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

Computers in Biology and Medicine 131 (2021) 104247

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Non-standard bioinformatics characterization of SARS-CoV-2

Dorota Bielińska-Wąż [a,*], Piotr Wąż [b]

[a] Department of Radiological Informatics and Statistics, Medical University of Gdańsk, 80-210, Gdańsk, Poland
[b] Department of Nuclear Medicine, Medical University of Gdańsk, 80-210, Gdańsk, Poland

ARTICLE INFO

Keywords:
Alignment-free methods
Moments of inertia
Similarity/dissimilarity analysis of DNA/RNA sequences

ABSTRACT

A non-standard bioinformatics method, *4D-Dynamic Representation of DNA/RNA Sequences*, aiming at an analysis of the information available in nucleotide databases, has been formulated. The sequences are represented by sets of "material points" in a 4D space - 4D-dynamic graphs. The graphs representing the sequences are treated as "rigid bodies" and characterized by values analogous to the ones used in the classical dynamics. As the graphical representations of the sequences, the projections of the graphs into 2D and 3D spaces are used. The method has been applied to an analysis of the complete genome sequences of the 2019 novel coronavirus. As a result, 2D and 3D classification maps are obtained. The coordinate axes in the maps correspond to the values derived from the exact formulas characterizing the graphs: the coordinates of the centers of mass and the 4D moments of inertia. The points in the maps represent sequences and their coordinates are used as the classifiers. The main result of this work has been derived from the 3D classification maps. The distribution of clusters of points which emerged in these maps, supports the hypothesis that SARS-CoV-2 may have originated in bat and in pangolin. Pilot calculations for Zika virus sequence data prove that the proposed approach is also applicable to a description of time evolution of genome sequences of viruses.

1. Introduction

The rapid growth of nucleotide databases, including GenBank, stimulated the development of methods aiming at the numerical characterization of the considered objects. One group of methods is composed of the *alignment-free* bioinformatics methods. This branch of bioinformatics constitutes an alternative for the standard approaches, based on the analysis of alignments of the considered sequences.

Alignment-free methods are, usually, fast and computationally simple. They are exceptionally useful for big data analysis. There exist many different alignment-free methods. For example, Zhou et al. constructed a complex network for similarity/dissimilarity analysis of DNA sequences [1]. Saw et al. perform DNA sequence comparison using the fuzzy integral with Markov chain [2]. Lichtblau uses Frequency Chaos Game Representation and signal processing for genomic sequence comparison [3]. He et al. propose a numerical representation of a DNA sequence called Subsequence Natural Vector and apply it for HIV-1 subtype classification [4]. Many other alignment-free methods are reviewed in Refs. [5,6]. Within this group of approaches, one can extract *graphical representations of biological sequences* applicable to both graphical and numerical similarity/dissimilarity analysis of biological sequences.

Similarity analysis is strictly related to the classification studies, which supply valuable information in various areas of science [7,8]. In particular, we classified different kinds of objects, such as stellar spectra in astrophysics [9], molecular spectra in the theory of molecular similarity [10,11], groups of individuals in social science [12], solutions of differential equations in the chaotic systems [13], biological sequences in bioinformatics [14]. The problem of similarity of complex objects is not unique. Multi-dimensional objects can be similar in one aspect and very different if some other aspects are considered. Additionally, different aspects of similarity can be important in descriptions of various problems. It is not obvious how to represent graphically multidimensional objects in two or three dimensions in such a way that their features are visible. It is also not obvious how to represent numerically such graphical objects.

The early graphical approaches to a description of DNA sequences were based on walks in three [15,16] and in two dimensions [17–19]. These works initiated a rapid development of graphical bioinformatics branch, and many different approaches have been designed, as for example [20–32] (for reviews see Refs. [33,34]).

In the alignment-free methods one can create a large number of different *descriptors* (numerical values characterizing the graphs). It is

* Corresponding author.
E-mail addresses: djwaz@gumed.edu.pl (D. Bielińska-Wąż), phwaz@gumed.edu.pl (P. Wąż).

https://doi.org/10.1016/j.compbiomed.2021.104247
Received 20 December 2020; Received in revised form 22 January 2021; Accepted 26 January 2021
Available online 1 February 2021

important (though sometimes difficult) that both graphs and the descriptors represent the sequence in a unique, degeneracy-free, manner.[1] The first descriptors based on graphical representations of sequences have been created by Raychaudhury and Nandy [35] and by Randić et al. [36]. In our works, several kinds of descriptors have been defined. One of them are statistical distribution moments derived from different statistical distributions describing DNA sequences [33]. Spectral distribution moments in sequence similarity studies were also used by Agüero-Chapin et al. [37] (for review see Ref. [38]). Another family of descriptors, introduced in the graphical representation methods called by us *Dynamic*, emerged by the inspiration taken from the classical dynamics. In particular, in 2D-Dynamic Representation of DNA/RNA Sequences, the sequences are represented by sets of "material points" in a 2D space [39]. This method has also been extended to three dimensions. Based on our method, two different approaches have been published under the same name – one by us [40] and another one by Aram and Iranmanesh [41].

In the present work, we extend the Dynamic Representations of DNA/RNA Sequences to four dimensions. The four-dimensional method belongs to the group of numerical alignment-free methods. Since we propose a new kind of visualization, the method can be also considered as a multidimensional graphical approach. In general, visualization of multidimensional methods is difficult, if possible at all. Nevertheless, visualizations of 4D [42,43] and 5D representations [44] have been introduced. In the extension of the Dynamic Representation method to four dimensions, each nucleobase is represented by a basis vector located on a separate axis, analogously as in method introduced by us for amino acid sequences [45]. A larger set of descriptors gives more complete view on the similarity problem. In the proposed approach, the sequence is represented by a set of 4D material points, *4D-dynamic graph*. For a numerical characterization, we treat the graph as a rigid body, analogously as in the classical dynamics. Nowadays, during the pandemic of coronavirus disease (COVID-19), studies on the 2019 novel coronavirus (SARS-CoV-2) are particularly important [46,47]. Therefore, the new method, *4D-Dynamic Representation of DNA/RNA Sequences*, has been introduced using coronavirus genome sequences.

2. Materials and methods

In the present work, we propose an approach in which DNA/RNA sequence is represented by a set of material points, called by us 4D-dynamic graph. The distribution of these points in the 4D space is determined by shifts according to unit vectors representing the nucleobases of the sequence. The first shift is performed from point $(0,0,0,0)$ according to the unit vector representing the first nucleobase in the sequence. The end of this vector marks the starting point for the next shift, defined by the second nucleobase in the sequence. The procedure is repeated until the last nucleobase is reached. At the end of each unit vector, a material point with mass $m = 1$, is set. An example of the assignment of the unit vectors is given in Table 1: The unit vector A=(1,0,0,0) represents adenine, C=(0,1,0,0) – cytosine, G=(0,0,1,0) – guanine, and T/U=(0,0,0,1) – thymine/uracil. Let us consider a model sequence AUGAC. Since the length of the sequence is five, the abstract graph consists of five 4D points. We start the walk at the origin of the Cartesian coordinate system. The first nucleobase, A, is represented by (1,0,0,0). We locate the first material point, at the end of this vector. Then, this point is the starting point for the next shift according to the unit vector representing the second nucleobase, U, and so on. The coordinates of material points of the 4D-dynamic graph corresponding to this sequence, and for the unit vectors collected in Table 1, are given in Table 2.

The final results related to similarity/dissimilarity analysis of the sequences are independent of assignments of particular unit vectors to the nucleobases.

Table 1
Representation of the nucleobases in the four-dimensional coordinate system.

Axis No.	Nucleobase	Symbol
1	Adenine	A
2	Cytosine	C
3	Guanine	G
4	Thymine/Uracil	T/U

Table 2
4D-dynamic graph representing a model sequence AUGAC for the basis vectors defined in Table 1.

m_i	$(x_i^1, x_i^2, x_i^3, x_i^4)$
m_1	(1, 0, 0, 0)
m_2	(1, 0, 0, 1)
m_3	(1, 0, 1, 1)
m_4	(2, 0, 1, 1)
m_5	(2, 1, 1, 1)

In order to visualize the 4D-dynamic graphs, we project them into 2D or 3D space. For example, if we put x_i^1 coordinates equal to zero, then we obtain a 3D projection, denoted as $x^2x^3x^4$-graph. The distribution of masses in 3D or 2D space gives some information about the locations of three or two nucleobases in the sequence.

