Eur. Phys. J. Spec. Top. (2023) 232:1051–1060
https://doi.org/10.1140/epjs/s11734-023-00791-y

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

A comparative study on structural proteins of viruses that belong to the identical family

A. A. Navish[a] and R. Uthayakumar[b]

Department of Mathematics, The Gandhigram Rural Institute-Deemed to be University, Gandhigram, Dindigul 624 302, Tamil Nadu, India

Received 8 September 2022 / Accepted 25 January 2023 / Published online 17 February 2023

Abstract Recent studies have focused on the similarity between *SARS Cov*-2 and various viruses from the Coronaviridae family (such as *MERS Cov*, *SARS Cov* and *Bat Cov RaTG*13) to uncover the mystery of *SARS Cov*-2. Specifically, some studies identified that the *SARS Cov*-2 is closely related to *Bat Cov RaTG*13 (a *SARS*-related coronavirus found in bats) rather than the other viruses in that family. These studies are mainly focusing on the biological techniques to show the similarity between the *SARS Cov*-2 and other viruses. Examining proteins is not easy for common researchers unless for biologists. To rectify this flaw, we have to convert the protein to one of the known formats, which are easy to understand. Consequently, this study uses viral structural proteins to analyse the relationship between *SARS Cov*-2 and the rest of the coronavirus with the help of mathematical and statistical parameters and explores the various graph representations of *MERS Cov*, *SARS Cov*, *Bat Cov RaTG*13 and *SARS Cov*-2 structural proteins, such as zig–zag curve, Protein Contact Map ($\mathcal{PCM}$) and Chaos Game Representation ($\mathcal{CGR}$). Though these graph interpretations are visually similar, a slight variation between the graphs reflects their structural and functional differences. Thus, we use an elegant parameter known as the fractal dimension to observe their minor changes. According to the nature of the graph, we employ different types of fractal dimensions, namely mass dimension and box dimension. Furthermore, we perform the similarity tests with normalized cross-correlation and cosine similarity to assess the comparability of the $\mathcal{PCM}$ and $\mathcal{CGR}$ graphs. The acquired $\mathcal{CC}_n$ values are near the sequence identity between *SARS Cov*-2 and *MERS Cov*, *SARS Cov*, *Bat Cov RaTG*13.

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MERS Cov	Middle East respiratory syndrome
SARS Cov	Severe acute respiratory syndrome coronavirus
*Bat Cov RaTG*13	Bat coronavirus rhinolophus affinis TG13
SARS Cov-2	Severe acute respiratory syndrome coronavirus-2
NCBI	National Center for Biotechnology Information

1 Introduction

Though the *SARS Cov*-2 initially originated from the Wuhan market, there is still no confirmed proof. To find drugs against a particular disease, discovering the characteristics of the corresponding disease-causing factor is very important [1, 2].

Protein plays numerous roles in the biological world, such as transporting nutrients and constructing the structures of living things. It will perform its functions by interacting with other molecules such as DNA or RNA proteins and small molecules. So, by discovering and disturbing/destroying the particular protein, we can reduce/stop the entire disease spread process.

Amino acids are the monomers that encode proteins. The protein comprises twenty different amino acids and has a polypeptide backbone with attached side chains. The protein structures are determined by the sequence of amino acids that compose the proteins and how the protein folds into more complex shapes. Each amino acid has the same fundamental structure and comprises

R. Uthayakumar has contributed equally to this work.

Framework of Fractals in Data Analysis: Theory and Interpretation. Guest editors: Santo Banerjee, A. Gowrisankar.

[a] e-mail: aa.navish2@gmail.com (corresponding author)
[b] e-mail: uthayagri@gmail.com

a core carbon atom renowned as the alpha (α) carbon, which is attached to an amino group (NH2), a carboxyl group (COOH) and a hydrogen atom. Thus, it acts as the point of attachment for the side chains of 19 of the 20 amino acids and is vital in protein folding.

The complete structure of a protein can be described by four levels of complexity, namely primary, secondary, tertiary, and quaternary.

Primary	The sequence of amino acids in a polypeptide chain
Secondary	Indicates the local interactions of polypeptide chain stretches that can form -helices and -sheets via hydrogen bonding interactions
Tertiary	It is the overall three-dimensional arrangement of its polypeptide chain in space
Quaternary	This is formed by assembling several protein chains or subunits into a densely packed array

Different visual representations of proteins can provide visible clues about the protein. The ternary (3D) structure of a protein is the three-dimensional shape of the protein chain, which aids in the elucidation of protein function, structure-based drug design and molecular docking. This tertiary structure is obtained using Nuclear magnetic resonance (NMR) spectroscopy/protein X-ray crystallography. Basically, these shapes are determined by the characteristics of the amino acids making up the chain. Moreover, the function of many proteins relies on their three-dimensional shapes.

Comprehending biological structures are not that much easy for everyone except biologist. But many researchers can understand the geometrical view of biological structures. Geometric structures are typically investigated using Euclidean and differential geometry. Since the conformation of protein molecules is exceedingly irregular, we cannot use them to analyse the geometrical structure of proteins. In this scenario, using different graph interpretations of proteins can aid with protein analysis.

In general, researchers perform a change detection analysis to observe the difference between the images. These are done on the basis of principal components, edge detection and some of the operations of images, etc [3]. Similarly, normalized correlation and cosine similarity are the measures used in template matching, a technique for identifying occurrences of a pattern or item inside an image.

