Human parainfluenza virus 4a viral cRNA	Influ4	0.3623035	14676
NC_006430.1	Parainfluenzavirus 5 strain W3A	Influ5	0.4226682	14676
NC_001906.3	Hendrahenipavirus	Hendra	0.394428	14676

Source: The authors.

Furthermore, the Table 1 show, the sequence classes belonging to the Coronaviridae family: Beta, Bet1, Bet2, Bet3, Bet4, Bet5, MERS, Alpha, Alf1, Alf2, Del, Gamma1, Gamma2, and those that belong to the Paramyxoviridae family: Influenza1, Influenza3, Influenza4, Influenza5 and Hendra.

Therefore, the variants belonging to the Coronaviridae family that have been the subject of research are divided in two groups, the variants that have humans and vertebrate animals as hosts, and those that have only humans as hosts. However, the Bet1 variant, which was isolated in England (de Groot et al. 2013), the MERS variant which originated from the HCoV-EMC strain and was isolated in 2012 in a patient in Saudi Arabia (van Boheemen et al. 2012), the Bet2 variant originated from the ATCC VR-759 strain (St-Jean et al. 2004), the Bet4 variant isolated from BCoV-ENT (Chouljenko et al. 2001), the Bet5 variant isolated from Tor2 (He et al. 2004), the Alf1 variant isolated from camel/Riyadh/Ry141/2015 and that originated from MERS-CoV in Saudi Arabia.

The Alf1 variant have three camel coronavirus species, and that according to Sabir et al. (2016) MERS-CoV and a human strain of CoV 229E species circulated simultaneously with high prevalence causing respiratory tract co-infections in dromedary camels, and camels shared the three covid species with humans.

The Alf2 variant, the Gamma1 variant dominant in birds such as duck, isolated from DK/GD/27/2014 (Zhuang et al. 2015), the Gamma2 variant isolated from an outbreak of enteritis in young turkeys (Gomaa et al. 2008), and the Del variant, originating in pigs in the United States from the

SDCV/USA/Illinois121/2014 strain (Marthaler et al. 2014), are variants with human and vertebrate animal hosts.

While the variants: Alpha, which has a complete genome sequence, which indicates that the virus does not come from recombination's (van der Hoek et al. 2004) and the Beta variant, first diagnosed in Shenzhen, China in a 71-year-old patient with pneumonia (Woo et al. 2005), have their occurrence only in human hosts.

On the other hand, the Paramyxoviridae family is also composed of Influenza 1 which is originated from the Washington strain (Newman et al. 2002), Influenza 3, Influenza 5 which originated from strain W3A, and Hendra which originated in Australia and was diagnosed for the first time in an outbreak that caused the death of horses (Wang et al. 2000). Furthermore, these strains of the Paramyxoviridae family occur in human hosts and vertebrate animals, however, Influenza 4 is the only strain that occurs only in humans and had its origin in the W-25 strain (Komada et al. 2011).

For each genome, sequences of the GC content were determined, with a sliding window of size $n=100$. Using the Daubechies wavelet with 4 null moments, the non-decimated wavelet transform was applied to the GC sequences, obtaining decomposition in six levels of resolution for each sequence.

To evaluate the similarity between the sequences under study, the Hurst exponent was calculated at each level of decomposition using five different methods, namely: the aggregated variance method, the differentiated aggregated variance method, the absolute moment's method, the Peng method, and the R/S analysis method. Finally, a cluster analysis was performed based on the Mahalanobis distance. All analyses were performed using R software (R Core Team 2017).

Following is presented all the methodology used in the analysis.

2.1 THE GC CONTENT

GC content is often used to map genome composition and understand the evolution of its coding sequence (Saini and Dewan 2016). To obtain the signal referring to the genomes under study, the GC content was estimated with a

sliding window of size $n = 100$. The GC content is calculated as the ratio of the sum of Guanine (G) and Cytosine (C) bases to the sum of the bases Adenine (A), Guanine (G), Cytosine (C) and Thine (T), given by equation 1.

$$GC_{content} = \frac{nG+nC}{nA+nG+nC+nT}, \quad (1)$$

Where

nA , nG , nC and nT represent the number of nucleotides of the bases of A, C, G, and T respectively. The proportions obtained according to Saini and Dewan (2016) can be interpreted, in which GC-rich regions indicate the inclusion of many protein-coding genes, that is, by determining the GC proportion, gene-rich regions in the genome are identified and at the same time as the signals of each sequence are obtained.

After obtaining the signal of each sequence, the decomposition phase followed according to the non-decimated wavelet transform.

2.2 WAVELETS

The wavelets, according to Denault et al. (2021), are mathematical functions that help to conduct transforms, such as the Fourier transform or Haar transform, among others. Brassarote et al. (2015) linked that one of the several applications of wavelets has been in areas of analysis of differential equation signals, this is mainly due to their ability to localize in time-frequency contrary to what occurs with Fourier analysis, bringing these forms a different and innovative perspective on the treatment of a non-stationary function database.