As the descriptors of the 4D-dynamic graphs we propose values analogous to the ones used in the classical dynamics. The coordinates of the center of mass of 4D-dynamic graphs are defined as

$$\mu^k = \frac{\sum_{i=1}^{N} m_i x_i^k}{\sum_{i=1}^{N} m_i},\ k = 1,2,3,4. \tag{1}$$

x_i^k are the coordinates of m_i in the 4D space. Assuming the mass $m_i = 1$ for each material point, the total mass of the 4D-dynamic graph is

$$N = \sum_{i=1}^{N} m_i, \tag{2}$$

where N is the length of the sequence. Then, the coordinates of the center of mass may be rewritten as

$$\mu^k = \frac{1}{N}\sum_{i=1}^{N} x_i^k. \tag{3}$$

The tensor of the moment of inertia is defined as a 4×4 matrix:

$$\widehat{I} = \begin{pmatrix} I_{1\,1} & I_{1\,2} & I_{1\,3} & I_{1\,4} \\ I_{2\,1} & I_{2\,2} & I_{2\,3} & I_{2\,4} \\ I_{3\,1} & I_{3\,2} & I_{3\,3} & I_{3\,4} \\ I_{4\,1} & I_{4\,2} & I_{4\,3} & I_{4\,4} \end{pmatrix}. \tag{4}$$

The matrix elements are

$$I_{j\,j} = \sum_{i=1}^{N} m_i \sum_{k=1}^{4} \left[\widehat{x}_i^k\left(1-\delta_{jk}\right)\right]^2, \tag{5}$$

$$I_{j\,k} = I_{k\,j} = -\sum_{i=1}^{N} m_i \widehat{x}_i^j \widehat{x}_i^k, \tag{6}$$

where

$$\delta_{jk} = \begin{cases} 1 & j = k, \\ 0 & j \neq k \end{cases}$$

is the Kronecker-Delta. $\widehat{x}_i^k$, $k = 1,2,3,4$ are the coordinates of m_i in the

[1] A description is degenerate (nonunique) if several different sequences are represented by the same graph or by the same set of descriptors.

Table 3
Similarity values [%] obtained using 4D-Dynamic Representation of DNA/RNA Sequences for SARS-CoV-2: MT106054 (Sa_1), MT159708 (Sa_2), MT192772 (Sa_3) and Embecovirus: AY391777 (Em_1), KM349744 (Em_2), FJ647223 (Em_3).

DES	Sequence	Sa_1	Sa_2	Sa_3	Em_1	Em_2	Em_3
r_1^{4D}	Sa_1	100	99.99	99.97	95.80	95.84	95.98
	Sa_2		100	99.97	95.81	95.85	95.98
	Sa_3			100	95.84	95.87	96.01
	Em_1				100	99.96	99.82
	Em_2					100	99.86
	Em_3						100
r_2^{4D}	Sa_1	100	99.99	99.97	95.80	95.84	95.98
	Sa_2		100	99.97	95.81	95.85	95.98
	Sa_3			100	95.83	95.88	96.01
	Em_1				100	99.96	99.81
	Em_2					100	99.86
	Em_3						100
r_3^{4D}	Sa_1	100	99.99	99.97	95.81	95.85	95.99
	Sa_2		100	99.97	95.82	95.86	96.00
	Sa_3			100	95.85	95.89	96.02
	Em_1				100	99.96	99.81
	Em_2					100	99.86
	Em_3						100
r_4^{4D}	Sa_1	100	100	99.92	76.59	65.19	67.41
	Sa_2		100	99.91	76.59	65.19	67.40
	Sa_3			100	76.66	65.25	67.46
	Em_1				100	85.12	88.01
	Em_2					100	96.72
	Em_3						100

Cartesian coordinate system for which the origin has been selected at the center of mass, i.e.

$$\widehat{x}_i^k = x_i^k - \mu^k. \tag{7}$$

The eigenvalue problem of the tensor of inertia with the eigenvectors ω_k and the eigenvalues I_k is defined as:

$$\widehat{I}\omega_k = I_k\omega_k,\ k = 1,2,3,4. \tag{8}$$

Solving the fourth-order secular equation

$$\det\left(\widehat{I} - I\widehat{E}\right) = 0, \tag{9}$$

where $\widehat{E}$ is 4×4 unit matrix, we obtain the eigenvalues I_k called the *principal moments of inertia*.

As numerical characterization of the 4D-dynamic graphs we propose normalized principal moments of inertia:

$$r_k^{4D} = \sqrt{\frac{I_k}{N}},\ k = 1,2,3,4. \tag{10}$$

For a numerical comparison of a pair of sequences labeled by i and j we apply the similarity measure introduced by us in Ref. [14]:

$$S_1(i,j) = S_1(j,i) = \frac{Min\{|DES(i)|, DES(j)|\}}{Max\{|DES(i)|, DES(j)|\}}100\%, \tag{11}$$

where $DES(i)$ denotes the descriptor representing the i-th sequence and $DES(j)$ denotes the same descriptor representing the j-th sequence.

For a pair of sequences with identical descriptors $DES(i) = DES(j)$, the similarity value $S_1 = 100\%$.

Alternatively, one can apply another similarity measure introduced by us in Ref. [48]:

$$S_2(i,j) = S_2(j,i) = 1 - exp(-|DES(i) - DES(j)|), \tag{12}$$

normalized in the same way as in the graphical representation methods:

$$0 \le S_2 \le 1.$$

In this case, $S_2 = 0$ if the descriptors of two sequences are the same, i. e. if $DES(i) = DES(j)$.

Table 4
Similarity values [%] obtained using 3D-Dynamic Representation of DNA/RNA Sequences for SARS-CoV-2: MT106054 (Sa_1), MT159708 (Sa_2), MT192772 (Sa_3) and Embecovirus: AY391777 (Em_1), KM349744 (Em_2), FJ647223 (Em_3).

DES	Sequence	Sa_1	Sa_2	Sa_3	Em_1	Em_2	Em_3
r_1^{3D}	Sa_1	100	99.99	99.96	92.19	94.05	94.97
	Sa_2		100	99.97	92.20	94.05	94.97
	Sa_3			100	92.22	94.08	95.00
	Em_1				100	98.03	97.08
	Em_2					100	99.03
	Em_3						100
r_2^{3D}	Sa_1	100	99.99	99.96	92.28	94.02	94.90
	Sa_2		100	99.97	92.28	94.03	94.90
	Sa_3			100	92.31	94.06	94.93
	Em_1				100	98.15	97.24
	Em_2					100	99.08
	Em_3						100
r_3^{3D}	Sa_1	100	99.92	99.89	79.13	98.56	91.03
	Sa_2		100	99.98	79.20	98.64	90.96
	Sa_3			100	79.21	98.67	90.94
	Em_1				100	80.29	72.03
	Em_2					100	89.72
	Em_3						100

The recent outbreak of the Covid-19 pandemia stimulated a strong interest in the characterization of the virus [49–51]. In this work, we present a new general method of the description of biological sequences and demonstrate its applicability to the characterization of the 2019 novel coronavirus. In the calculations, all complete genome sequences of the 2019 novel coronavirus, Sarbecovirus, Embecovirus, Merbecovirus, Nobecovirus, and Hibecovirus, available in GenBank on May 25, 2020, have been used. For a comparison, we have also used all complete genome sequences of Deltacoronavirus available in GenBank in March 2020. Many sequences contain unknown nucleobases. Since the descriptors are different also for very similar sequences, the sequences containing one or more unknown nucleobases have been rejected. The sequence data used for the calculations are listed in Supplementary Materials.

Table 5
Similarity values [%] obtained using 2D-Dynamic Representation of DNA/RNA Sequences for SARS-CoV-2: MT106054 (Sa_1), MT159708 (Sa_2), MT192772 (Sa_3) and Embecovirus: AY391777 (Em_1), KM349744 (Em_2), FJ647223 (Em_3).