A visually insignificant modification in the examined object can significantly impact its fractal measure. Consequently, fractals are aiding in numerous protein research to reveal the protein's characteristics and distinguish defective proteins. Some researchers have computed the fractal dimension using invariant parameters such as fluorescence energy transfer [4], dihedral angles [5] and provided globally acceptable results. However, few researchers have simply discovered any fractal dimension for the backbone images or the 3D view of the proteins and protein surfaces [6]. Though these are unique, their shapes are variants under rotation. This shape variation is reflected in their fractal dimension since some of the dimensions like box dimension [7] and mass dimension [8] are calculated based on the shape of the images. In that case, the gathered results would be authentic only locally rather than globally. So, we must adhere to the invariant graph interpretations of proteins to standardize the conclusion.

This work aids the few invariant and precise graph representations, namely backbone, $\mathcal{PCM}$ and $\mathcal{CGR}$ of proteins. Then, find their fractal dimension using the opted fractal dimension method for the structural proteins of distinct corona viruses belonging to the same family.

Moreover, we discovered a similarity between the structural proteins. Typically, computer-based classification techniques or mathematical criteria are used to evaluate the similarity. Though Euclidean distance, structural similarity index, mean squared error and correlation are a few mathematical measures to evaluate the similarity of images, we assure the similarity with the employment of well recognized mathematical parameters, namely normalized cross-correlation and cosine similarity. As they are the normalized values and the time-minimized user-friendly methods, they provide elegant outputs compared with other methods [9, 10]. Researchers get the idea of developing a drug against *SARS Cov*-2 by identifying similarities between SARS *SARS Cov*-2 and other proteins because a drug against these structural proteins has already been developed for the well-known virus.

For fast computational purposes, we employ the following software, namely, SWISS-MODEL, NAPS, MATLAB R2021a and ImageJ. In short, this work reveals the complexity and similarity of structural proteins and enriches pathological studies. Consequently, this study leads us to drug development and brings us one step closer to gain a depth knowledge.

2 Prologue

2.1 Structural proteins of virus

The Structural proteins act a fundamental role in forming cells, tissues and organisms. Structural proteins have a characteristic amino acid sequence that repeats to form a higher-order structure via intermolecular and intramolecular hydrogen bonding [11–13]. All the viruses that belong to the Coronaviridae family have four structural proteins ($\mathcal{SP}$): Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N).

$\mathcal{SP}_1$ The S protein is a crucial multifunctional and clover-shaped protein that interacts with particular cellular receptors to indicate the host-specificity of different coronaviruses. Furthermore, it occasionally causes fusion between the viral envelope and host-cell membranes, as well as cell-to-cell fusion. The S protein in *SARS Cov*-2 has 76%, 97.4% and 29.8% sequence identity with the S protein of *SARS Cov*, *Bat Cov RaTG*13 and *MERS Cov* respectively.

$\mathcal{SP}_2$ The E protein is a tiny membrane protein that encourages virion formation and viral pathogenicity. Virions are considered necessary for viral assembly and viral release. The sequence identity between *SARS Cov*-2 and *SARS Cov* is 94.7%. Similarly with *Bat Cov RaTG*13 and *MERS Cov* are 100% and 37.33%.

$\mathcal{SP}_3$ The M protein is indispensable for viral assembly in infected cells. Also, it transports and releases the virus from host cell organelles. The sequence identity of *SARS Cov*-2 with *SARS Cov*, *Bat Cov RaTG*13 and *MERS Cov* are 99.5%, 90.5% and 42.3% respectively.

$\mathcal{SP}_4$ The N proteins are indicated in host samples at the prior stage of infection. It improves viral entry and carries out post-fusion cellular processes required for viral survival in the host. The N protein in *SARS Cov*-2 has 90.5% and 95.9% sequence identity with the N protein of *SARS Cov* and *Bat Cov RaTG*13.

2.2 Raw data description

The raw data of protein sequences used in this manuscript are rooted from the NCBI website https://www.ncbi.nlm.nih.gov where the data are submitted by various researchers and are officially published by NCBI under the NCBI Viral RefSeq Project with an open license. The relevant NCBI accession numbers for the raw data are already presented in table 1. With the help of these numbers, you can directly copy or download the protein sequences in easily accessible file formats. Since a part of this study exploits the 3D structures of structural proteins, the raw protein sequences are converted into 3D proteins using homology modelling.

2.3 Homology modelling

Homology modelling, also known as comparative modelling is one of the computational structure prediction approaches used to identify a protein's 3D structure from its amino acid sequence. Since the protein structures are modeled using a template of a well-known experimental structure of a homologous protein, it is recognized as one of the most accurate structure prediction approaches.

Several techniques are available to make homology modelling. Hither, a web-based service named SWISS-MODEL [14] (https://swissmodel.expasy.org/) is exploited to make the homology model of the considered proteins. The process of creating a SWISS-based model consists of the following five significant steps:

S_1 **Input data:** The input/target protein sequences are gathered from various sources. We have compiled the data from the NCBI data in FASTA format. The corresponding IDs related to the raw data are detailed in Table 1.

S_2 **Template searching:** Searching for the appropriate template is essential to obtain a precise homology model in which the templates are sought based on the maximum sequence similarity with the target sequence.

