Computational analysis of the SARS-CoV-2 and other viruses based on the Kolmogorov’s complexity and Shannon’s information theories

J. A. Tenreiro Machado · João M. Rocha-Neves · José P. Andrade

Received: 26 April 2020 / Accepted: 14 June 2020 / Published online: 4 July 2020
© Springer Nature B.V. 2020

Abstract This paper tackles the information of 133 RNA viruses available in public databases under the light of several mathematical and computational tools. First, the formal concepts of distance metrics, Kolmogorov complexity and Shannon information are recalled. Second, the computational tools available presently for tackling and visualizing patterns embedded in datasets, such as the hierarchical clustering and the multidimensional scaling, are discussed. The synergies of the common application of the mathematical and computational resources are then used for exploring the RNA data, cross-evaluating the normalized compression distance, entropy and Jensen–Shannon divergence, versus representations in two and three dimensions. The results of these different perspectives give extra light in what concerns the relations between the distinct RNA viruses.

Keywords COVID-19 · Kolmogorov complexity theory · Shannon information theory · Hierarchical clustering · Multidimensional scaling

1 Introduction

In December 2019, a mysterious pneumonia with unknown etiology was reported in the city of Wuhan, Province of Hubei, China [1]. The International Committee on Taxonomy of Virus (ICTV) named the virus as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2]. Globally, to the 30th of July 2020, according to the World Health Organization Coronavirus Disease 2019 (COVID-19) Situation Report, 10,185,374 confirmed cases of COVID-19 are reported, resulting in 503,862 deaths.

The scientific community reacted as never before, and many researchers focused on this urgent topic [3–10]. The mathematical and computer science communities are also studying this challenging problem, and we testimony the recent emergence of new models and algorithmic approaches. This common multidisciplinary research will allow the human kind
to implement a robust and fast response to a problem that is causing a severe global socioeconomic disruption, and most probably will lead to the largest global recession since the Great Depression.

The present paper follows this trend by studying the genetic information by means of mathematical and computational tools. The starting point is the information encoded in the ribonucleic acid (RNA), a polymeric molecule essential in various biological roles.

The evolutionary origin and divergence of eucaryotes is mostly recoverable from their genetic relationships. The phylogeny of core genes, such as those for ribosomal proteins, provides a reasonable representation of many billions of years [11]. Unfortunately, the diversity of viruses prevents such a reconstruction of virus evolutionary histories as they lack any equivalent set of universally conserved genes on which to construct a phylogeny. Viral diversity is far greater than that of other organisms, with significant differences in their genetic material, RNA or deoxyribonucleic acid (DNA), and configurations (double or single stranded), as well as the orientation of their encoded gene. The smallest virus genomes [12] contain over 2,500 genes [13]. The RNA is a nucleic acid that is essential to all forms of life, and it is found in nature generally as single-stranded (ss) rather than a paired double-stranded (ds) like DNA. In DNA, there are four bases: adenine (A) complementary to thymine (T) and cytosine (C) complementary to guanine (G). In the RNA, uracil (U) is used instead of thymine.

Like DNA, RNA can carry genetic information. An RNA virus is a virus that has RNA as its genetic material encoding a number of proteins. This nucleic acid is usually single-stranded RNA (ssRNA), but there are double-stranded RNA (dsRNA) viruses. These viruses exploit the presence of RNA-dependent RNA polymerases in eucaryotes cells for replication of their genomes. Many human diseases are caused by such RNA viruses such as influenza, severe acute respiratory syndrome (SARS), COVID-19, Ebola disease virus, chikungunya, Zika virus, influenza B and Lassa virus [14].

The RNA is usually sequenced indirectly by copying it into complementary DNA (cDNA), which is often amplified and subsequently analyzed using a variety of DNA sequencing methods. Therefore, the genomic sequences of the RNA viruses are published presenting the four bases, namely adenine (A), cytosine (C), guanine (G) and thymine (T).

This genetic information is analyzed by means of the Kolmogorov’s complexity and Shannon’s information theories. In the first case, the so-called normalized information distance (NID) is adopted. In the second, a statistical approach is considered, by constructing histograms for the relative frequency of three consecutive bases (triplets). The histograms are interpreted under the light of entropy, cumulative residual entropy and Jensen–Shannon divergence. The results obtained for each theory, that is, the values assessing the virus genetic code under the light of the Kolmogorov’s complexity and the Shannon’s information, are further processed by means of advanced computational representation techniques. The final visualization is obtained using the hierarchical clustering (HC) [15–20] and multidimensional scaling (MDS) techniques [21–27]. Three alternative representations, namely dendrograms, hierarchical trees and three-dimensional loci, are considered.