DES	Sequence	Sa_1	Sa_2	Sa_3	Em_1	Em_2	Em_3
r_1^{2D}	Sa_1	100	99.92	99.90	80.16	99.27	90.55
	Sa_2		100	99.98	80.23	99.35	90.48
	Sa_3			100	80.25	99.38	90.45
	Em_1				100	80.75	72.59
	Em_2					100	89.89
	Em_3						100
r_2^{2D}	Sa_1	100	99.98	99.88	56.11	43.41	40.13
	Sa_2		100	99.86	56.10	43.40	40.12
	Sa_3			100	56.18	43.46	40.18
	Em_1				100	77.37	71.53
	Em_2					100	92.45
	Em_3						100

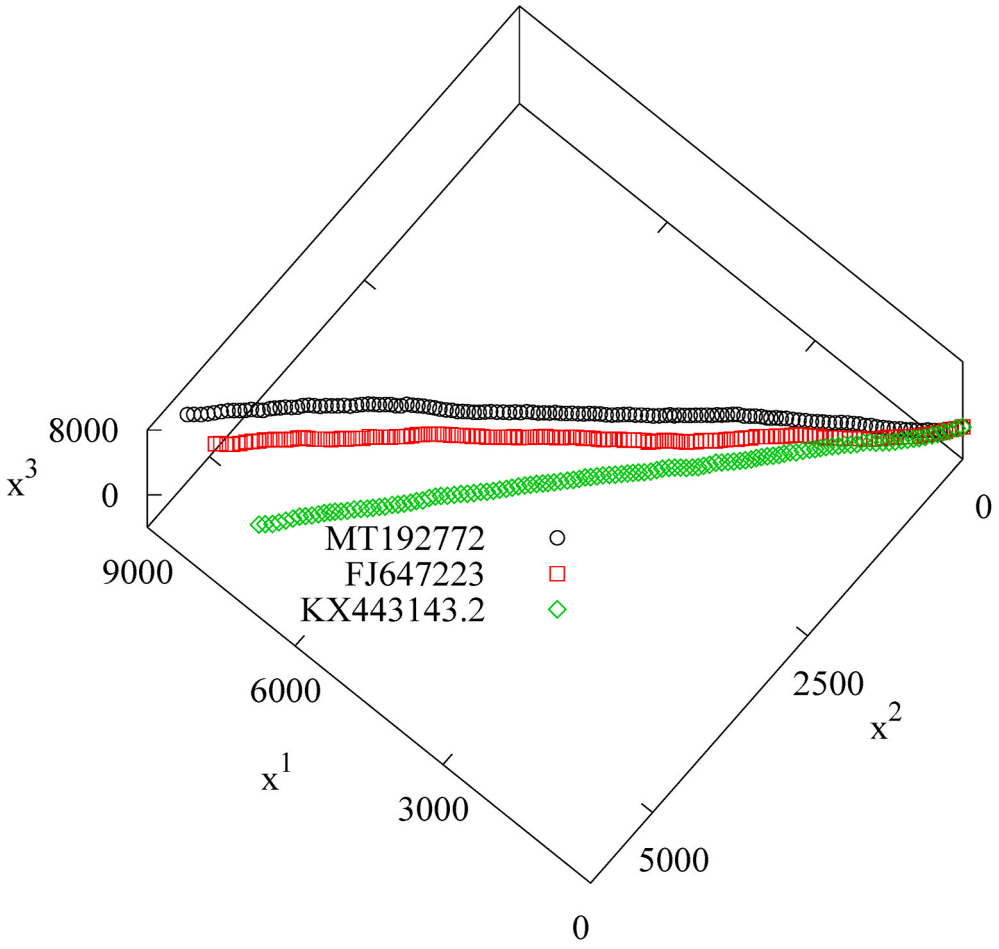

Fig. 1. $x^1x^2x^3$-graphs representing the complete genome sequences: FJ647223 (Embecovirus), MT192772 (SARS-CoV-2), and KX443143.2 (Deltacoronavirus).

3. Results and discussion

Fig. 1 shows examples of the 3D projections of the 4D-dynamic graphs, $x^1x^2x^3$-graphs. The lengths of the sequences are large (more than 25 thousands, see Supplementary Materials). Note that, nevertheless, one can clearly see differences between SARS-CoV-2, Deltacoronavirus, and Embecovirus graphs. The locations in the space of the graphs representing different SARS-CoV-2 sequences are similar – examples of seven 2D-graphs representing sequences of SARS-CoV-2 nearly overlap (Fig. 2). The graphs representing Embecovirus are different.

A convenient way to show similarities/dissimilarities of the sequences are 2D classification maps $DES_1 - DES_2$, or 3D classification maps $DES_1 - DES_2 - DES_3$, where $DES_1/DES_2/DES_3$ denote the descriptors represented in the axes. A sequence corresponding to a point located in the 2D map may be identified by a *classifier*, i.e. by the pair of descriptors corresponding to the coordinates of this very point (the first descriptor from the horizontal axis and the second one from the vertical axis). Analogously, a *classifier* in the 3D map is a triple of descriptors. If the points are close to each other, then the degree of similarity of the corresponding sequences, in the aspects described by the descriptors represented in the coordinate axes, is high. At the limit of very high similarity, the points nearly overlap. Figs. 3–7 show 3D classification maps based on the coordinates of the centers of mass and on the normalized four-dimensional moments of inertia (Eq. (10)): $\mu^k - \mu^l - \mu^m$; $k,l,m = 1,2,3,4; k \neq l \neq m$ (Fig. 3), $r_1^{4D} - r_4^{4D} - \mu^k$; $k = 1,2,3,4$ (Fig. 4), $r_2^{4D} - r_4^{4D} - \mu^k$; $k = 1,2,3,4$ (Fig. 5), and $r_3^{4D} - r_4^{4D} - \mu^k$; $k = 1,2,3,4$ (Fig. 6). The principal moments of inertia and the lengths of the sequences are shown in Supplementary Materials A.1–A.7.

Embecovirus, Sarbecobirus, SARS-CoV-2, Merbecovirus, Nobecovirus, Hibecovirus belong to the group of Betacoronaviruses. As we can see, the proposed descriptors correctly classify the sequences. The descriptors representing Deltacoronaviruses and Betacoronaviruses are located in different parts of all the maps. A pair of sequences can be similar in one aspect of similarity and different in another aspect. Each aspect is reflected by a particular descriptor. Some similarities within the sequences of Betacoronoviruses may be seen in the classification maps - the corresponding descriptors overlap. In particular, in Fig. 3, the points representing SARS-CoV-2 overlap with the points representing Sarbecovirus, but they are also close to Merbecovirus and Hibecovirus (top, right panel) and to Embecovirus (bottom, left panel). Analogously, in Figs. 4–6, SARS-CoV-2 overlaps with Sarbecovirus, but it is also close to Merbecovirus and Nobecovirus (top, bottom, right panels).

Let us focus on the overlapping of the points representing the sequences of Sarbecovirus and of SARS-CoV-2, which in fact have been already classified as Sarbecovirus. Fig. 7 is an enlargement of Fig. 3. The ranges of the maps are chosen so to display all the sequences of the 2019 novel coronavirus. In fact, only five points representing Sarbecovirus sequences are located close to SARS-CoV-2. The closest point represents the full-length genome sequence of bat coronavirus RaTG13 (MN996532). According to the classification maps, this sequence together with the sequences of SARS-CoV-2 can be considered as a separate group within Betacoronaviruses. No other points overlap with these clusters of points. The second group of points located close to SARS-CoV-2 (but not so close as the sequence of bat coronavirus) represents four full-length genome sequences of pangolin coronavirus (MT040333, MT040334, MT040335, MT040336).