S_3 **Template fixing:** Users have the option for fixing a template. It is done through parameters such as GMQE [15] (Global Model Quality Estimate) and QSQE [16] (Quaternary Structure Quality Estimate) with the values ranging from [0, 1]. Selecting the template with GMQE and QSQE values closer to 1 will provide the exact template.

S_4 **Model building:** The 3D protein model is instinctively created for fixed templates by first transferring conserved atom coordinates defined by the target template alignment. Then the loop modelling generates residual coordinates corresponding to insertions and deletions in the alignment. It is done via the ProMod3 [17] modelling engine since the SWISS-MODEL depends on the OpenStructure computational structural biology framework.

S_5 **Model quality estimation:** In this step, the parameter QMEAN [18] defined in the range [0, 1] is often used to identify the highly reliable 3D model. Finally, the appropriate 3D model is downloaded in the .pdb format.

3 Implementation of chaotic measures on the various interpretation of structural proteins

The 3D proteins can be examined using various portrayals. Accordingly, not only at the image level but also mathematically and statistically the 3D proteins can be analysed. Though the 3D proteins look like images, we can decode the hidden pieces of information with software help and make various representations.

However, this work considered three different kinds of graph representation to analyse structural proteins as shown in Fig. 1, where the first two graphs are different interpretations of 3D proteins while the third one is designed based on the protein sequences. Further, the construction of the different graphs of structural proteins are explained, followed by describing the appropriate fractal dimension approach to uncover the characteristics of the graphs since they followed a fractal-like nature. Finally, the statistical tests are performed to identify the similarity between the various graph interpretations.

Table 1 Various structural proteins and their associated fractal dimensions

Name of Virus	Proteins	Accession number	$FracDim_Z$	$FracDim_M$	$FracDim_B$
MERS Cov	S	YP_009047204.1	1.4499	1.5045	1.5945
	E	YP_009047209.1	1.2130	1.2512	1.2933
	M	YP_009047210.1	1.2103	1.2387	1.2899
	N	QBM11755.1	1.3201	1.4212	1.4580
Bat Cov RaTG13	S	QHR63300.2	1.4545	1.5076	1.5981
	E	QHR63302.1	1.2206	1.2637	1.3016
	M	QHR63303.1	1.2128	1.2411	1.2908
	N	QHR63308.1	1.3496	1.4215	1.4685
SARS Cov	S	YP_009825051.1	1.4405	1.4981	1.5874
	E	YP_009825054.1	1.1984	1.2195	1.2682
	M	YP_009825055.1	1.2093	1.2387	1.2899
	N	YP_009825061.1	1.3434	1.4126	1.4614
SARS Cov-2	S	YP_009724390.1	1.4592	1.5108	1.6016
	E	YP_009724392.1	1.2206	1.2637	1.3016
	M	YP_009724393.1	1.2230	1.2596	1.3041
	N	YP_009724397.2	1.3558	1.4305	1.4756

S spike, *E* envelope, *M* membrane, *N* nucleocapsid

3.1 Zig–zag curve representation

It is a simplistic indication of 3D proteins, enabling the polymer chain structure to be formed by connecting the alpha carbons of the proteins [7]. The MATLAB molviewer [19] (https://in.mathworks.com/help/bioinfo/ref/moleculeviewer-app.html) acquired the zig-zag curve by taking the homology modelled .pdb file as input.

Method 3.1 Fractal dimension

The fractal dimension is depicted for a structure with a self-similar characteristic that is invariant under a scale transformation. Since proteins are unbranched polymers, they have a statistically self-similar property.

In this method [20], the protein molecule is considered a function of the fineness or coarseness of the scale r. This is similar to the walker dimension method. Herewith, the zig-zag curve is acquired by connecting the alpha atoms of the proteins with the interval of r residues starting from the alpha atoms of the N-terminal residue. Continue the process until spot the less residues to make the next move. Repeat the process with different residues r.

The length of the protein molecule on a scale of r (i.e.,$\mathcal{L}(r)$) is obtained by the sum of the length of the zig-zag line $\mathcal{L}_z$ and the correction term $\mathcal{C}_t$, which are calculated using the following equations:

$$\mathcal{L}(r) = \mathcal{L}_z + \mathcal{C}_t$$

with

$$\mathcal{C}_t = \frac{N(u)}{r+1} \times \mathcal{L}_m$$

Here $N(u)$ indicates the number of remaining disconnected residues and $\mathcal{L}_m$ represents the average length of the fractional lines that make up the zig-zag lines. Then the fractal dimension is acquired using the relation

$$N = \left(\frac{\mathcal{L}(r)}{r}\right)^{\mathrm{FracDim}_Z},$$

where N represents the total number of residues.

3.2 Protein contact map ($\mathcal{PCM}$)

Over the last 3 decades, the research on $\mathcal{PRN}$ have evolved. Initially, this is a kind of mathematical formulation of 3D proteins derived from graph theory. In a nutshell, $\mathcal{PRN}$ is a way of assessing protein structure in which residues are seen as nodes and edges reflect the interaction between them. The structure and function of proteins are greatly assisted by determining the centrality of these networks. Furthermore, this $\mathcal{PRN}$ study revealed several valuable approaches and tactics for discovering the information of proteins.