According to the ICTV, human coronavirus belongs to the Betacoronavirus genus, a member of the Coronaviridae family, categorized in the order Nidovirales [28]. It has been categorized into several genera, based on phylogenetic analyses and antigenic criteria, namely: (i) Alphacoronavirus, responsible for gastrointestinal disorders in humans, dogs, pigs and cats; (ii) Betacoronavirus, including the bat coronavirus, the human severe acute respiratory syndrome (SARS) virus, the Middle Eastern respiratory syndrome (MERS) virus and now the SARS-CoV-2; (iii) Gammacoronavirus, which infects avian species; and (iv) Deltacoronavirus [2, 29].

Four coronaviruses broadly distributed among humans (229E, OC43, NL63 and, HKU1) frequently cause only common cold symptoms [2, 12]. The two other strains of coronaviruses are linked with deadly diseases and zoonotic in origin, i.e., the SARS-CoV and MERS-CoV [30]. In 2002–2003, there was an outbreak of SARS beginning in Guangdong Province in China and affecting 27 countries subsequently [31]. It was considered the first pandemic event of the twenty-first century, due to the SARS-CoV infecting 8098 individuals with 774 deaths [31]. In 2012, in the Middle East, MERS-CoV caused a severe respiratory disease that affected 2494 individuals and 858 deaths [2]. In both epidemics, bats were the original host of these two coronaviruses [2].
Coronaviruses contain a positive-sense, single-stranded RNA genome. The genome size for coronaviruses ranges from 26.4 to 31.7 kilobases, and it is one of the largest among RNA viruses.

The rapid spread of SARS-CoV-2 raises intriguing questions such as whether its evolution is driven by mutations, recombination or genetic variation [32]. This information is now being applied to the development of diagnostic tools, effective antiviral therapies and in the understanding of viral diseases. Although numerous studies have been done from the biological and medical perspective, to help further understand SARS-CoV-2 and trace its origin, this paper reports the use of multidimensional scale techniques for the finding of the similarities and the relationships among COVID-19 strains themselves and between other described viruses.

For that, we have collected a set of 37 complete genome sequences of SARS-CoV-2 virus obtained in several countries from patients with COVID-19. To help verify whether it is possible to trace the original or intermediate host of SARS-CoV-2, we obtained the genomic sequences of SARS-CoV-2 virus in other hosts, including bats (16 genomic sequences), pangolins (8 genomic sequences) and environment (market of Wuhan) [13]. We have also collected 23 genomic sequences of the coronavirus that cause mild symptoms related to common cold in man (229E, OC43, NL63 and, HKU1). The genomic sequences of SARS-CoV (10 genomic sequences) and a MERS-CoV (13 genomic sequences) were also gathered. For comparison and control, we also obtained sequences of other deadly pathogenic RNA viruses such as Lassa (6 genomic sequences), Ebola (6 genomic sequences), dengue (7 genomic sequences), chikungunya (1 genomic sequence) and influenza B (2 genomic sequences).

The diagram of Fig. 1 summarizes the main historical flow of coronavirus in what concerns the twenty-first-century epidemics.

To the authors’ best knowledge, this paper analyzes for the first time a large number of viruses associated with the combination of several distinct mathematical and computational tools. In short, we have the cross-comparison of:

- The genomic sequences of 133 viruses.
- The data treatment by means of Kolmogorov’s complexity and Shannon’s information theories, using normalized compression distance, entropy, cumulative residual entropy and Jensen–Shannon divergence.
- Two clustering computational techniques, namely the hierarchical clustering and multidimensional scaling.
- The visualization of the results by means of dendrograms, trees and point loci.

The results indicate clearly the superior performance of the approaches based on the Kolmogorov’s complexity measure and the MDS three-dimensional visualization. Moreover, the characteristics of the coronaviruses within the large set of tested cases are highlighted. We conclude that:

- The association between the Kolmogorov perspective and the three-dimensional MDS representation leads to be the best results.
- The clusters are easily distinguishable and we observe the relation between the new SARS-CoV-2 virus and some CoV found in bats and in the pangolin.
Bearing these ideas in mind, the paper is organized as follows. Section 2 introduces the mathematical and computational tools adopted in the follow-up. Section 3 analyzes the data by means of the complexity and information theories accompanied by the HC and MDS computational visualization resources. Finally, Sect. 4 summarizes the main conclusions.

2 Fundamental tools

2.1 Distance metrics

Evaluating the similarity degree between several objects, each having a number of features requires the definition of a distance. The calculation of a function \( d \) of two objects \( x_A \) and \( x_B \) can be interpreted as a distance if \( d(x_A, x_B) \geq 0 \) and satisfies the following axioms [33]:

- identity: \( d(x_A, x_B) = 0 \) if \( x_A = x_B \)
- symmetry: \( d(x_A, x_B) = d(x_B, x_A) \)
- triangle inequality: \( d(x_A, x_B) \leq d(x_A, x_C) + d(x_C, x_A) \).