In recent bioinformatics studies related to the 2019 novel

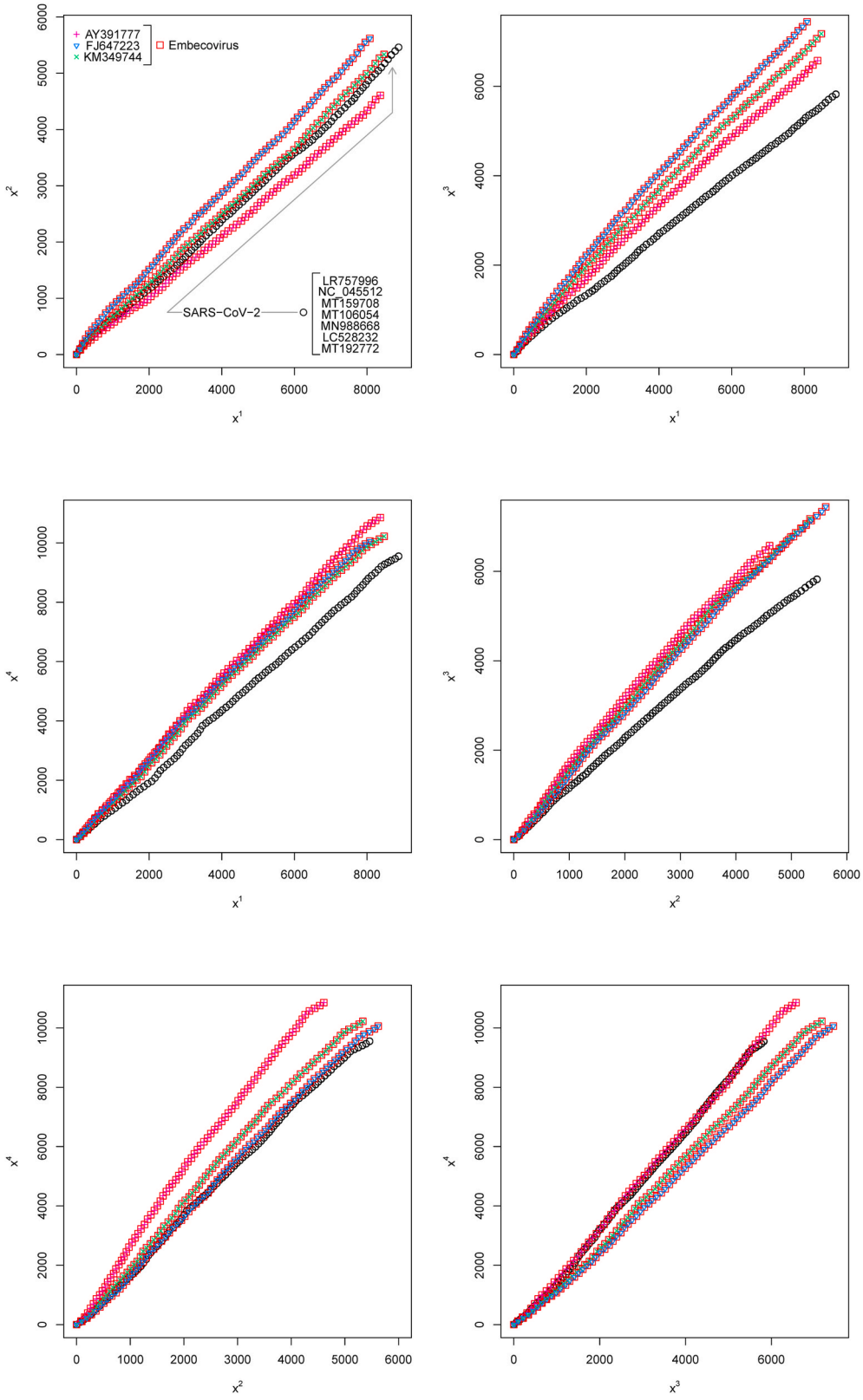

Fig. 2. $x^k x^l$ -graphs ($k, l = 1, 2, 3, 4; k \neq l$) representing selected complete genome sequences.

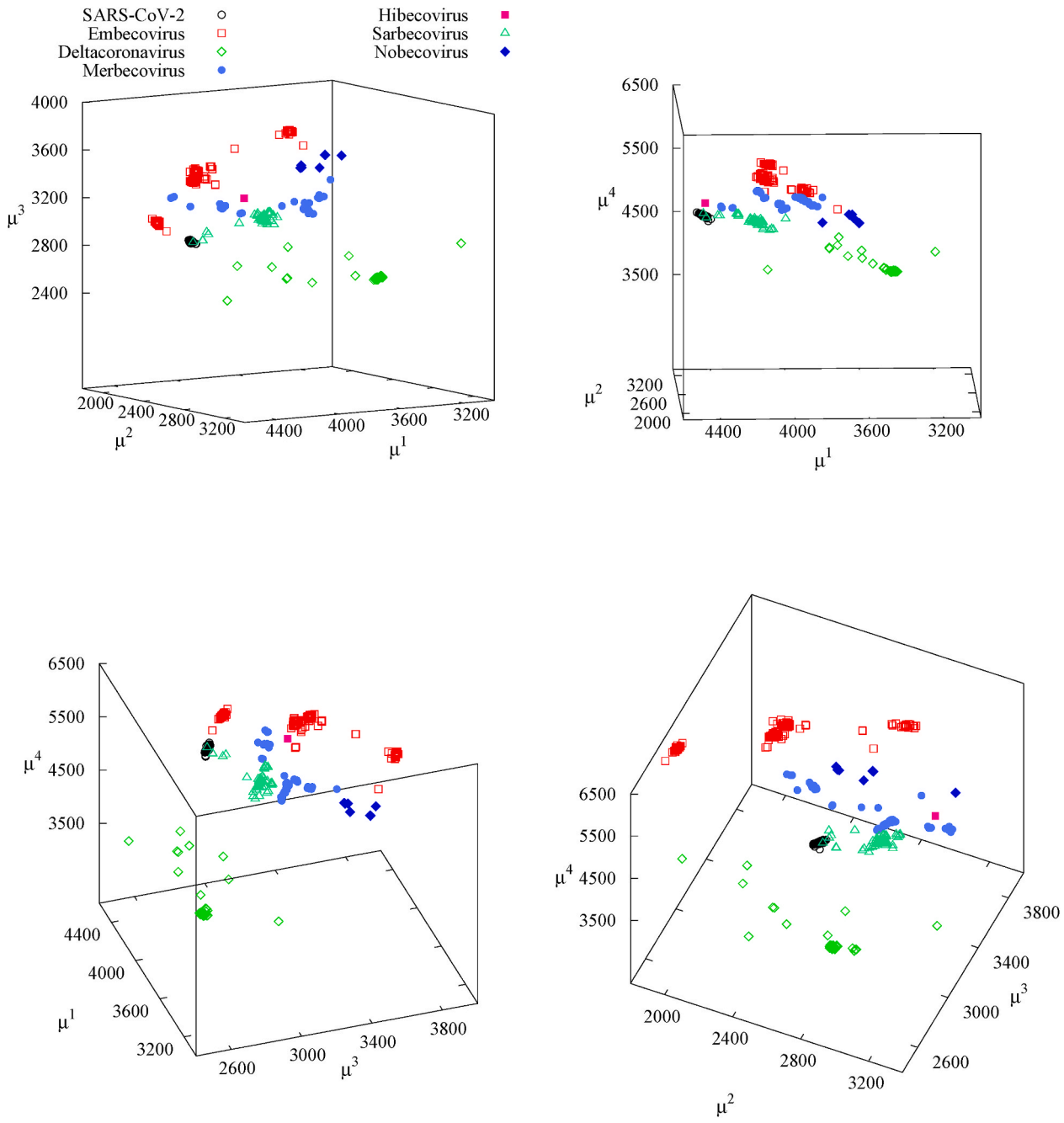

Fig. 3. Classification maps $\mu^k - \mu^l - \mu^m$ $(k,l,m = 1,2,3,4; k \neq l \neq m)$.

coronavirus, four complete genome sequences of Embecovirus are considered: AY391777, KM349744, FJ647223, and MK167038 [49–51]. Since the sequence MK167038 contains an unknown nucleobase, it is rejected from the consideration. The corresponding sequence data are shown in Supplementary Materials A.3.

Some details of similarity/dissimilarity between these three sequences and three randomly selected sequences of SARS-CoV-2 are shown in 2D classification maps (Fig. 8). The differences between the two groups of sequences are reflected by all descriptors taken into account. In all maps the border lines are explicitly shown.

A numerical representation of similarity/dissimilarity between three sequences of Embecovirus and three sequences of SARS-CoV-2 is shown in similarity/dissimilarity matrices (Tables 3–5). In all these matrices, the similarity values are obtained using the similarity measure defined in Eq. (11). In all these matrices the same sequences are used. The data collected in the tables correspond to several matrices. In the diagonals self-similarity values, i.e. 100% are inserted. Since the matrices are symmetric, only upper triangles are displayed.