In this sequence, $\mathcal{PCM}$ is used to explore the contact/connectivity between all possible amino acid residue pairs in a three-dimensional protein structure using a two-dimensional binary matrix. In other words, this is a kind of adjacency matrix representation of $\mathcal{PRN}$. This can be expressed as follows.

Let r_m, r_n be any two residues. The edge between any two residues are denoted by $e_{r_m r_n}$. Then, the $\mathcal{PCM}$ of

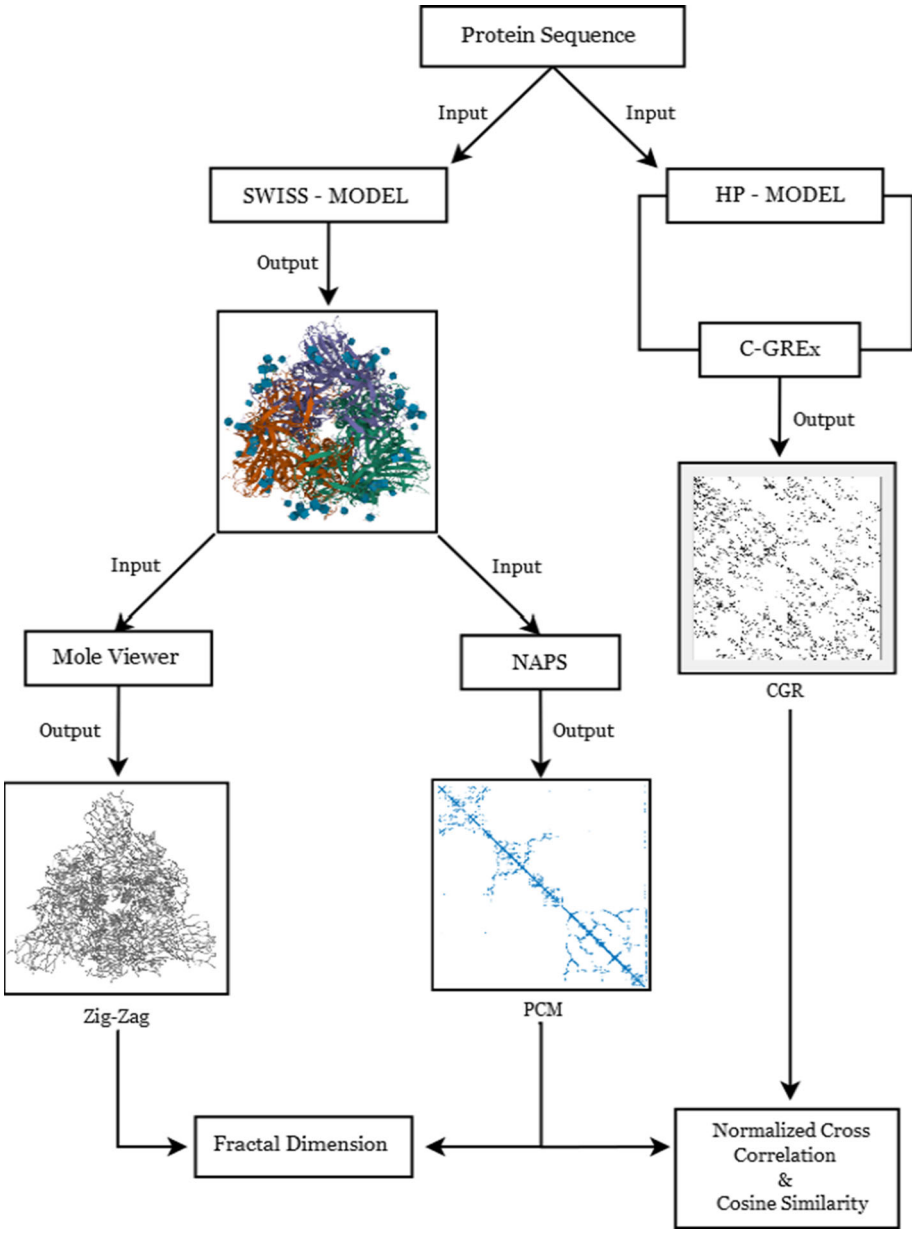


Fig. 1 Flowchart for a fast overview of the technical processes used in this study

a 3D protein is defined as

$$\mathcal{PCM} = \begin{cases} 1 & \text{if } e_{r_m r_n} \text{ exists} \\ 0 & \text{Otherwise} \end{cases}$$

With the help of contact maps, the similarity between proteins are quickly evaluated. Moreover, the contact maps are translation and rotation invariant. As a consequence, the coordinates of proteins are reconstructed without interruption.

Both $\mathcal{PRN}$ and $\mathcal{PCM}$ are acquired from NAPS (Network Analysis of Protein Structures) [21], a web-based tool that provides the desired output by giving input or uploading the PDB file on the web page http://bioinf.iiit.ac.in/NAPS/index.php .

Method 3.2 Mass dimension
The mass dimension ($FracDim_M$) is like a traditional box dimension. Instead of counting the boxes (with size $\check{\varepsilon}$) that covers the taken object, we count the average number of pixels $\mu_{\check{\varepsilon}}$ in the boxes. It quantifies the connectivity of $\mathcal{PRN}$ and can be calculated for each pixel in a given image and used to identify the irregularities of heterogeneous geometrical objects. The following equation can be used to calculate $FracDim_M$.

$$FracDim_M = \lim_{\check{\varepsilon} \to 0} \frac{ln\ \mu_{\check{\varepsilon}}}{ln\ \check{\varepsilon}}.$$

3.3 Chaos game representation ($\mathcal{CGR}$)

The $\mathcal{CGR}$ could be used to translate the amino acid sequence into bi-dimensional real values. It helps to

preserve the statistical properties of the sequences and offers information on the global and local patterns of the sequences. Each sequence member has a corresponding point representation in $\mathcal{CGR}$. As a result, each amino sequence has a unique $\mathcal{CGR}$ [22].