We find in the literature a variety of different functions that can tackle datasets and shed light to distinct characteristics [34]. For a set of numerical vectors \( (x_1, \ldots, x_N)^T \in \mathbb{R}^n \), the Minkowski norm \( L_p : (\sum_{k=1}^{N}|x_k|^p)^{\frac{1}{p}} \), and in particular the Manhattan and Euclidean cases, \( L_1 \) and \( L_2 \) for \( n = 1 \) and \( n = 2 \), are often used [35]. In the case of DNA analysis, these norms support different algorithms [36–38] that allow the comparison of data sequences. We can also mention metrics such as the Chi-square [39], Hamming [40] and edit [41] distances. Nonetheless, the selection of the optimal distances for a specific application poses relevant challenges [42–45]. In fact, the adoption of a given metric often depends on the user experience that performs several numerical experiments before selecting one or more distances. Due to these issues, assessing the similarity of several objects is not a straightforward process and we can adopt both non-probabilistic and probabilistic information measures to obtain distinct perspectives between object similarities.

2.2 Kolmogorov complexity theory

The Kolmogorov complexity, or algorithmic entropy, addresses the information measurement without relying on probabilistic assumptions. The information measurement focuses an individual finite object described by a string and takes into consideration that ‘regular’ strings are compressible [46]. The Kolmogorov complexity of a string \( x \), denoted as \( K(x) \), can be defined as the length of the shortest binary program that, given an empty string \( \psi \) at its input, can compute \( x \) on its output in a universal computer, and then halts. Therefore, \( K(x) \) can be interpreted as the length of the ultimate compressed version of \( x \).

The information distance of two strings (or files) \( \{x_A, x_B\} \in \Sigma \), where \( \Sigma \) denotes the space of the objects, can also be computed by means of the conditional Kolmogorov complexity \( K(x_A | x_B) \) [47, 48]. This concept can be read as the length of the shortest program to obtain \( x_A \), if \( x_B \) is provided for input. In heuristic terms, if the two strings are more/less similar, then the calculation is less/more difficult and, consequently, the size of the program is smaller/larger. Therefore, the following inequality always holds

\[
K(x_A | x_B) \leq K(x_A).
\]

Under the light of these concepts, the universal distance metric [47] denoted as normalized information distance (NID):

\[
\text{NID}(x_A, x_B) = \frac{\max \{K(x_A | x_B), K(x_B | x_A)\}}{\max \{K(x_A), K(x_B)\}}, \quad \{x_A, x_B\} \in \Sigma,
\]

was proposed.

From equation (2), we have that the NID may take values in the range \([0, 1]\). Moreover, we have \( \text{NID}(x_A, x_A) \approx 0 \) and \( \text{NID}(x_A, \psi) \approx 1 \), where \( \psi \) is an empty object that has no similarity to \( x_A \). It is shown [47] that the NID is a distance because it satisfies the axioms defined in (1), up to some additive precision, but is non-computable [47]. In spite of this limitation, the NID gives the basis for the so-called normalized compression distance (NCD), which is a computable distance [45]. The computability comes with the cost of using a good approximation of the Kolmogorov complexity by a standard compressor.
Information theory [56] has been applied in a variety of scientific areas. The most fundamental piece of the theory is the information content $I$ of a given event $x_i$ having probability of occurrence $P(X = x_i)$:

\[
I(X = x_i) = - \ln(P(X = x_i)),
\]

where $X$ is a discrete random variable.

The expected value of the information, the so-called Shannon entropy [57, 58], is given by:

\[
H(X) = E(- \ln(X)) = - \sum_i P(X = x_i) \ln(P(X = x_i)),
\]

where $E(\cdot)$ denotes the expected value operator.

Expression (6) obeys the four Khinchin axioms [59, 60]. Several generalizations of Shannon entropy have been proposed, relaxing some of those axioms, and we can mention the Rényi, Tsallis and generalized entropy [61–63], just to name a few.

A recent and interesting concept is the cumulative residual entropies $\varepsilon$ given by [64, 65]:

\[
\varepsilon(X) = - \sum_i P(X > x_i) \log(P(X > x_i)).
\]

Within the scope of information theory, we can formulate also the concept of distance discussed in Sect. 2.1. The Kullback–Leibler divergence between the probability distributions $X_1$ and $X_2$ is defined as [34, 66–69]:

\[
D_{KL}(X_1 \parallel X_2) = \sum_i P(X_1 = x_i) \ln \left( \frac{P(X_1 = x_i)}{P(X_2 = x_i)} \right).
\]

From this, we obtain the Jensen–Shannon divergence, $JSD(X_1 \parallel X_2)$, or distance, given by:

\[
JSD(X_1 \parallel X_2) = \frac{1}{2} [D_{KL}(X_1 \parallel X_{12}) + D_{KL}(X_2 \parallel X_{12})],
\]

where $X_{12} = \frac{1}{2}(X_1 + X_2)$.