In Table 3, the results given by the present method are shown, i.e. $DES \equiv r_k^{4D}$, where $k = 1$ for the first matrix, $k = 2$ for the second one, and so on. As we can see in Fig. 8, the largest differences between the descriptors representing the sequences of Embecovirus are for r_4^{4D}. This fact is numerically reflected in the values of matrix elements for different k. The differences between the similarity values are small for $k = 1, 2, 3$. The largest differences between the similarity values are displayed in the last matrix of Table 3 ($k = 4$).

Tables 4 and 5 show similarity/dissimilarity matrices obtained using 3D-Dynamic Representation of DNA/RNA Sequences and 2D-Dynamic Representation of DNA/RNA Sequences, respectively [39,40]. As the descriptors, the normalized moments of inertia of 3D-dynamic graphs

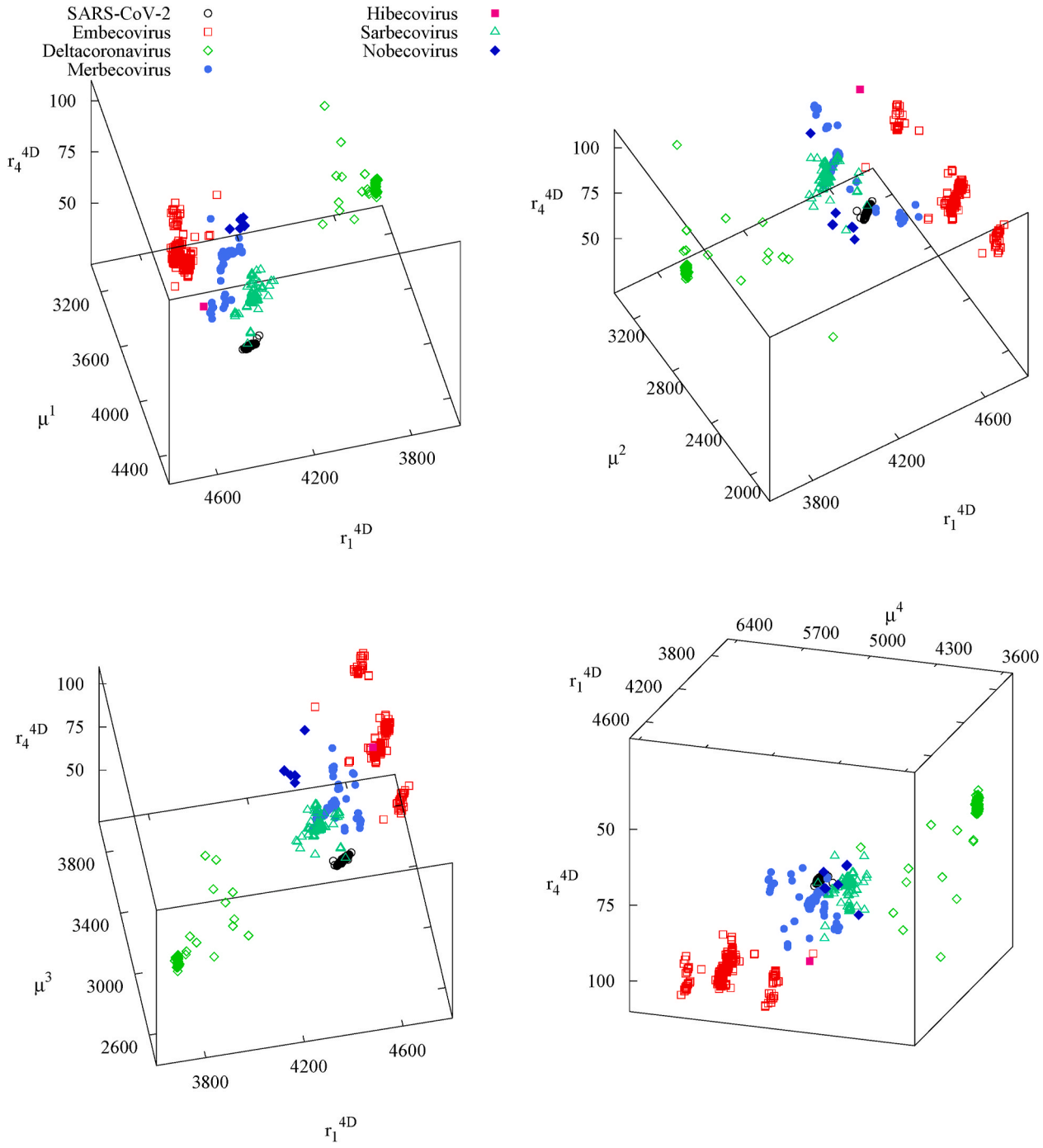

Fig. 4. Classification maps $r_1^{4D} - r_4^{4D} - \mu^k$ $(k = 1, 2, 3, 4)$.

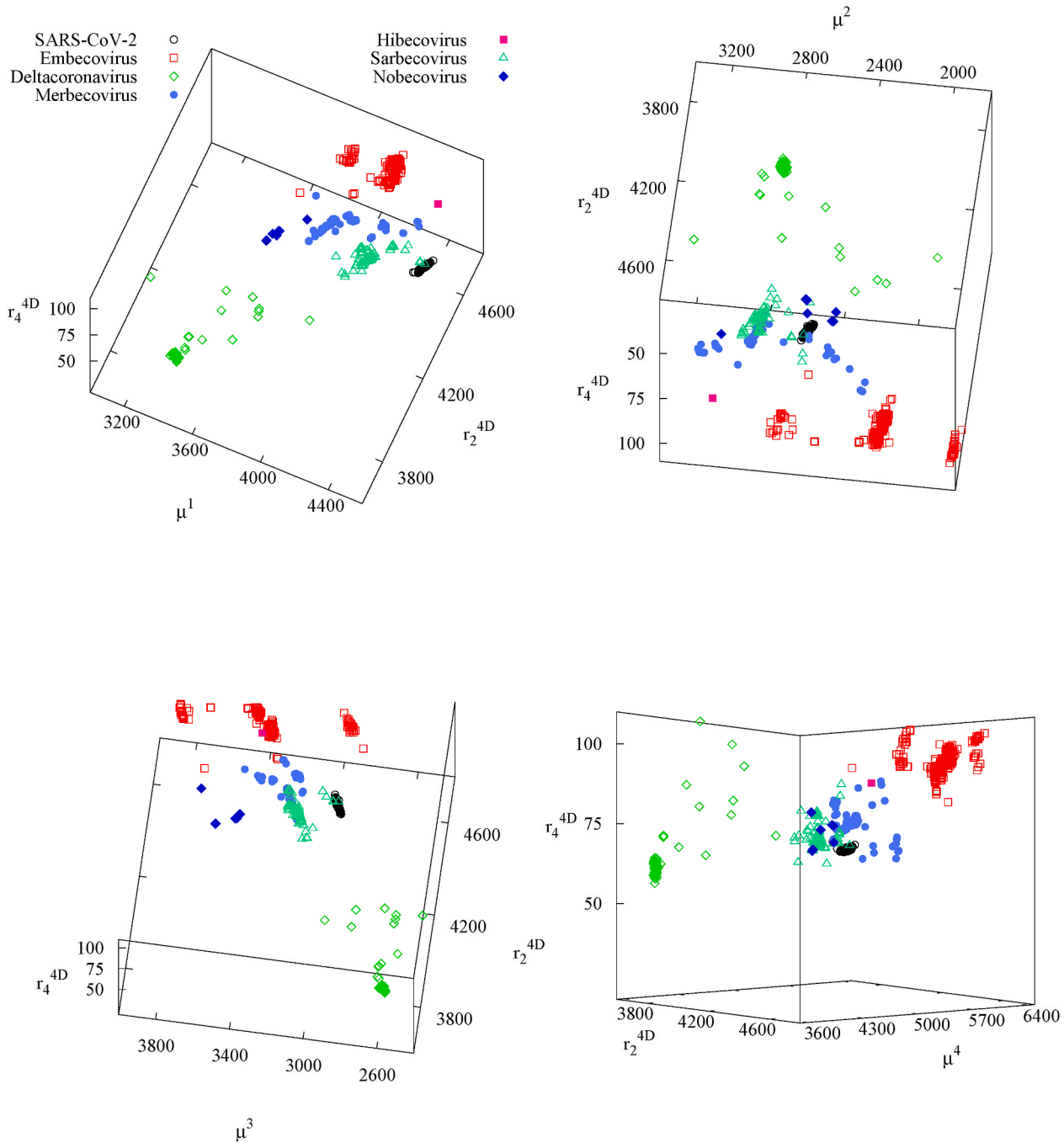

Fig. 5. Classification maps $r_2^{4D} - r_4^{4D} - \mu^k$ $(k = 1, 2, 3, 4)$.