Initially, protein sequences consist of 20 different amino acids, including phenylalanine (F), histidine (H), isoleucine (I), lysine (K), leucine (L), methionine (M), threonine (T), valine (V), tryptophan (W), alanine (A), aspartic acid (D), glutamic acid (E), asparagine (N), serine (S), cysteine (C), glycine (G), proline (P), glutamine (Q), arginine (R) and tyrosine (Y). Among these, the first nine amino acids are essential, the succeeding five are non-essential amino acids and the remaining six are conditional amino acids.

These sequences can be represented as a $\mathcal{CGR}$ graph using the HP model. It is a lattice-based model introduced in the year of 1985 in which the amino acids are categorized into four groups: the eight amino acids A, I, L, M, F, P, W, and V are categorized as a non-polar class; the seven amino acids N, C, Q, G, S, T, and Y are labeled as an uncharged polar class; the three amino acids R, H, and K are considered a positive polar class; and the remaining two amino acids D and E are designated as a negative polar class. The only premise is that two non-polar amino acids interact iff they are spatially close to each other, whereas the other techniques are based on statistically predicting interactions from the primary sequence. Thus, the HP model is more precise and that is why the HP model is widely chosen [23].

To convert a given protein's amino acid sequence into $\mathcal{CGR}$ coordinates, consider the given protein sequence as follows:

$$Seq = \{S_1 S_2 \dots S_N\}, \tag{1}$$

where S_i, $i = 1, 2, ...N$ indicates the elements in the amino acid sequence and N represents the length of the sequence.

Let define

$$s_i = \begin{cases} 0 \text{ if } S_i \text{ is non polar} \\ 1 \text{ if } S_i \text{ is negative polar} \\ 2 \text{ if } S_i \text{ is uncharged polar} \\ 3 \text{ if } S_i \text{ is positive polar} \end{cases}$$

Then Eq. (1) is reconstructed as

$$\mathcal{X}(S) = s_1 s_2 ... s_N, \qquad s_i \in \{0, 1, 2, 3\}. \tag{2}$$

To plot the $\mathcal{CGR}$, consider the $[0, 1] \times [0, 1]$ square with 4 corners that reflects the values 0, 1, 2 and 3. Then, plot the first element in the sequence $\mathcal{X}(S)$ by positioned the point halfway between the centre of the square. Similarly, plot the i^{th} point in the halfway between the $(i-1)th$ point. Continuing in this way, the desired $\mathcal{CGR}$ of the protein sequence is constructed.

The sample protein sequence and the corresponding reconstructed sequences are presented in the following figures 2 and 3. This interpretation is made via the software C-GREx (https://sites.google.com/site/cgrexonline/).

Method 3.3 Box counting dimension
The fractal dimension of $\mathcal{CGR}$ is determined using the standard box count approach, where the provided $\mathcal{CGR}$ is covered by tiny boxes of size $(\check{\varepsilon})$. The number of boxes $(\mathcal{N}(\check{\varepsilon}))$ that cover the item varies depending on the box size. The following expression can be used to compute the box dimension:

$$FracDim_B = \lim_{\check{\varepsilon} \to 0} - \frac{\log \mathcal{N}(\check{\varepsilon})}{\log \check{\varepsilon}}$$

4 Methods for finding similarity between proteins

This section deals with a few statistical methods that are used to investigate the changes in the $\mathcal{PCM}$ and $\mathcal{CGR}$ interpretation of the structural proteins. It will aid in understanding the worthiness of the fractal measures. Since $\mathcal{PCM}$ and $\mathcal{CGR}$ have translational and rotational invariant, the normalized cross-correlation and cosine similarity are performed to analyse the similarity between the structural proteins of *SARS Cov* and *SARS Cov*-2.

Method 4.1 Normalized cross-correlation

The normalized correlation coefficient is a number that represents the similarity of two images in terms of pixel intensity, where the two taken images $\mathcal{A}$ and $\mathcal{B}$ are converted into two distinct data sets X and Y [24].

MKSSHHHHHHENLYFQSNATKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQT
QGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKL
DDKDPNFKDQVILLNKHIDAY

Fig. 2 Amino acid representation of sample protein

032233333312020222023320010233032332023022022002332012122202121003222123300200200020200020230201020222002222003011310203120000233010 2

Fig. 3 A reconstructed sequence representation of Fig 2

Then the normalized cross-correlation ($\mathcal{CC}_n$) is calculated using the forthcoming equation:

$$\mathcal{CC}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Here, $\mathcal{CC}_n$ is the correlation coefficient and $x_i \in X$ & $y_i \in Y$. The value of $\mathcal{CC}_n$ is in the range in $[-1, 1]$, where $\mathcal{CC}_n = -1$ reflect that the taken two images are negatively correlated; $\mathcal{CC}_n = 0$ implies that there is no correlation between the considered images; $\mathcal{CC}_n = 1$ denotes that the taken images are exactly identical. The $\mathcal{CC}_n$ calculation in this work is done through the MATLAB normxcorr2() function (https://in.mathworks.com/help/images/ref/normxcorr2.html).