Alternatively, the $JSD(X_1 \parallel X_2)$ can be calculated as:

\[
JSD(X_1 \parallel X_2) = \frac{1}{2} \sum_i P(X_1 = x_i) \ln(P(X_1 = x_i)) + \frac{1}{2} \sum_i P(X_2 = x_i) \ln(P(X_2 = x_i)) - \sum_i P(X_{12} = x_i) \ln(P(X_{12} = x_i)).
\]
2.4 Hierarchical clustering, multidimensional scaling and visualization

The HC is a computational technique that assesses a group of \( N \) objects \( X_i, i = 1, \cdots, N \), in a \( q \)-dimensional space, and tries to rearrange them in a visual structure with objects \( Y_i \) highlighting the main similarities between them in the sense of some predefined metric [70].

Let us consider \( N \) objects defined in a \( q \)-dimensional real-valued space and a distance (or dissimilarity) measure \( \delta_{ij} \) between pairs of objects. The first step is to calculate a \( N \times N \)-dimensional matrix, \( \Delta = [\delta_{ij}] \), where \( \delta_{ii} \in \mathbb{R}^+ \) for \( i \neq j \) and \( \delta_{ii} = 0 \), \( (i, j) = 1, \cdots, N \), stands for the object to object distances [71]. The HC uses the input information in matrix \( \Delta \) and produces a graphical representation consisting in a dendrogram or a hierarchical tree.

The so-called agglomerative or divisive clustering iterative techniques are usually adopted for processing the information. In the first approach, each object starts in its own cluster and the computational iterations merge the most similar items until having just one cluster. In the second, all objects start their own cluster and the computational iterations separate items, until each object has his own cluster. For both approaches, the numerical iterations follow a linkage criterion, based on the distances between pairs, for quantifying the dissimilarity between clusters. The maximum, minimum and average linkages are possible criteria. The distance \( d(x_A, x_B) \) between two objects \( x_A \in A \) and \( x_B \in B \) can be assessed by means of several metrics such as the average linkage given by [72]:

\[
 d_{av}(A, B) = \frac{1}{\|A\|\|B\|} \sum_{x_A \in A, x_B \in B} d(x_A, x_B). \tag{11}
\]

The clustering quality can be assessed by means of the cophenetic correlation [73]:

\[
 c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(y(i,j) - \bar{y})}{\sqrt{\sum_{i<j}(x(i,j) - \bar{x})^2 \sum_{i<j}(y(i,j) - \bar{y})^2}}, \tag{12}
\]

where \( x(i, j) \) and \( y(i, j) \) stand for the distances between the pairs \( X_i \) and \( X_j \), in the initial measurements, and \( Y_i \) and \( Y_j \), in the HC chart, respectively, and \( \bar{x} \) denotes the average of \( x \).

Values of \( c \) close to 1 (to 0) indicate a good (weak) cluster representation of the original data. In MATLAB, \( c \) is computed by means of the command cophenet.

In this paper, we adopt the agglomerative clustering and the average linkage [74, 75] for tackling the matrix of distances based on the \( JSD \) metric (10).

The MDS is also a computational technique for clustering and visualizing multidimensional data [76]. Similarly to what was said for the HC, the input of the MDS is the matrix \( D = [\delta_{ij}], \ (i, j) = 1, \cdots, N \). The MDS is to adopt points for representing the objects in a \( d \)-dimensional space, with \( d < q \), that try to reproduce the original distances, \( \delta_{ij} \). The MDS performs a series of numerical iterations rearranging the points in order to optimize a given cost function called stress \( S \). We have, for example, the residual sum of squares and the Sammon criteria:

\[
 S = \left[ \frac{\sum_{i<j}(\phi_{ij} - \delta_{ij})^2}{\sum_{i<j}^2} \right]^{\frac{1}{2}}, \tag{13a}
\]

\[
 S = \left[ \frac{\sum_{i<j}(\phi_{ij} - \delta_{ij})^2}{\sum_{i<j}^2 \phi_{ij}^2} \right]^{\frac{1}{2}}. \tag{13b}
\]

The resulting MDS points have coordinates that produce a symmetric matrix \( \Phi = [\phi_{ij}] \) of distances that approximate the original one \( A = [\delta_{ij}] \). In MATLAB, the commands cmdscale and Sammon can be adopted for the classical MDS and the Sammon stress criterion.