According to Abreu and Speckbacher (2022) wavelets satisfy the indispensable condition of admissibility (equation 2):

$$C_{\psi} = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty. \quad (2)$$

According to Alshammari et al. (2023), there are two types of wavelets, namely father wavelets and mother wavelets, which are represented in equations 3 and 4 respectively:

$$\phi_{j,k} = 2^{\left(\frac{-j}{2}\right)} \phi \left(t - \frac{2^j k}{2^j} \right), \quad (3)$$

$$\psi_{j,k} = 2^{\left(\frac{-j}{2}\right)} \psi \left(t - \frac{2^j k}{2^j} \right),$$

where

$j = 1, 2, 3, \dots, J$, and J are the levels of decomposition.

$$\int \phi(t) dt = 1, \quad \int \psi(t) dt = 0 \quad (4)$$

It is recommended that wavelets have compact support. Still from the perspective of Alshammari et al. (2023), the general mathematical model is represented in equation 5:

$$F_{j,k} = \int \psi_{j,k} f(t) dt \quad (5)$$

2.2.1 Daubechies Wavelets

Daubechies wavelets belong to the family of orthogonal wavelets and are characterized by the number of null moments according to an established support, in addition, they define a transformation of discrete wavelets, according to Rashid et al. (2020), the Daubechies wavelet can make discrete wavelet analysis feasible. They are defined by the number of null moments (m), which corresponds to the filter length ($2m$) (Rashid et al. 2020). It is advantageous to work with wavelet functions and Daubechies scale functions as functional bases for the multi-resolution approximation as they can radically change the formulation and solution of the equations of motion (Nastos and Saravanos 2021).

2.2.2 Non-Decimated Discrete Wavelet Transform

According to Wang et al. (2019) the non-decimated discrete wavelet transform not only retains the properties of the decimal wavelet transform, but also has the property of being invariant to displacement, that is, the non-

decimated discrete wavelet transform does not use the down sampling operation during the signal decomposition process, it inserts zeros every two coefficients in order to make the filter expansion in the high-pass and low-pass filtering process.

According to Yu et al. (2021), at each decomposition level of the non-decimated discrete wavelet transform, the scale filter ($H_0(z)$) and the wavelet filter ($H_1(z)$) are upsampled by filling zeros for each response coefficient to the impulse of the filters. The coefficients of filters at a certain level j are calculated by equations 6 and 7.

$$d_j^{2n}(-k) = d_{j-1}^n(k) * h_{0j}(k) \quad (6)$$

$$d_j^{2n+1}(-k) = d_{j-1}^n(k) * h_{1j}(k) \quad (7)$$

2.3 HURST EXPONENT

According to Soterroni et al. (2008), the Hurst exponent (H) is a tool capable of providing information related to correlation and persistence in a time series.

According to Rout et al. (2022), the Hurst exponent is defined in the range (0.1) and is strictly less than 0.5 in sequences that are anticorrelated and is in the range of $0.5 < H < 1$ when the sequences are positively correlated. For cases where the Hurst exponent is equal to 0.5, it can be concluded that the sequence presents random patterns (Rout et al. 2022).

The Hurst exponent can be estimated considering 5 different techniques:

2.3.1 R/S Analysis

According to Lara-Musule et al. (2021), the R/S analysis measures the amplitude of the deviations of the partial sums from the mean rescaled by the standard deviation of the series. The R/S analysis method is described by Bărbulescu et al. (2010) and by Lara-Musule et al. (2021). Therefore, according to Lara-Musule et al (2021), when we take a time series $Z_N = Z_i$ with length N , we consider a subsequence $X_{N_S} = X_i$ with length N_S where $N_S < N$. Therefore, the R/S analysis for X_{N_S} is calculated as follows:

$$(R/S) = \frac{1}{\sigma_S(N_S)} \left\{ \max_{1 \leq i \leq M} \sum_{k=1}^i (X_k - \bar{X}_{N_S}) - \min_{1 \leq i \leq M} \sum_{k=1}^i (X_k - \bar{X}_{N_S}) \right\}. \quad (8)$$

Where

$\bar{X}_{N_S}$ is the mean of the subsequence and $\sigma_S(N_S)$ is the standard deviation of the sample, which are defined as:

$$\bar{X}_{N_S} = \frac{1}{N_S} \sum_{i=1}^{N_S} X_i \quad (9)$$

and

$$\sigma_S(N_S) = \sqrt{\frac{1}{N_S} \sum_{k=1}^{N_S} (X_k - \bar{X}_{N_S})^2} \quad (10)$$

Still from the perspective of Lara-Musule et al. (2021) the R/S analysis is guided by the power law, which states that:

$$(R/S) = aN_S^H \quad (11)$$

Where

a is a constant and H is the Hurst exponent. A log-log plot of (R/S) versus $N_S \in (N_{(S,min)}, N_{(S,max)})$, gives a straight line with slope H . Therefore, if the plot is a straight line with a slope $H = 0.5$, means that the time series data are independent. Furthermore, if $H > 0.5$ indicates that the time series is persistent and at the same time presents a long-term autocorrelation. And if $H < 0.5$, it means that the autocorrelations in the signal are anti-persistent.

2.3.2 Aggregate Variance Method

The aggregate variance method is based on the self-similarity property of the samples of a process (Aldea and Tarniceriu 2013). When considering X a time series with length N , it is divided into d subseries of length m , and for each aggregate series, composed by an average given by Aldea and Tarniceriu (2013):

$$X^m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2, \dots, d \quad (12)$$

and by the sample variance given by Bărbulescu et al. (2010):

$$\text{Var}X^m = \frac{1}{d} \sum_{k=1}^d (X^{(m)}(k) - \bar{X})^2. \quad (13)$$

Bărbulescu et al. (2010) also state that for successive values of m , the sample variance is computed in a log-log plot against m . And the least squares will be fitted on a line for all points on the graph, and the Hurst coefficient will be calculated considering the slope of the straight line which is given by $2H-2$.