Method 4.2 Cosine similarity

Cosine similarity is a mathematical metric used in many machine learning algorithms that measures the similarity between two sequences of numbers. This approach calculates the cosine similarity by taking into account the cosine of the angle between the vectors that make up the number sequence [25].

Thus, to find out the cosine similarity ($\mathcal{S}_\mathcal{C}$) between the two considered images $\mathcal{A}$ and $\mathcal{B}$, they are first converted into vectors A and B. Then the cosine similarity is calculated using the following formula:

$$\mathcal{S}_\mathcal{C} = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

where $A_i \in A$ and $B_i \in B$. The range of $\mathcal{S}_\mathcal{C}$ is $[-1, 1]$. It represents the similarity level between the images $\mathcal{A}$ and $\mathcal{B}$. If the $\mathcal{S}_\mathcal{C}$ is 1, then the images $\mathcal{A}$ and $\mathcal{B}$ are the same, i.e., the vectors are proportional, whereas the value -1 indicates that the images are totally different. In precisely, we can consider its range in [0, 1], since the $\mathcal{S}_\mathcal{C}$ reflects that there is no similarity between images, i.e., the vectors are orthogonal. As $\mathcal{S}_\mathcal{C}$ is based on the angles between the vectors, it is the best method to measure the similarity even if the images are of different sizes. The $\mathcal{S}_\mathcal{C}$ calculation is done via the Python software (https://soumilshah1995.blogspot.com/2020/07/computing-similarity-on-images-using.html).

5 Results and discussion

The fractal dimension is an elegant parameter to identify minor changes. According to this fact, visually different images can have the same fractal dimensions. However, if the virtually similar images (with the same image quality and size) have different dimensions, then there must be a minute change between the images.

Though the $\mathcal{PCM}$ and $\mathcal{CGR}$ of taken structural proteins appear to be the same (refer to Appendix), their fractal dimensions vary, which means that there is a minute change between the $\mathcal{PCM}$ and $\mathcal{CGR}$ of structural proteins. i.e., There is a change in the corresponding residue-residue interaction and the amino acid sequences.

Descriptively, the alpha carbons of proteins generate the zig-zag curves and $\mathcal{PCM}$, whereas the protein sequence is used to evolve the $\mathcal{CGR}$. As a result, even small changes in the alpha carbons and protein sequences are reflected in the fractal dimension. This allows us to detect identical proteins.

Based on the literature [13], the E protein of *Bat Cov RaTG*13 and *SARS Cov*-2 are identical, i.e., both have same type of envelope protein. We can see this reflection in the fractal dimension. In our investigation, the E protein of *Bat Cov RaTG*13 and *SARS Cov*-2 has identical fractal dimensions. This information leads us to conclude that the fractal dimensions can be used to find the identical proteins.

The obtained fractal dimension values are presented in the following Table 1.

The fractal dimension reflects the complexity of the objects. The fractal dimension increases as the thing becomes more complicated. Table 1 shows that the fractal dimension corresponding to the different representations of *SARS Cov*-2 structural proteins is greater than the other. As a result, we can identify that *SARS Cov*-2 structural proteins are more complicated than the other structural proteins where the structural complexity of proteins depends on the interactions among their residues. The acquired graph interpretation becomes denser and more complicated when the interactions are more. As the fractal dimension quantifies the small changes essentially, we said that fractals aid in finding the structural complexity of structural proteins.

The normalized cross-correlation and cosine similarity of the $\mathcal{PCM}$ and $\mathcal{CGR}$ explores the match up of *SARS Cov* − 2 with *MERS Cov*, *Bat Cov RaTG*13, and *SARS Cov*.

The $\mathcal{CGR}$ is a kind of representation of amino acids. The acquired $\mathcal{CGR}$'s normalized cross-correlation values and cosine similarity index are virtually identical to the percent of sequence identity of structural proteins (refer the Sect. 2.1), where the sequence identity refers to the presence of the same nucleotide or amino acid in structural proteins. The $\mathcal{PCM}$ reflects the residue–residue interaction of the structural proteins. Through the $\mathcal{CC}_n$ and $\mathcal{S}_\mathcal{C}$ of $\mathcal{PCM}$, we can get information about how the residue–residue interaction varies between the structural proteins of virus.

From Tables 1, 2, 3, 4, 5, 6, 7, we assure that the normalized cross-correlation and cosine similarity of $\mathcal{PCM}$ is near the sequence identity. As a result, we can utilize $\mathcal{CGR}$ and $\mathcal{PCM}$ to analyse sequence identity rather

Table 2 Normalized cross-correlation between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *MERS Cov* and *SARS Cov*-2

Proteins	$\mathcal{CC}_n(\mathcal{PCM})$	$\mathcal{CC}_n(\mathcal{CGR})$
S	0.25573	0.2813
E	0.3387	0.3578
M	0.3918	0.4175
N	0.5601	0.5713

Table 3 Normalized cross-correlation between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *Bat Cov RaTG*13 and *SARS Cov*-2

Proteins	$\mathcal{CC}_n(\mathcal{PCM})$	$\mathcal{CC}_n(\mathcal{CGR})$
S	0.9327	0.9685
E	0.9815	1.0000
M	0.9535	0.9785
N	0.9260	0.9469