The interpretation of the MDS locus is based on the patterns of points [77, 78]. Therefore, objects with strong (weak) similarities are represented by fairly close (distant) points. The MDS locus is interpreted on the relative positions of the point coordinates. So, the absolute position of the points or the shape of the clusters has usually a special meaning, and we can magnify, translate and rotate the locus for achieving a good visualization. In the same line of reasoning, the axes of the plot do not have units or physical meaning. The quality of the produced MDS locus can be evaluated using the stress plot and the Shepard diagram. The stress plot shows \( S \) versus \( d \) and decreases monotonically. Usually, low values of \( d \) are adopted, and present computational resources.
allow a direct three-dimensional representation, but often some rotations and magnification are required to achieve the best visual perspective. The Shepard diagram plots \( \phi_{ij} \) against \( \delta_{ij} \) for a given value of \( d \), and a narrow (large) scatter of points indicates a good (weak) fit between the original and the reproduced distances.

3 The dataset

The information of 133 publicly released genomic sequences was collected in the Global Initiative on Sharing Avian Influenza Data (GISAID) and the GenBank of the National Center for Biotechnology Information (NCBI) databases (https://www.gisaid.org/, https://www.ncbi.nlm.nih.gov/genbank). The information regarding the sequences and serial numbers is given in Table of “Appendix.”

The mathematical tools of Kolmogorov and Shannon theories are used to compare and extract relationships among the data and to identify viral genomic patterns. The viral genomes are analyzed from the perspective of dynamical systems using HC and MDS. Dendrograms and trees are generated by HC algorithms, and a three-dimensional visualization through MDS visualization. Several clusters with medical and epidemiological interest were visualized. The MDS loci provide a key visualization of the relation of SARS-CoV-2 with the other known coronavirus affecting humans.

The several non-coronavirus pathogenic viruses analyzed are very well delimited in several independent and easily delineated clusters (Zika, Chikungunya, Dengue, Lassa, Ebola, Influenza B).

The several types of coronaviruses that cause mild disease (common flu) are very well separated from the other coronavirus clusters. Interestingly, the types, 229E, HkU1, NL63 and OC43, form separated mini-clusters within a big well-defined cluster.

There are two well-defined CoV-19 or SARS-CoV clusters. The two environmental SARS-CoV-2 were aggregated with the human SARS-CoV-2. To differentiate the human MERS-CoV, a zoom was made to isolate it forming a separate cluster. The MERS-CoV obtained from camels was included in this cluster. SARS-CoV formed another independent and well-segregated cluster. The CoV from bats and pangolin is dispersed among these several clusters of coronavirus forming sometimes clusters of few elements. Note that the bat CoV RaTG13 is near to one of the SARS-CoV clusters. This coronavirus is the closest known relative of SARS-CoV-2 [79].

For processing the RNA information consisting of ASCII files with the four nitrogenous bases, we consider two approaches. The first one follows the Kolmogorov complexity theory described in Sect. 2.2. Therefore, we consider the compressors zlib and bzip2 (see https://www.zlib.net and https://sourceware.org/bzip2/) followed by the NCD distance (4). Nonetheless, the two algorithms give almost identical results, and therefore, only the zlib is considered in the follow-up.

The application of the NCD produces a symmetrical matrix \( A, 133 \times 133 \) dimensional that is tackled and visualized by means of a dendrogram and a HC tree obtained by the program Phylip http://evolution.genetics.washington.edu/phylip.html and MDS using MATLAB https://www.mathworks.com/products/matlab.html. The corresponding charts are represented in Figs. 2, 3 and 4.

It is known that the three-dimensional plots require some rotation in the computer screen. Therefore, Fig. 5 shows the MDS locus in a different perspective, without point labels and the cluster of SARS-CoV-2 connected by a line.

The second approach follows the Shannon information theory described in Sect. 2.3. Therefore, we start by considering non-overlapping (codon or anticodon) triplets of bases and the histograms of relative frequency for each virus. In a second phase, after having the histogram for the complete set of virus, we process them using entropy concepts. Before continuing, we must clarify that we tested the adoption of n-tuples, with \( n = 1, 2, 3, 4 \), in the DNA information analysis of a large set of superior species such as mammals, where the genetic information is considerably larger and the construction of histograms for large values of \( n \) does not pose problems of statistical significance (since the number of histograms bins grows as \( 4^n \)). It was observed that \( n = 3 \) is a ‘good value,’ since \( n = 1 \) leads to poor results, \( n = 2 \) improves things but is still insufficient, while \( n = 4 \) gives almost identical results to \( n = 3 \).

In a first experiment, we consider the Shannon entropy \( H \) vs fractional cumulative residual entropy \( \varepsilon \) as possible descriptors of the 133 histograms. Figure 6
Fig. 2 Dendrogram for the set of 133 viruses using the NCD and zlib based on the Kolmogorov complexity theory
shows the resulting two-dimensional plot, that is, the Shannon entropy $H$ vs the fractional cumulative residual entropy $e$, for the set of 133 viruses. We observe some separation between virus, but we have some difficulties due to the high density of points in some areas.