2.3.3 Differentiated Aggregate Variance Method

This methodology was described by Teverovsky and Taqqu (1997), in which they examined the effect of certain types of non-stationarity on the detection of long-range dependence on Hurst's estimate of parameter H , when the variance estimator is applied. The purpose of applying this method is to distinguish the case in which H is close to 0.5 and there is a certain non-zero trend, from the cases in which H is far from 0.5, or otherwise, ends up being significantly greater than 0.5.

According to Teverovsky and Taqqu (1997), to detect a long-range dependence it is necessary to analyse the sample variation of a time series $X = \{X_i, i = 1, 2, \dots\}$, at several levels. of aggregation m , in the aggregate series of order m . By taking the following equation,

$$X^m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2, \dots, d \quad (14)$$

Therefore, the sample variation of the aggregate series X^m will be given by:

$$\widehat{\text{var}} X^{(m)} = \frac{1}{N/m} \sum_{k=1}^{N/m} \{X^{(m)}(k)\}^2 - \left\{ \frac{1}{N/m} \sum_{k=1}^{N/m} X^{(m)}(k) \right\}^2 \quad (15)$$

The aggregate variance method can acquire the estimates through the logarithm of the first order difference. When considering the sequence given by m_1, m_2 from the values of m , and taking the difference, we obtain the following (Teverovsky and Taqqu 1997):

$$\frac{d\widehat{var}X^{(m)}}{dm} \approx \beta C_2 m^{\beta-1} \quad (16)$$

and it is assumed that

$$\Delta\widehat{var}X^{(m)} \approx \frac{d\widehat{var}X^{(m)}}{dm} \Delta m$$

Then

$$\log(\Delta\widehat{var}X^{(m)}) \approx \log\left(\frac{d\widehat{var}X^{(m)}}{dm}\right) + \log \Delta m$$

Analogously it is necessary to consider that

$$\frac{d}{dm} Var[\bar{X}_m(k)] \approx (2H - 2)Cm^{2H-3}.$$

Since we will work with log-log graphs, the following approximations will be used:

$$\log \Delta Var[\bar{X}_m(k)] = (2H - 3) \log m + \log(2H - 2)C + \log m + C_1 = (2H - 2) \log m + C_2 \quad (17)$$

When applying this method, it is expected that a line with slope equal to $2H-2$ will also be obtained.

2.3.4 Absolute Moments Method

From the perspective of Garcin (2022), this method allows the researcher to make a good assessment of the model specification, and is a generalization of the aggregate variance method as it uses the same principle as X^m (Aldea and Tarniceriu 2013).

The absolute moments are calculated for the aggregate series (Bărbulescu et al., 2010), and the n th absolute moment is given by:

$$AM_n^{(m)} = \frac{1}{(N/m)} \sum_{k=1}^{(N/m)} |\bar{X}_m(k) - \bar{X}_N|^n \quad (18)$$

where

$AM_n^{(m)}$ is asymptotically proportional to $m^n(H-1)$.

In this method the estimates of the Hurst exponent will be calculated starting from the calculation of $AM_n^{(m)}$ considering different values of m and generating a log-log plot against m . What is expected when applying this methodology is that the points are spread along a straight line, with slope given by $n(H-1)$.

2.3.5 Peng Method

This method is discussed in the literature by Adler et al. (1998), in which four fundamental steps can be applied:

1st Step: observe each block of size m and calculate the partial sum within the blocks according to equation 19.

$$Y(k)^m = \sum_{t=(k-1)m+1}^{km} X(t), \quad k = 1, 2, \dots, (N/m) \quad (19)$$

2nd Step: this step comprises the adjustment phase of a regression line of the type $g = a + bk$, which will help us to calculate the variance of the residual, which is given by equation 20.

$$s_r^{(m)} = \frac{1}{m} \sum_{k=1}^{N/m} (Y(k)^m - a - bk)^2 \quad (20)$$

Step 3: Sketch the graphs of $\log s_r^{(m)}$ versus $\log m$

4th Step: Finally, the slope to be obtained must be equal to 2H.

Cluster Analysis

Cluster analysis is the classification of objects by groups, that is, each group must contain objects with similar characteristics, according to some statistical distance function.

From the perspective of Manly (2008), two forms of grouping can be used, the hierarchical and division grouping method.

According to Frei (2006), hierarchical methods follow a specific algorithm: Search for the two most similar objects in the similarity matrix.

Objects i and j are removed and form a group, in which the corresponding row and column i and j are eliminated.

A row and a column are defined, which are obtained by the distances between the group i j and the remaining objects, according to the adopted algorithm procedure.

The previous steps are repeated n-1 times, so that all n objects have a group until the end of the algorithm.

Furthermore, hierarchical techniques were used, as they allow the construction of dendrograms, and the implementation of these hierarchical techniques begins with the calculation of the distances of each object to all other objects. The distances from each object to all objects will be calculated using the mahalanobis distance function.