Table 4 Normalized cross-correlation between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *SARS Cov* and *SARS Cov*-2

Proteins	$\mathcal{CC}_n(\mathcal{PCM})$	$\mathcal{CC}_n(\mathcal{CGR})$
S	0.7325	0.7585
E	0.9115	0.9457
M	0.8807	0.8992
N	0.8872	0.9084

Table 5 Cosine similarity between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *MERS Cov* and *SARS Cov*-2

Proteins	$\mathcal{S}_\mathcal{C}(\mathcal{PCM})$	$\mathcal{S}_\mathcal{C}(\mathcal{CGR})$
S	0.2843	0.3025
E	0.3525	0.3781
M	0.4125	0.4302
N	0.5797	0.6039

Table 6 Cosine similarity between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *Bat Cov RaTG*13 and *SARS Cov*-2

Proteins	$\mathcal{S}_\mathcal{C}(\mathcal{PCM})$	$\mathcal{S}_\mathcal{C}(\mathcal{CGR})$
S	0.9658	0.9803
E	0.9987	1.0000
M	0.9632	0.9878
N	0.9425	0.9602

than the other techniques since both biologists and non-biologist researchers easily understand the $\mathcal{CGR}$ and $\mathcal{PCM}$ graph interpretation. Among which $\mathcal{CGR}$'s $\mathcal{CC}_n$ and $\mathcal{S}_\mathcal{C}$ provide the more appropriate sequence identity because both the sequence identity and $\mathcal{CGR}$ are designed based on the protein sequences. In fact, compared with $\mathcal{CC}_n$, $\mathcal{S}_\mathcal{C}$ is closer to the sequence identity since it is based on the angle between the points rather than the distance. Therefore, a slight change is reflected in the $\mathcal{S}_\mathcal{C}$. Based on the obtained outcomes, we assure that *SARS Cov*-2 is closely related to *Bat Cov RaTG*13.

Table 7 Cosine similarity between $\mathcal{PCM}$ and $\mathcal{CGR}$ of *SARS Cov* and *SARS Cov*-2

Proteins	$\mathcal{S}_\mathcal{C}(\mathcal{PCM})$	$\mathcal{S}_\mathcal{C}(\mathcal{CGR})$
S	0.7445	0.7652
E	0.9392	0.9568
M	0.8906	0.9205
N	0.8932	0.9095

6 Conclusion

We have examined alternative ways to understand the protein sequences. Through the fractal dimension, we have tracked the visually minor or negligible changes and have identified that *SARS Cov*-2 possesses a more complex structure. In addition, with the help of normalized cross-correlation and cosine similarity, we have ensured that *Bat Cov RaTG*13 and *SARS Cov*-2 have similar structural proteins rather than other virus. Thus, we can ensure that *SARS Cov*-2 originated from *Bat Cov RaTG*13 and can utilize the *Bat Cov RaTG*13 drug combination ideas to develop the drugs against *SARS Cov*-2. This study does not need any deep biological knowledge to analyse the structural proteins. We hope this work will give a new perspective on protein analysis for researchers from non-biological backgrounds.

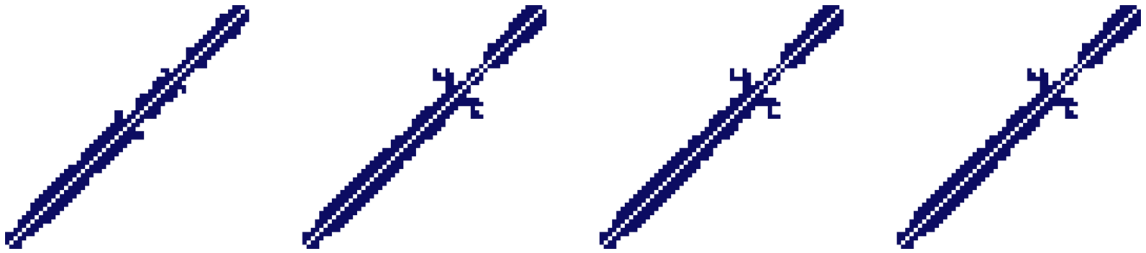
Data Availability Statement The raw data sets used during the current study are available at https://www.ncbi.nlm.nih.gov without restriction. The homology modelling done in this study is done through the web page https://swissmodel.expasy.org/ and the protein contact maps are generated via http://bioinf.iiit.ac.in/NAPS/index.php.

Declarations

Conflict of interest There is no conflict of interest.

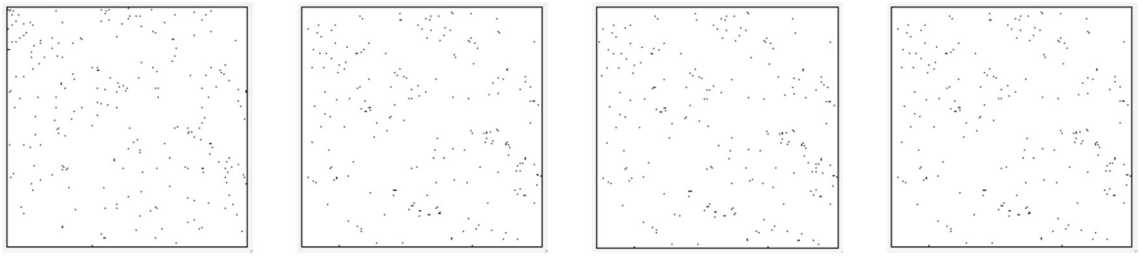
Appendix

See Figs. 4 and 5.