$(r_k^{3D}, k = 1, 2, 3)$ and of 2D-dynamic graphs $(r_k^{2D}, k = 1, 2)$ are taken.

For a pair of sequences SARS-CoV-2 the similarity values are close to 100%: $S_1(MT106054, MT159708) = 99.99\%$ for r_1^{4D} and for r_1^{3D} and $S_1(MT106054, MT159708) = 99.92\%$ for r_1^{2D}. The diversity of the applied methods is essential in the detailed similarity studies - different methods expose different aspects of similarity. The 2D-Dynamic Representation of DNA/RNA Sequences is based on shifts in a 2D space, analogously as it was proposed in the Nandy plots [18]. The nucleobases are represented by the basis vectors: A = (−1,0), G = (1,0), C = (0,1) and T/U= (0,-1). In the Nandy plots, if a walk is performed back and forth along the same trace, then some parts of the sequence are hidden. As a consequence, different sequences are represented by the same plot. In order to remove this degeneracy coming from the repetitive walks, in the 2D-Dynamic Representation of DNA/RNA Sequences we introduced masses different than 1, if the ends of the vectors meet several times at the same point. The accuracy of 2D-dynamic graphs is larger than that of the Nandy plots, but still one axis contains a combined information about two nucleobases. In 3D-Dynamic Representation of DNA/RNA Sequences, the information about four nucleobases is combined to three directions. The corresponding moments of inertia are calculated for simplified “rigid bodies” both in 2D and 3D methods. In the present method, we have more options. We can either consider four components of the descriptors separately or to combine the information coming from different sources. More details are given in Ref. [14].

Summarizing, the problem of similarity is multidimensional. A pair of sequences can be very similar in one aspect and simultaneously very different in some other aspects. It may happen that finding a difference for very similar sequences or a similar aspect for very different se-

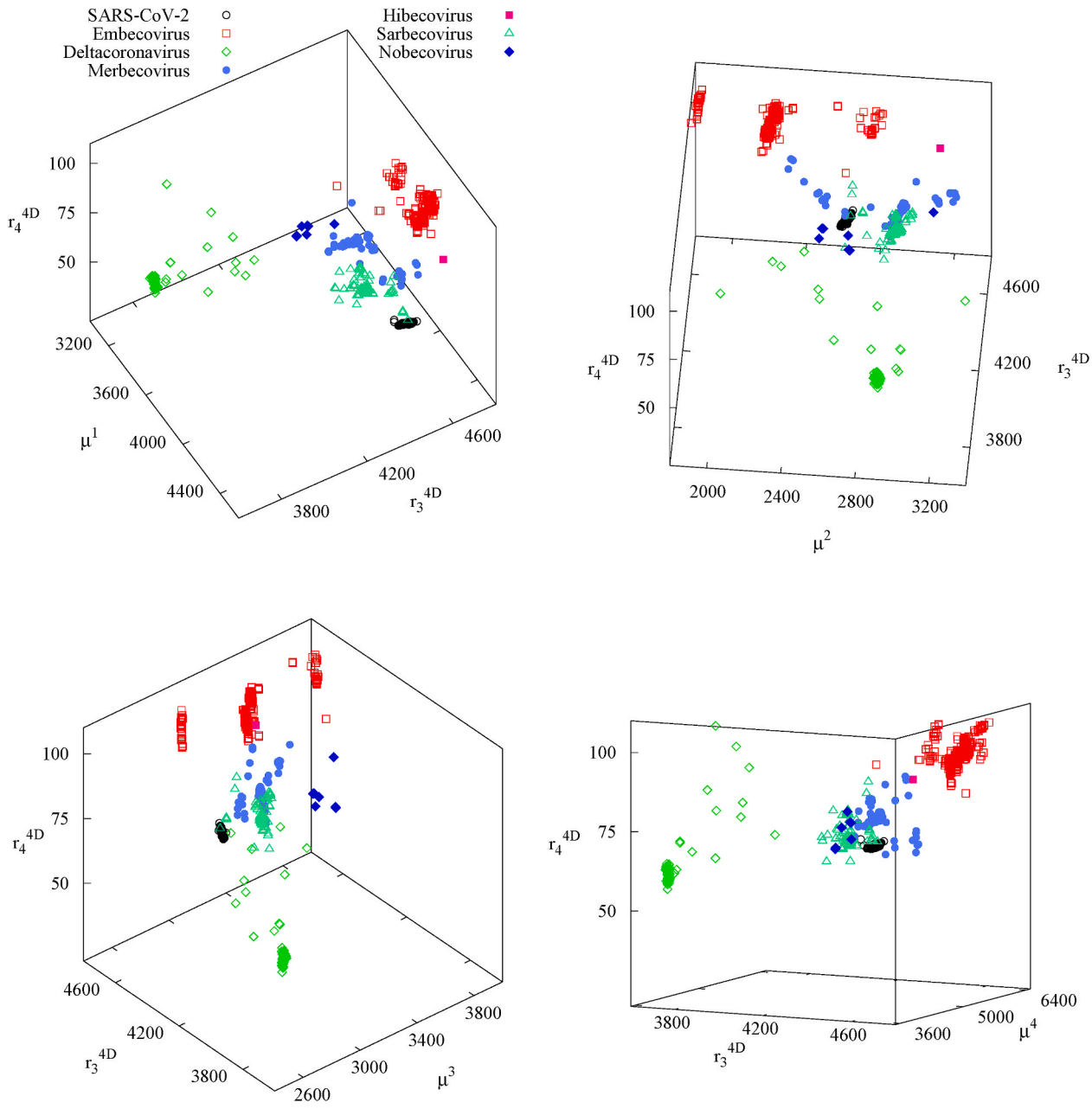

Fig. 6. Classification maps $r_3^{4D} - r_4^{4D} - \mu^k$ $(k = 1, 2, 3, 4)$.

quences may be an important classifier. Different aspects of similarity may be revealed by different descriptors. As an additional point to the present studies, we visualize the problem of multidimensionality of the similarity using the sequences most commonly used in the literature by the authors introducing new graphical representation methods. This choice seems to be very intuitive: the first exons of β-globin gene of different species. The problem of the multidimensionality of the similarity of complex objects has already been discussed by us [32,48] and by other authors [52]. In Fig. 9, we compare the results derived from the present method using the similarity measure S_2 and $DES = r_4^{4D}$ (red line), with results obtained by several other methods [31,48,53–58]. In this figure the similarity values human-other species for 11 species listed in Supplementary Materials A.8 and labeled by index j, are shown. The similarity value human-human, corresponding to $j = 1$, is equal to 0. The similarity values are normalized so that they are equal to 1 for $j = 4$, i.e. for human-gallus. The border line $S = 1$ corresponds to the similarity value for these two species. The larger are the similarity values, the smaller is the degree of similarity. As one can see, the present method, follows reasonably well the intuitive expectation: the smallest degree of similarity is for the only non-mammalian species, i.e. human-gallus. Since each of the considered methods describes different aspects of similarity, the results are also different, and often do not meet our intuitive expectations.