We now try the other computational techniques introduced in Sect. 2. In the follow-up, we consider identical input information, that is, the $JDS$ and the corresponding matrix $A$ for the set of 133 viruses, and we compare the distinct computational clustering and visualization techniques.

Figures 7, 8 and 9 show the dendrogram, the HC tree and the three-dimensional MDS locus, respectively. The dendrogram is the simplest representation and it is straightforward to interpret. However, this technique does not take full advantage of the space. The HC tree uses more efficiently the two-dimensional space, but we now have more difficulties in reading the most dense clusters. The three-dimensional MDS takes advantage of the computational representation, but requires some rotation/shift/magnification in the computer screen.

Successive magnifications can be done and are necessary to achieve a more distinct visualization. Therefore, we can say that all representations have their own pros and cons, although the three-dimensional MDS is, a priori, the most efficient method.
Since the three-dimensional plot requires some rotation in the computer screen, Fig. 10 shows the MDS locus in a different perspective, without point labels and the cluster of SARS-CoV-2 connected by a line.

The MDS 3D following the Kolmogorov method isolated the several groups of virus in several extended groups. The cluster of SARS-CoV-2 is very extended. Note that the SARS-CoV-2 virus found in the market of Wuhan is very near of the SARS-CoV-2 cluster. The CoV found in the pangolin also forms a cluster mixing with the upper part of the SARS-CoV-2 cluster. Bat CoV virus forms also a diffuse cluster lacking some contiguity. The bat CoV Yunnan RaTG13 is very near the SARS-CoV-2 cluster. There is a big cluster formed by the different CoV that causes a mild respiratory disease (common cool).

Other pathogenic RNA virus clusters, i.e., Lassa virus, Zika, influenza B and Ebola, can be easily separated from the CoV clusters. However, with this methodology, the cluster formed by the dengue virus is not very far, in some views, from a part of the cluster formed by the CoV that causes mild respiratory disease.

The SARS CoV virus forms another independent cluster. MERS-CoV from humans forms another extended cluster and in the vicinity of the CoV virus from camel.
4 Conclusions

This paper explored the information of 133 RNA viruses available in public databases. For that purpose, several mathematical and computational tools were applied. On the one hand, the concepts of distance and the theories developed by Kolmogorov and Shannon for assessing complexity and information were recalled. The first involved the normalized compression distance and the zlib compressor for the DNA. The second understood the use of histograms of triplets of bases and their assessment through entropy, cumulative residual entropy and Jensen–Shannon divergence. On the other hand, the advantages of hierarchical clustering and multidimensional scaling algorithmic techniques were also discussed. Representations in two and three dimensions, namely by dendrograms and trees, and loci of points or points and lines, respectively, were compared. The results revealed their pros and cons for the specific application of the set of viruses under comparison.

The MDS 3D in the Kolmogorov perspective leads to be the best visualization method. We have not only the clusters easily distinguishable, but we find also the relation between the new SARS-CoV-2 virus and some CoV found in bats (the primary host of the virus) and in the pangolin, the likely intermediate host. The SARS-CoV-2 found in the environment, namely the Market of Wuhan where the epidemic probably started, is indistinguishable from the SARS-CoV-2 found in humans.

This type of methodology may help to study how an animal virus jumped the boundaries of species to infect humans, and pinpoint its origin, as this
knowledge can help to prevent future zoonotic events [80]. The statistical and computational techniques allow different perspectives over viral diseases that may be used to grasp the dynamics of the emerging COVID-19 disease. These methodologies [79] may help interpreting future viral outbreaks and to provide additional information concerning these infectious agents and understand the dynamics of viral diseases.

**Fig. 6** Shannon entropy $H$ vs fractional cumulative residual entropy $e$ for the set of 133 viruses, with the cluster of SARS-CoV-2 connected by a line.
Computational analysis of SARS-CoV-2 genome

Fig. 7 Dendrogram for the set of 133 viruses using the JSD based on the Shannon information theory
Fig. 8 HC tree for the set of 133 viruses using the JSD based on the Shannon information theory
Fig. 9  MDS three-dimensional locus for the set of 133 viruses using the JSD based on the Shannon information theory, with the cluster of SARS-CoV-2 connected by a line
The authors thank all those who have contributed and shared sequences to the GISAID database (https://www.gisaid.org/). The authors also thank those who have contributed to the GenBank of the National Center for Biotechnology Information (NCBI) databases (https://www.ncbi.nlm.nih.gov/genbank). The authors also thank Romulo Antão for the help in handling the information with the compressors zlib and bz2.

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.
## Appendix

| ![](image1.png) | ![](image2.png) |
|----------------|----------------|
| **1** | **2** | **3** | **4** |
| **Title** | **References** | **Details** | **Counts** |
| --- | --- | --- | --- |
| 1 | 123456 | Computational analysis of SARS-CoV-2 genome | 1747 |

### Notes
- The table above summarizes the computational analysis of the SARS-CoV-2 genome, highlighting key findings and references for further reading.
- This analysis is crucial for understanding the viral dynamics and the development of potential vaccines and treatments.