According to Yang and Delpha (2022), the Mahalanobis distance of a sample is calculated considering two sample vectors X, Y from the same distribution and is given by:

$$d_M(X) = \sqrt{(X - \mu)^T \Sigma^{-1} (X - \mu)} \quad (21)$$

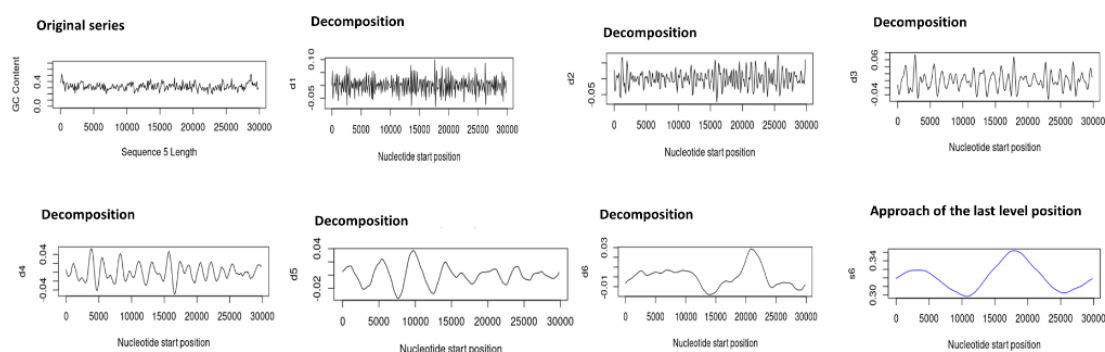
where

μ is a mean vector, and Σ is the covariance matrix. Therefore, the Mahalanobis distance is always a non-negative value, which is calculated in units of deviations, and takes variability into account (McLachlan 1999).

3 RESULTS

Table 1 summarizes the important aspects of the analyzed genomes. The decomposition of the GC content sequence is exemplified by Betacoronavirus England and is shown in Figure 1.

Figure 1: Decomposition of the human coronavirus HKU1(Beta) sequence



Source: The authors.

Table 2 illustrates the Hurst exponent values for the aggregated absolute value method for the decomposition levels of each GC sequence under study.

Table 2. Hurst Exponent Values by absolute moments method

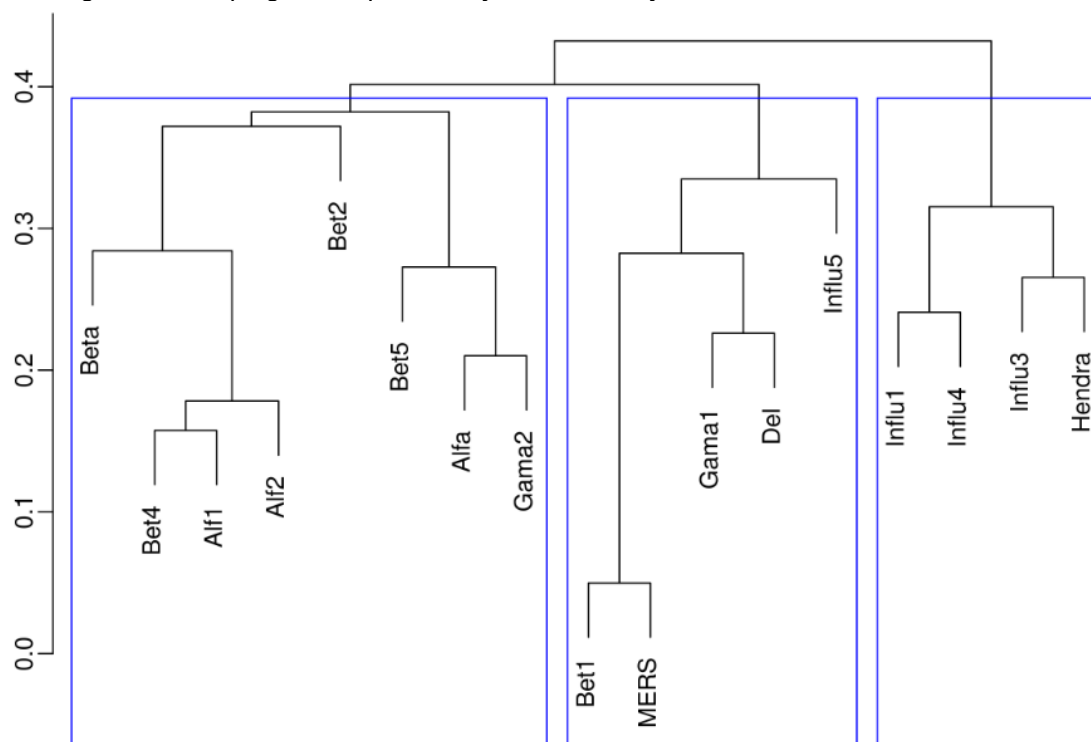
Sequences	Level1	Level2	Level3	Level4	Level5	Level6
Beta	0.15925	0.06039	0.03276	0.03453	0.18273	0.83592
Bet1	0.11744	0.14182	0.07244	-0.02129	0.36787	0.73179
Bet2	0.14523	-0.11753	0.10988	-0.10595	0.28212	0.69620
Bet3	0.14523	-0.11753	0.10988	-0.10595	0.28212	0.69620
Bet4	0.11044	0.04285	-0.11222	0.01375	0.27903	0.69817
Bet5	-0.15239	-0.26166	-0.06326	-0.075288	0.25063	0.79646
Alfa	-0.07652	-0.12263	-0.00088	-0.04076	0.28340	0.70964
Alf1	0.06487	0.02972	-0.13859	0.06475	0.23351	0.72348
Alf2	0.08155	0.14401	-0.20865	0.08037	0.18944	0.72569
MERS	0.16466	0.16676	0.07986	-0.02699	0.36907	0.73739
Gama1	1.05926	0.16426	0.21280	-0.12138	0.27105	0.83348
Gama2	-0.11325	-0.04656	0.07619	0.02878	0.33219	0.83523
Del	-0.01649	0.28606	0.06015	-0.04101	0.31838	0.86559
Influ1	-0.03829	-0.17001	0.03567	0.07965	0.26689	0.99896
Influ3	-0.03885	0.09411	-0.02732	0.13057	0.25049	1.03683
Influ4	0.067420	-0.17437	0.05549	0.07060	0.46435	1.09032
Influ5	0.19312	0.20963	-0.06607	-0.14383	0.29561	1.01124
Hendra	0.11724	0.06061	-0.10491	0.09032	0.33878	0.99014

Source: The authors.