As a final point of the studies, we would like to discuss some applications of the alignment-free methods in biomedical sciences. These methods can be applied to all problems which require similarity/dissimilarity analysis of the sequences, in particular to the mutation analysis, to the identification of protein coding regions, to the construction of phylogenetic trees using new descriptors and new similarity measures. For example, a phylogenetic analysis has been performed using graphical bioinformatics method "3DD-curve" [59] and new three-dimensional graphical representations of DNA sequences have been proposed and applied to phylogenetic analysis [30,60]. It has also been shown that a method of comparison of protein sequences,

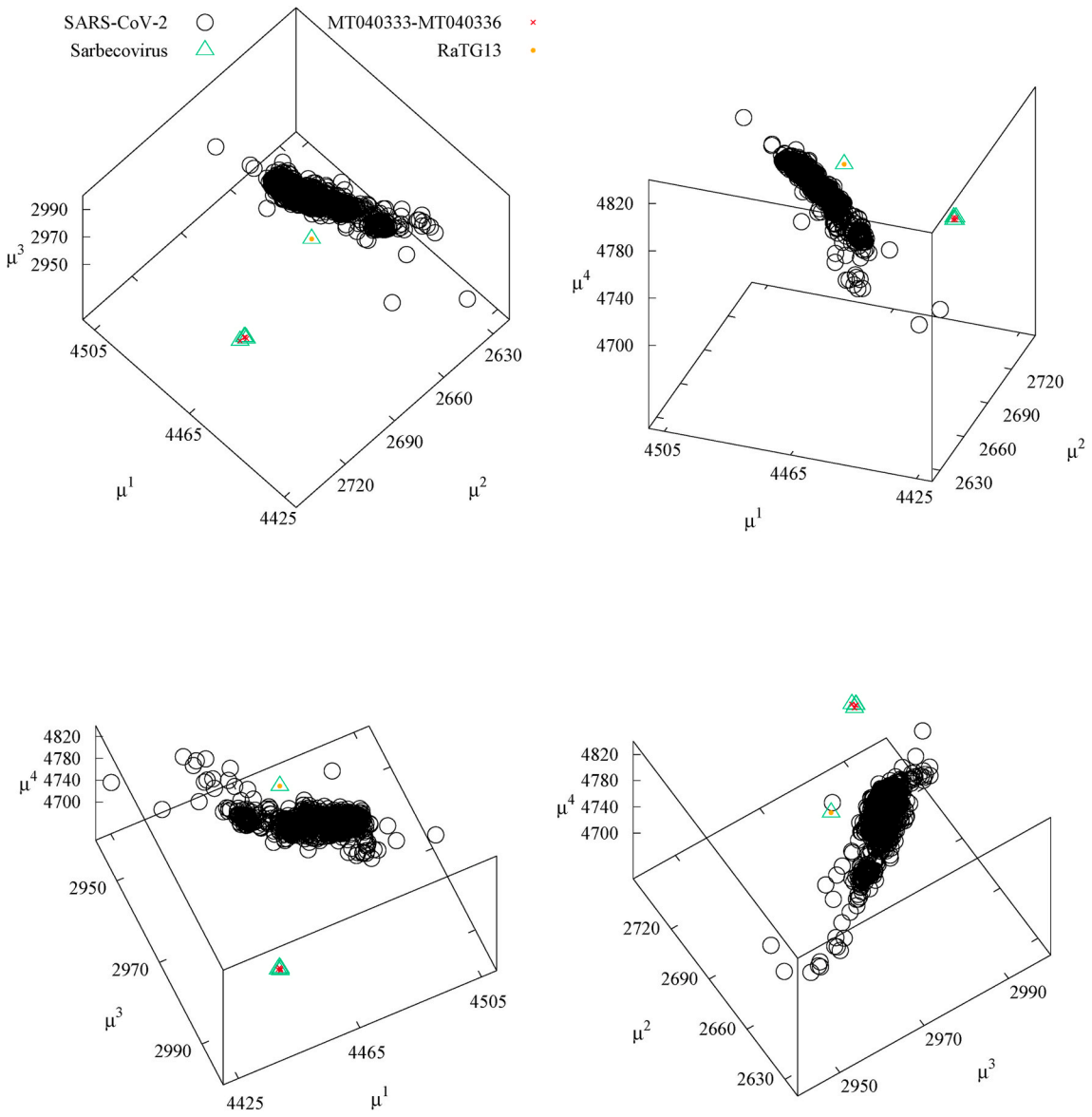

Fig. 7. An enlargement of Fig. 3.

introduced by us, is a convenient tool for the creation of phylogenetic trees [45]. Huang and Wang created a mathematical model describing different kinds of mutations: substitutions, insertions and deletions [60]. They demonstrated a high quality of this approach using a model sequence: by the removing guanine and thymine from the initial sequence they got examples of deletions and by the substitution of adenine by guanine – an example of substitution.

Another interesting aspect of approaches aimed at the description of single sequences, is a possibility of tracking the time evolution of these sequences. Our recent studies have shown a correlation between the values of descriptors of the 2D-Dynamic method representing the complete genome sequences of the Zika virus and time (for the years of the collection 1947 through 2015) [61]. The dates of collection of the 2019 novel coronavirus cover a period of two years only (2019 and 2020). Then, it is too early to study the time evolution of the genome sequences of SARS-CoV-2. Therefore, we present an application of the descriptors of the 4D-Dynamic Representation of DNA/RNA Sequences to the studies of the time evolution of the genome sequences of the Zika virus. For the calculations, the same sequence data have been used as in our previous work [61]. The results are shown in Table 6. The coordinates of the centers of mass of the graphs, μ^k (Eq. (1)), and the eigenvectors ω_k^l, $k, l = 1, 2, 3, 4$, (Eq. (8)), have been chosen as the descriptors. Three kinds of correlation measures have been considered: Pearson, Spearman, and Kendall. Only statistically significant (p-value smaller than 0.05) results for the considered descriptors are shown in the Table. As one can see, in many cases the correlations are strong (negative or positive). Then, the proposed method is also applicable to the studies of time evolution of genome sequences of viruses.

4. Conclusions

Nowadays, we live in difficult times of pandemic. COVID-19 disease spreads to the whole globe [62–64]. We still do not know the scenario for the next months. Will it be eventually overcome as, for example, Acute Respiratory Syndrome (SARS) in 2002 [65]? Therefore, studies aimed at a detailed characterization of the 2019 novel coronavirus, are crucial. In particular, developing alignment-free methods with broad applications including vaccine design is an important task [66,67].

The present approach shows, in a nonstandard way, some aspects of similarity of the genome sequences of viruses. As a result, we obtained 2D and 3D similarity maps in which points representing particular groups of sequences cluster. One of the advantages of this approach is,