---

**References**

- 123456: "Computational analysis of SARS-CoV-2 genome" (2023).
- Additional references may be included in the full document for comprehensive analysis.

---

**Appendix**

#### 1. Introduction

The SARS-CoV-2 virus has become a significant global health concern, necessitating comprehensive computational analysis to understand its genetic makeup and potential for transmission.

#### 2. Methods

This analysis involves computational tools to study the viral genome, focusing on sequence alignment, mutation detection, and phylogenetic relationships.

#### 3. Results

- **Pathway 1:** Detailed analysis of specific gene expressions.
- **Pathway 2:** Comparative genomic studies with SARS-CoV.

#### 4. Conclusion

The findings contribute to a deeper understanding of the virus's adaptability and its potential impact on public health strategies.

---

**Supplementary Material**

- Additional data tables and figures are available in the full document for a thorough understanding of the computational methods and results.
References

1. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W.: A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382(8), 727–733 (2020). https://doi.org/10.1056/nejmoa2001017

2. ur Rehman, S., Shafique, L., Ihsan, A., Liu, Q.: Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. Pathogens 9(3), 240 (2020). https://doi.org/10.3390/pathogens9030240

3. Kandeil, A., Shehata, M.M., Shesheny, R.E., Gomaa, M.R., Ali, M.A., Kayali, G.: Complete genome sequence of middle east respiratory syndrome coronavirus isolated from a dromedary camel in Egypt. Genome Announc. (2016). https://doi.org/10.1128/genomea.00309-16

4. Kucharski, A.J., Russell, T.W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R.M., Sun, F., Jit, M., Munday, J.D., Davies, N., Gimma, A., van Zandvoort, K., Gibbs, H., Hellewell, J., Jarvis, C.I., Clifford, S., Guiry, B.J., Bosse, N.L., Abbott, S., Klepac, P., Flasche, S.: Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect. Dis. (2020). https://doi.org/10.1016/s1473-3099(20)30144-4

5. Lam, T.T.Y., Shum, M.H.H., Zhu, H.C., Tong, Y.G., Ni, X.B., Liao, Y.S., Wei, W., Cheung, W.Y.M., Li, W.J., Li, L.F., Leung, G.M., Holmes, E.C., Yu, Y.L., Guan, Y.: Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature (2020). https://doi.org/10.1038/s41586-020-2169-0

6. Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M.: Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science (2020). https://doi.org/10.1126/science.abb5793

7. Li, C., Yang, Y., Ren, L.: Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infect. Genet. Evol. 82, 104285 (2020). https://doi.org/10.1016/j.meegid.2020.104285

8. Peng, L., Fang, G., Liu, W.B., Kou, Z.: Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect. Dis. Poverty (2020). https://doi.org/10.1186/s40429-020-00649-8

9. Qiang, X.L., Xu, P., Fang, G., Liu, W.B., Kou, Z.: Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect. Dis. Poverty (2020). https://doi.org/10.1186/s40429-020-00649-8

10. Liu, Y., Liu, B., Cui, J., Wang, Z., Shen, Y., Xu, Y., Yao, K., Guan, Y.: COVID-19 evolves in human hosts (2020). https://doi.org/10.20944/preprints202003.0316.v1

11. Segata, N., Huttenhower, C.: Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS ONE 6(9), e24704 (2011). https://doi.org/10.1371/journal.pone.0024704

12. Al-Khannaq, M.N., Ng, K.T., Oong, X.Y., Pang, Y.K., Takebe, Y., Chook, J.B., Hanafi, N.S., Kamarulzaman, A., Tee, K.K.: Molecular epidemiology and evolutionary histories of human coronavirus OC43 and HKU1 among patients with upper respiratory tract infections in Kuala Lumpur, Malaysia. Virol. J. (2016). https://doi.org/10.1186/s12985-016-0488-4

13. Abergel, C., Legendre, M., Claverie, J.M.: The rapidly expanding universe of giant viruses: mimivirus, pandoravirus, pithovirus and mollivirus. FEMS Microbiol. Rev. 39(6), 779–796 (2015). https://doi.org/10.1093/femsre/fuv037

14. Acheson, N.H.: Fundamentals of Molecular Virology. Wiley, New York (2011)

15. Defays, D.: An efficient algorithm for a complete link method. Comput. J. 20(4), 364–366 (1977). https://doi.org/10.1093/comjnl/20.4.364

16. Székely, G.J., Rizzo, M.L.: Hierarchical clustering via joint between-within distances: extending Ward’s minimum variance method. J. Classif. 22(2), 151–183 (2005). https://doi.org/10.1007/s00357-005-0012-9