The same calculation was made for the other techniques. The package Nbclust of R was used and applied in all five methodologies to verify the optimal

number of clusters to be applied in the formation of the groups, and from Figure 2 it can be observed that for the method of aggregated absolute value we obtained a grouping with 3 groups, and some subgroups.

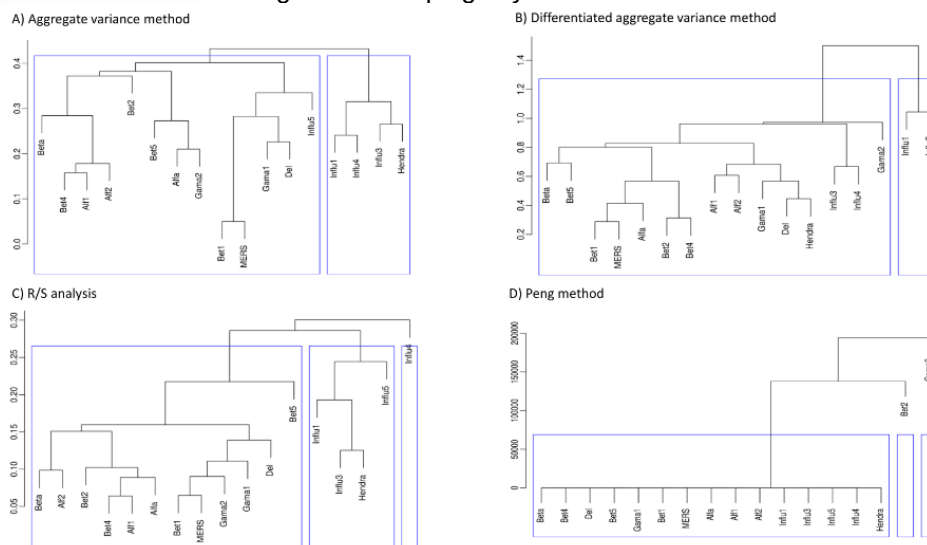
Figure 2: Groupings of sequences by similarities by the method of absolute moments



Source: The authors.

Figure 3 reflects the groups formed by the aggregated variance, differentiated aggregated variance, Peng method and R/S analysis methods, and the aggregated variance and differentiated aggregated variance methods presented the formation of two groups, the R/S analysis presented the formation of three groups and the Peng method presented the formation of four groups. Furthermore, the R/S method and the Peng method had Hurst exponent estimates $H > 0.5$ at levels 4, 5 and 6, which means that the sequences are positively correlated, and levels 1 and 2 presented $H < 0.5$, which indicates the existence of long-range dependence.

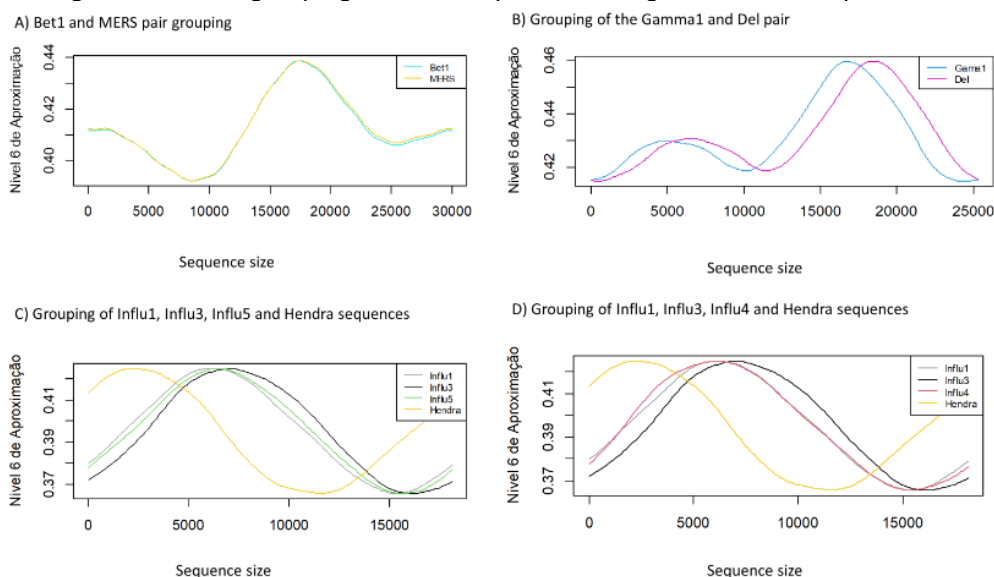
Figure 3: Groupings by other methods



Source: The authors.