(a) *MERSS Cov*-2 (b) *Bat Cov RaTG13* (c) *SARS Cov* (d) *SARS Cov*-2

Fig. 4 $\mathcal{PCM}$ interpretation of E protein



(a) *MERSS Cov*-2 (b) *Bat Cov RaTG13* (c) *SARS Cov* (d) *SARS Cov*-2

Fig. 5 $\mathcal{CGR}$ interpretation of E protein

References

1. F.K. Yoshimoto, Protein J. **39**(3), 198–216 (2020). https://doi.org/10.1007/s10930-020-09901-4
2. R. Yadav, J.K. Chaudhary, N. Jain, P.K. Chaudhary, S. Khanra, P. Dhamija, S. Handu, Cells **10**(4), 821 (2021). https://doi.org/10.3390/cells10040821
3. S.P. Mishra, J. Theor. Appl. Inform. Technol **64**(3), 820–836 (2014)
4. T.G. Dewey, M.M. Datta, Biophys. J. **56**(2), 415–420 (1989). https://doi.org/10.1016/S0006-3495(89)82687-6
5. Xin Peng, Wei Qi, Mengfan Wang, Su. Rongxin, Zhimin He, Commun. Nonlinear Sci. Numer. Simul. **18**(12), 3373–3381 (2013). https://doi.org/10.1016/j.cnsns.2013.05.005
6. D. Craciun, A. Isvoran, N. Avram, Phys. Ser. **52**, 116 (2008)
7. R. K. Rout, P. Pal Choudhury, S. P. Maity, B. S. Daya Sagar, S. S. Hassan, Comput. Methods Biomech. Biomed. Eng. Imaging Visual. 6(2), 192-203 (2018). https://doi.org/10.1080/21681163.2016.1214850
8. M.B. Enright, D.M. Leitner, Phys. Rev. E. **71**(1), 011912 (2005). https://doi.org/10.1103/PhysRevE.71.011912
9. U. Sara, M. Akter, M.S. Uddin, J. Comput. Commun. **7**(3), 8–18 (2019). https://doi.org/10.4236/jcc.2019.73002
10. H. V. Nguyen, L. Bai, In: *Asian conference on computer vision*, Springer. pp. 709-720 (2010). https://doi.org/10.1007/978-3-642-19309-5_55
11. J. Verma, N. Subbarao, Arch. Virol. **166**(3), 697–714 (2021). https://doi.org/10.1007/s00705-021-04961-y
12. E.A. Aldaais, S. Yegnaswamy, F. Albahrani, F. Alsowaiket, S. Alramadan, Biochem. Biophys. Rep. **26**, 101023 (2021). https://doi.org/10.1016/j.bbrep.2021.101023
13. M.E.A. Mohammed, J. Proteins Proteom. **12**(2), 81–91 (2021). https://doi.org/10.1007/s42485-021-00060-3
14. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, T. Schwede, Nucl. Acids Res. **46**(W1), W296–W303 (2018). https://doi.org/10.1093/nar/gky427
15. M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, T. Schwede, Nucl. Acids Res. **42**(W1), W252–W258 (2014). https://doi.org/10.1093/nar/gku340
16. M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, Sci. Rep. **7**(1), 1–15 (2017). https://doi.org/10.1038/s41598-017-09654-8
17. G. Studer, G. Tauriello, S. Bienert, M. Biasini, N. Johner, T. Schwede, PLoS Comput. Biol. **17**(1), e1008667 (2021). https://doi.org/10.1371/journal.pcbi.1008667
18. P. Benkert, S.C. Tosatto, D. Schomburg, Proteins Struct. Funct. Bioinform **71**(1), 261–277 (2008). https://doi.org/10.1002/prot.21715
19. N. Niknam, H. Khakzad, S.S. Arab, H. Naderi-Manesh, Comput. Biol. Med. **72**, 151–159 (2016). https://doi.org/10.1016/j.compbiomed.2016.03.012
20. F. Torrens, Molecules **7**(1), 26–37 (2002). https://doi.org/10.3390/70100026
21. B. Chakrabarty, V. Naganathan, K. Garg, Y. Agarwal, N. Parekh, Nucl. Acids Res. **47**(W1), W462–W470 (2019). https://doi.org/10.1093/nar/gkz399

22. Z.G. Yu, V. Anh, K.S. Lau, J. Theor. Biol. **226**(3), 341–348 (2004). https://doi.org/10.1016/j.jtbi.2003.09.009
23. K.A. Dill, Biochemistry **24**(6), 1501–1509 (1985). https://doi.org/10.1021/bi00327a032
24. F. Zhao, Q. Huang, W. Gao, In: *2006 IEEE international conference on acoustics speech and signal processing proceedings*. 2, II-II (2006). https://doi.org/10.1109/ICASSP.2006.1660446
25. M. J. Falato, B. T. Wolfe, T. M. Natan, X. Zhang, R. M. Marshall, Y. Zhou, Z. Wang (2022). https://doi.org/10.48550/arXiv.2205.04609. arXiv preprint arXiv:2205.04609