17. Fernández, A., Gómez, S.: Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. J. Classif. 25(1), 43–65 (2008). https://doi.org/10.1007/s10711-007-9040-4

18. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2nd edn. Springer, New York (2009)

19. Lopes, A.M., Machado, J.A.T.: Tidal analysis using time-frequency signal processing and information clustering. Entropy 19(8), 390 (2017). https://doi.org/10.3390/e19080390

20. Machado, J.A.T., Lopes, A.: Rare and extreme events: the case of COVID-19 pandemic. Nonlinear Dyn. (2020). https://doi.org/10.1007/s11071-020-05680-w

21. Torgerson, W.: Theory and Methods of Scaling. Wiley, New York (1958)

22. Shepard, R.N.: The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27(1 and II), 219–246 and 219–246 (1962)

23. Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29(1), 1–27 (1964)

24. Kruskal, J.B., Wish, M.: Multidimensional Scaling. Sage Publications, Newbury Park (1978)

25. Borg, I., Groenen, P.J.: Modern Multidimensional Scaling—Theory and Applications. Springer, New York (2005)

26. Ionescu, C., Machado, J.T., Keyser, R.D.: Is multidimensional scaling suitable for mapping the input respiratory impedance in subjects and patients? Comput. Methods Programs Biomed. 104(3), e189–e200 (2011)

27. Machado, J.A.T., Ding, E., Balleau, D.: Analysis of UV spectral bands using multidimensional scaling. SViP 9(3), 573–580 (2013). https://doi.org/10.1007/s11760-013-0485-7

28. Lai, M.M., Cavanagh, D.: The molecular biology of coronaviruses. In: Kielian, M., Mettenleiter, T., Roossink, M. (eds.) Advances in Virus Research, pp. 1–100. Elsevier, Amsterdam (1997). https://doi.org/10.1016/s0065-3527(08)60286-9

29. Schoeman, D., Fielding, B.C.: Coronavirus envelope protein: current knowledge. Virol. J. (2019). https://doi.org/10.1186/s12985-019-1182-0

30. Cui, J., Li, F., Shi, Z.L.: Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17(3), 181–192 (2018). https://doi.org/10.1038/s41579-018-0118-9

31. Lau, S.K.P., Woo, P.C.Y., Li, K.S.M., Huang, Y., Tsui, H.W., Wong, B.H.L., Wong, S.S.Y., Leung, S.Y., Chan, J. A. T. Machado et al.
65. Xiong, H., Shang, P., Zhang, Y.: Fractional cumulative residual entropy. Commun. Nonlinear Sci. Numer. Simul. 78, 104879 (2019). https://doi.org/10.1016/j.cnsns.2019.104879
66. Sibson, R.: Information radius. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 14(2), 149–160 (1969)
67. Taneja, I., Pardo, L., Morales, D., Ménandez, L.: Generalized information measures and their applications: a brief survey. Qüestiio 13(1–3), 47–73 (1989)
68. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37(1), 145–151 (1991). https://doi.org/10.1109/18.61115
69. Cha, S.H.: Measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. 1(4), 300–307 (2007)
70. Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975)
71. Tenreiro, J.A., Machado, A.M.L., Galhano, A.M.: Multidimensional scaling visualization using parametric similarity indices. Entropy 17(4), 1775–1794 (2015). https://doi.org/10.3390/e17041775
72. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space. Springer, New York (2001)
73. Sokal, R.R., Rohlf, F.J.: The comparison of dendrograms by objective methods. Taxon 10, 33–40 (1962). https://doi.org/10.2307/1217208
74. Felsenstein, J.: PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein (1993)
75. Tuimala, J.: A Primer to Phylogenetic Analysis Using the PHYLIP Package. CSC—Scientific Computing Ltd., Espoo (2006)
76. Saied, N., Haewoon, I.M., Saqib, D.B.M.: A survey on multidimensional scaling. ACM Comput. Surv. CSUR 51(3), 47 (2018). https://doi.org/10.1145/3178155
77. Machado, J.A.T.: Relativistic time effects in financial dynamics. Nonlinear Dyn. 75(4), 735–744 (2014). https://doi.org/10.1007/s11071-013-1100-8
78. Lopes, A.M., Andrade, J.P., Machado, J.T.: Multidimensional scaling analysis of virus diseases. Comput. Methods Programs Biomed. 131, 97–110 (2016). https://doi.org/10.1016/j.cmpb.2016.03.029
79. Cyranoski, D.: Profile of a killer: the complex biology powering the coronavirus pandemic. Nature 581(7806), 22–26 (2020). https://doi.org/10.1038/d41586-020-01315-7
80. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F.: The proximal origin of SARS-CoV-2. Nat. Med. 26(4), 450–452 (2020). https://doi.org/10.1038/s41591-020-0820-9

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.