The differentiated aggregate variance method presented the estimate of $H > 0.5$ at levels 5 and 6, while at levels 1, 2, 3 and 4, $H < 0.5$, the aggregated variance and aggregated absolute value methods presented the estimates of $H > 0.5$ at level 6, while at levels 1, 2, 3 and 4 (except for the Gamma1 variant) the $H < 0.5$. Finally, the graphs of the last level of approximation of the groups that were similar throughout the application of the techniques were verified, as shown in Figure 4.

Figure 4: Similar groupings that are repeated throughout the techniques used



Source: The authors.

4 DISCUSSION

Figure 1 shows the decomposition process that was carried out at each of the levels where the non-decimal discrete wavelet transform was applied. The Hurst exponent as a measure to verify similar genomic sequences has been previously addressed by Ferreira et al. (2020), in which they apply the 5 techniques discussed here to assess similarities in strains of the tuberculosis genome.

Figure 2 shows the formation of 3 clusters, after the implementation of the method of absolute moments, and in these clusters, we must highlight some patterns that are also repeated in other methodologies. The Bet1 and MERS pair proved to be similar in almost all methods, except for the Peng method. Furthermore, previous studies indicate that the phylogenetic analysis of SARS-CoV-2 belongs to the genus Betacoronavirus that includes the SARS bat types, SARS-Cov and MERS-SAR (Petrosillo et al. 2020). When analyzing the similarities in sequences of the Coronaviridae and Paramyxoviridae families, a high similarity was observed between the Bet1 and MERS pair, and from figure 4 A) the graph of the Bet1 and MERS sequences almost overlap.

The other pair of sequences that proved to be similar by the absolute moment methods and by the aggregated variance method was the pair Gama1 and Del. Previous studies indicate that the Gamma variant is immune to the Coronavac vaccine, therefore, this strain is immune to neutralizing antibodies generated in order to respond to polyclonal stimulation against the variants that circulated before the Gamma variant (Souza et al. 2021), on the other hand, according to Bian et al. (2021), the Delta variant contains mutations L452R, T478K, D614G and P681R that may also affect resistance to specific antibodies and that, according to Hu et al. (2021) after two injections of Coronavac the neutralizing titer decreased 2.7-fold.

The group composed of Infl1, Infl4, Infl3 and Hendra was similar by the methods of absolute moments, aggregated variance, and in the R/S analysis method, there was a substitution of Infl4 by Infl5. Furthermore, according to Wong et al. (2016) Hendra virus entry into the paramyxovirus cell is mediated by the F fusion protein, in response to the binding of a host receptor by the binding protein. Therefore, the structure of Hendra virus clearly shows a very high

similarity with the structure of the parainfluenza virus5. On the other hand, the parainfluenza 5 pair of fusion proteins (F) together with the parainfluenza 3 and Newcastle disease virus binding protein pair behave according to the fusion activation association model (Wong et al. 2016), on the other hand, in Figure 4D) the Infla1 and Infla4 sequences almost overlap. In addition, the human parainfluenza virus types 1, 3 and 4 cause mild, severe, or prolonged infections in the respiratory tract, such as bronchiolitis and pneumonia.

However, Liu et al. (2013) developed studies to explore the epidemiological characteristics and clinical manifestations of parainfluenza virus variants in types 1, 2, 3 and 4 to establish clinical distinctions, and found that the only significant difference in clinical presentation between types of parainfluenza was sputum, and that the most frequent types in patients were Infla 3 and Infla1, this theory is confirmed in research by Denny and Clyde (1986), Glezen et al. (1984) and Murphy et al. (1980), and that the Infla 1 and Infla 3 types present many similar clinical features that somehow differ from parainfluenza 4 (Infla 4). Therefore, one of the aspects to be considered in this research is the fact that it was found that the genetic sequencing of the Infla1 virus is like the genetic sequencing of Infla4, a fact that had not been mentioned in previous research.

5 CONCLUSIONS

The non-decimated Daubechies discrete wavelet transform allowed the decomposition to be performed on each of the GC sequences. Furthermore, when the Hurst exponent estimation techniques were implemented, patterns with greater similarity were repeated throughout the applied methodologies, and the absolute moments and aggregate variance methods presented similar clusters compared to the other implemented techniques. On the other hand, the two virus families, despite having different ancestors, present some similarity in the strains because the methodology used was able to distinguish variants of Coronaviridae and Paramyxoviridae with similarities, which is the case of Infla5, which is identical to the Bet1 and MERS pair, and to the pair Gama1 and Del at least 30%. Not all strains mixed with other variants from other families, in the case of Infla1, it remained separate from Coronaviridae variants in all techniques, apart from the Peng method.

The absolute moment, aggregated variance and R/S analysis methods managed to separate the genomes of the families under study so that not all strains were mixed with other variants from other families, and the Peng method, in turn, did not show satisfactory results. The study shows that some sequences remained more isolated from the others, which is the case of Influenza that remained separated from the variants of Coronaviridae in all the techniques, except in the Peng method. Furthermore, the methods of absolute moment, aggregated variance and R/S analysis showed better performance than the other applied methodologies.

REFERENCES

ABREU, L. D. and SPECKBACHER, M. Affine density, von Neumann dimension and a problem of Perelomov. *Advances in Mathematics*. 2022, 407, 108564. <https://doi.org/10.1016/j.aim.2022.108564>

ADLER, R., FELDMAN, R. and TAQQU, M. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Springer Science and Business Media. 1998, pp. 534.

ALDEA, R. and TARNICERIU, D. G. Estimating the hurst exponent in motor imagery-based brain computer interface. *7th Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013, Cluj-Napoca, Romania*. 2013, 1–6. <https://doi.org/10.1109/SpeD.2013.6682656>

ALSHAMMARI, T., et al. Forecasting Stock Volatility Using Wavelet-based Exponential Generalized Autoregressive Conditional Heteroscedasticity Methods. *Intelligent Automation & Soft Computing*. 2023, 35(3). <https://doi.org/10.32604/iasc.2023.024001>

AROUXET, M. B., et al. Covid-19 impact on cryptocurrencies: Evidence from a wavelet-based Hurst exponent. *Physica A: Statistical Mechanics and Its Applications*. 2022, 596, 127170. <https://doi.org/10.1016/j.physa.2022.127170>

ASSAF, A., et al. Multivariate long memory structure in the cryptocurrency market: The impact of COVID-19. *International Review of Financial Analysis*. 2022, 82, 102132. <https://doi.org/10.1016/j.irfa.2022.102132>

BĂRBULESCU, A., SERBAN and MAFTEI, C. Evaluation of Hurst exponent for precipitation time series. Latest trends on computers: Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference – Volume II. 2010, 2, 7.

BIAN, L., et al. Impact of the Delta variant on vaccine efficacy and response strategies. *Expert Review of Vaccines*. 2021, 1. <https://doi.org/10.1080/14760584.2021.1976153>

BRASSAROTE, G. De O. N., et al. Análise multiescala a partir de wavelets não decimadas: Investigação dos efeitos da cintilação ionosférica nos sinais GPS. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*. 2015, 3(1). <https://doi.org/10.5540/03.2015.003.01.0476>

CHOULJENKO, V. N., et al. Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. *The Journal of General Virology*. 2001, 82(12), 2927–2933. <https://doi.org/10.1099/0022-1317-82-12-2927>

DAS, S. and KUMAR, A. Long-term dependency between sovereign bonds and sectoral indices of India: Evidence using Hurst exponent and wavelet analysis. *Managerial Finance*. 2021, 47(10), 1448–1464. <https://doi.org/10.1108/MF-12-2020-0596>

De Groot, R. J., et al. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group. *Journal of Virology*. 2013, 87(14), 7790–7792. <https://doi.org/10.1128/JVI.01244-13>

DENAULT, W. R. P., et al. A fast wavelet-based functional association analysis replicates several susceptibility loci for birth weight in a Norwegian population. *BMC Genomics*. 2021, 22(1), 321. <https://doi.org/10.1186/s12864-021-07582-6>

DENNY, F. W. and CLYDE, W. A. Acute lower respiratory tract infections in non-hospitalized children. *The Journal of Pediatrics*. 1986, 108, 635–646. [https://doi.org/10.1016/s0022-3476\(86\)81034-4](https://doi.org/10.1016/s0022-3476(86)81034-4)

FERREIRA, L. M., SÁFADI, T. and FERREIRA, J. L. Evaluation of genome similarities using a wavelet-domain approach. *Revista Da Sociedade Brasileira de Medicina Tropical*. 2020, 53, e20190470. <https://doi.org/10.1590/0037-8682-0470-2019>

FREI, F. 2006. *Introdução à análise de agrupamentos: Teoria e prática*. São Paulo: UNESP, pp. 112.

GARCIN, M. A comparison of maximum likelihood and absolute moments for the estimation of Hurst exponents in a stationary framework. *Communications in Nonlinear Science and Numerical Simulation*. 2022, 114, 106610. <https://doi.org/10.1016/j.cnsns.2022.106610>

GLEZEN, W. P., et al. Parainfluenza Virus Type 3: Seasonality and Risk of Infection and Reinfection in Young Children. *The Journal of Infectious Diseases*. 1984, 150(6), 851–857. <https://doi.org/10.1093/infdis/150.6.851>

GOMAA, M. H., et al. Complete genomic sequence of turkey coronavirus. *Virus Research*. 2008, 135(2), 237–246. <https://doi.org/10.1016/j.virusres.2008.03.020>

HE, R., et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochemical and Biophysical Research Communications*. 2004, 316(2), 476–483. <https://doi.org/10.1016/j.bbrc.2004.02.074>

HU, J., et al. Reduced neutralization of SARS-CoV-2 B.1.617 variant by inactivated and RBD-subunit vaccine. *bioRxiv*. 2021. <https://doi.org/10.1101/2021.07.09.451732>

KOMADA, H., et al. Completion of the full-length genome sequence of human parainfluenza virus types 4A and 4B: Sequence analysis of the large protein

genes and gene start, intergenic and end sequences. *Archives of Virology*. 2011, 156(1), 161–166. <https://doi.org/10.1007/s00705-010-0834-6>

LARA-MUSULE, A., et al. Diagnosis and Monitoring of Volatile Fatty Acids Production from Raw Cheese Whey by Multiscale Time-Series Analysis. *Applied Sciences*. 2021, 11(13). <https://doi.org/10.3390/app11135803>

LI, G. and GUO, Y. Exploring the Medication Pattern of Chinese Medicine for Peptic Ulcer Based on Data Mining. *Journal of Healthcare Engineering*. 2021, 9072172. <https://doi.org/10.1155/2021/9072172>

LIU, W.-K., et al. Epidemiology and clinical presentation of the four human parainfluenza virus types. *BMC Infectious Diseases*. 2013, 13, 28. <https://doi.org/10.1186/1471-2334-13-28>

MANLY, B. F. J. *Métodos estatísticos multivariados: Uma introdução*. 3th ed. Bookman, 2008.

MARTHALER, D., et al. Complete Genome Sequence of Strain SDCV/USA/Illinois121/2014, a Porcine Deltacoronavirus from the United States. *Genome Announcements*. 2014, 2(2), e00218-14. <https://doi.org/10.1128/genomeA.00218-14>

MCLACHLAN, G. J. Mahalanobis distance. *Resonance*. 1999, 4(6), 20–26. <https://doi.org/10.1007/BF02834632>

MURPHY, B., et al. Seasonal Pattern of Childhood Viral Lower Respiratory Tract Infections in Melbourne. *Medical Journal of Australia*. 1980, 1(1), 22–24. <https://doi.org/10.5694/j.1326-5377.1980.tb134568.x>

NASTOS, C. V. and SARAVANOS, D. A. Multiresolution Daubechies finite wavelet domain method for transient dynamic wave analysis in elastic solids. *International Journal for Numerical Methods in Engineering*. 2021, 122(23), 7078–7100. <https://doi.org/10.1002/nme.6822>

NEWMAN, J. T., et al. Sequence analysis of the Washington/1964 strain of human parainfluenza virus type 1 (HPIV1) and recovery and characterization of wild-type recombinant HPIV1 produced by reverse genetics. *Virus Genes*. 2002, 24(1), 77–92. <https://doi.org/10.1023/a:1014042221888>

PETROSILLO, N., et al. COVID-19, SARS and MERS: Are they closely related? *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*. 2020, 26(6), 729–734. <https://doi.org/10.1016/j.cmi.2020.03.026>

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: Vienna: R Foundation for Statistical Computing, 2017. Available from: <http://www.R-project.org/>

RASHID, O., AMIN, A. and LONE, M. R. Performance Analysis of DWT Families. *3rd International Conference on Intelligent Sustainable Systems (ICISS)*. 2020, 1457–1463. <https://doi.org/10.1109/ICISS49785.2020.9315960>

ROUT, R. K., et al. Feature-extraction and analysis based on spatial distribution of amino acids for SARS-CoV-2 Protein sequences. *Computers in Biology and Medicine*. 2022, 141, 105024. <https://doi.org/10.1016/j.compbio-med.2021.105024>

SABIR, J. S. M., et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science Express*. 2016, 351(6268), 81–84. <https://doi.org/10.1126/science.aac8608>

SAINI, S. and DEWAN, L. Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *SpringerPlus*. 2016, 5(1), 64. <https://doi.org/10.1186/s40064-016-1668-9>

SOTERRONI, A. C., DOMINGUES, M. O. and RAMOS, F. M. Estimativa do expoente de hurst de séries temporais caóticas por meio da transformada wavelet discreta. In: *Proceedings of the Thematic Congress on Dynamics, Control and Applications-DINCON, Presidente Prudente, Brazil*. 2008, 7-9.

SOUZA, W. M., et al. Neutralisation of SARS-CoV-2 lineage P.1 by antibodies elicited through natural SARS-CoV-2 infection or vaccination with an inactivated SARS-CoV-2 vaccine: An immunological study. *The Lancet Microbe*. 2021, 2(10), e527–e535. [https://doi.org/10.1016/S2666-5247\(21\)00129-4](https://doi.org/10.1016/S2666-5247(21)00129-4)

TEVEROVSKY, V., & TAQQU, M. Testing for long-range dependence in the presence of shifting means or a slowly declining trend, using a variance-type estimator. *Journal of Time Series Analysis*. 1997, 18(3), 279–304. <https://doi.org/10.1111/1467-9892.00050>

VAN BOHEEMEN, S., et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio*. 2012, 3(6), e00473-12. <https://doi.org/10.1128/mBio.00473-12>

VAN DER HOEK, L., et al. Identification of a new human coronavirus. *Nature Medicine*. 2004, 10(4), 368–373. <https://doi.org/10.1038/nm1024>

WANG, L. F., et al. The exceptionally large genome of Hendra virus: Support for creation of a new genus within the family Paramyxoviridae. *Journal of Virology*. 2000, 74(21), 9972–9979. <https://doi.org/10.1128/jvi.74.21.9972-9979.2000>

WANG, X., et al. The UDWT image denoising method based on the PDE model of a convexity-preserving diffusion function. *EURASIP Journal on Image and Video Processing*. 2019, (1), 81. <https://doi.org/10.1186/s13640-019-0480-1>

WONG, J. J. W., et al. Structure and stabilization of the Hendra virus F glycoprotein in its prefusion form. *Proceedings of the National Academy of Sciences*. 2016, 113(4), 1056–1061. <https://doi.org/10.1073/pnas.1523303113>

YANG, J. and DELPHA, C. An incipient fault diagnosis methodology using local Mahalanobis distance: Fault isolation and fault severity estimation. *Signal Processing*. 2022, 200, 108657. <https://doi.org/10.1016/j.sigpro.2022.108657>

YU, Y., et al. A Two-Stage Wavelet Decomposition Method for Instantaneous Power Quality Indices Estimation Considering Interharmonics and Transient Disturbances. *IEEE Transactions on Instrumentation and Measurement*. 2021, 70, 1–13. <https://doi.org/10.1109/TIM.2021.3052554>

ZHUANG, Q.-Y., et al. Genomic Analysis and Surveillance of the Coronavirus Dominant in Ducks in China. *PloS One*, 2015, 10(6), e0129256. <https://doi.org/10.1371/journal.pone.0129256>