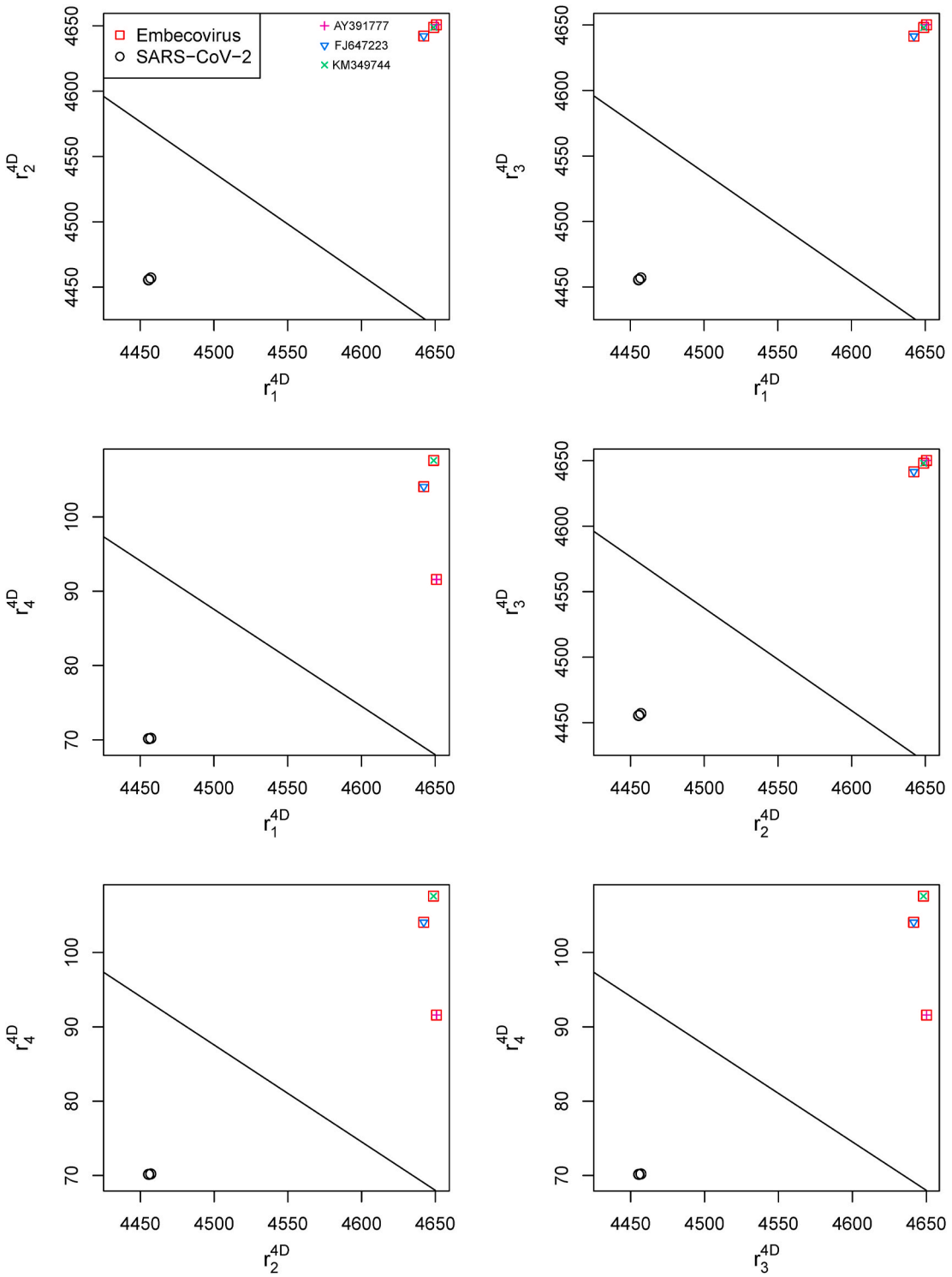

Fig. 8. Classification maps $r_k^{4D} - r_l^{4D}$ $(k, l = 1, 2, 3, 4; k \neq l)$ for the complete genome of 6 sequences used in the calculations and collected in Tables 3–5.

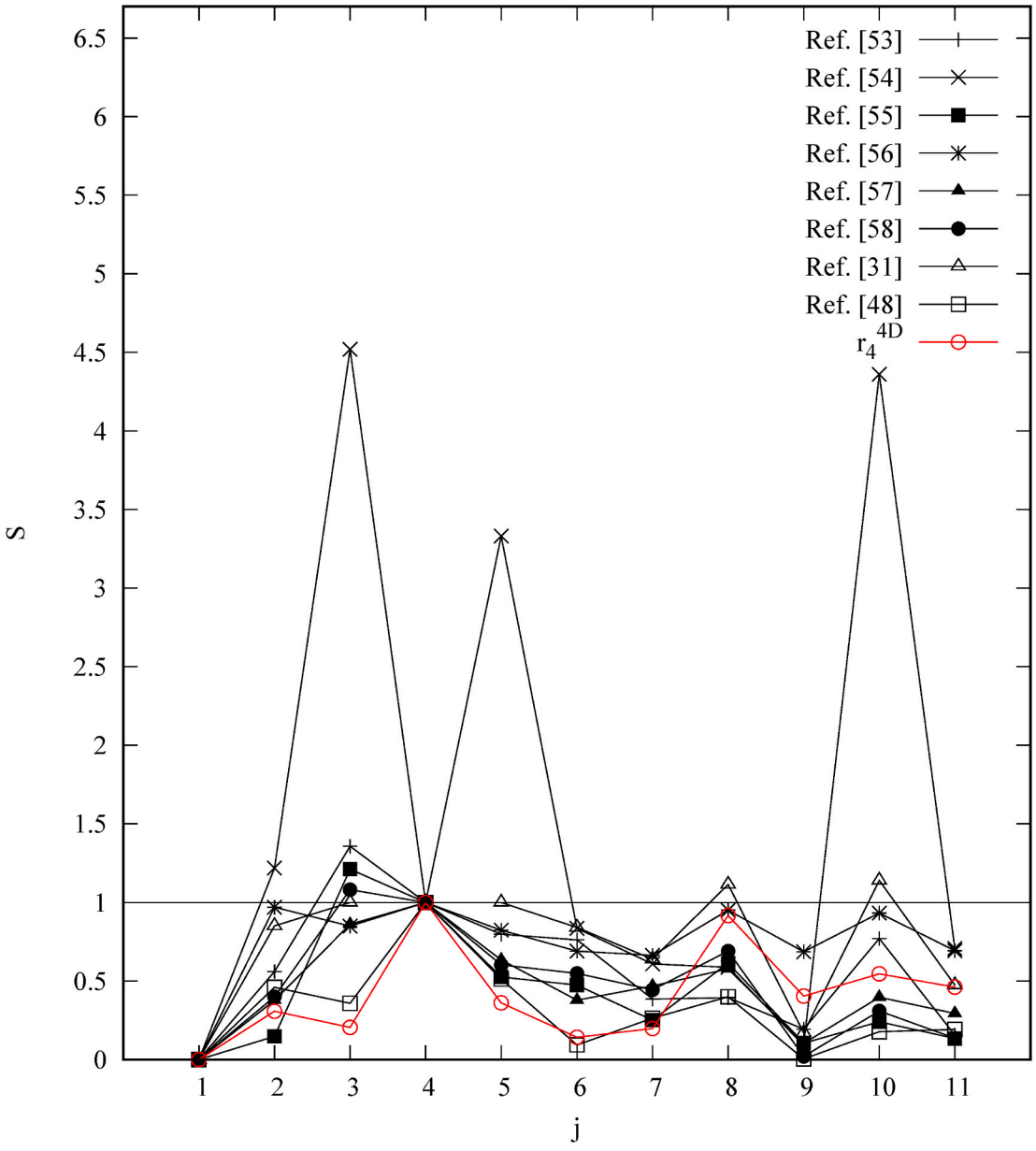

Fig. 9. Similarity values for the first exons of β-globin gene of human-other species using different methods. Integers j in the horizontal axis stand for the labels of different species.

that the axes of the similarity maps have simple interpretation – they represent the descriptors calculated using analytical expressions characterizing the 4D-dynamic graphs. The method is not time-consuming and may deal with large-scale sequence data. The reduction of the dimension from four to two and three is necessary to obtain a visualization. Four-dimensional objects represented in lower dimensional space may be degenerate, but the full set of all projections constitutes a complete, non-degenerate, information about the considered system. The sets of coordinates of points in the maps, i.e. the descriptors of the sequences, are the classifiers. In the presented maps, the points representing the complete genome sequences of SARS-CoV-2 are located close to the ones of Sarbecovirus.

According to some classification schemes, all sequences of SARS-CoV-2 and one sequence of bat coronavirus RaTG13 can be defined as a separate group within the family of Betacoronaviruses (Fig. 7). A similar analysis, performed in Ref. [68], provides some evidence that 2019 novel coronavirus may have originated in bats. The four points of the second group, located in the maps close to SARS-CoV-2 (in Fig. 7), represent the full-length genome sequences of pangolin coronavirus. This result supports the hypothesis on the pangolin origin of SARS-CoV-2, formulated in Ref. [69].

The present method, the 4D-Dynamic Representation of DNA/RNA Sequences, is a generalization of our previous 2D and 3D approaches. In the present approach, four components of the descriptors are considered instead of two and three, respectively. Splitting the information to a larger set of the components gives an opportunity to analyze separately more aspects of similarity/dissimilarity of the sequences and to extract these components which may be correlated with the considered

Table 6
Correlations of the descriptors of 4D-dynamic graphs with time for the complete genome sequences of Zika virus.

Descriptor	Correlation coefficient	Method	p-value
μ^2	0.655	Spearman	0.0110
μ^2	0.479	Kendall	0.0247
ω_1^1	−0.793	Pearson	0.0007
ω_1^1	−0.678	Spearman	0.0077
ω_1^1	−0.503	Kendall	0.0183
ω_1^2	0.781	Pearson	0.0010
ω_1^2	0.703	Spearman	0.0050
ω_1^2	0.479	Kendall	0.0247
ω_1^3	−0.598	Spearman	0.0240
ω_1^4	−0.794	Pearson	0.0007
ω_1^4	−0.754	Spearman	0.0018
ω_1^4	−0.601	Kendall	0.0048
ω_2^1	−0.533	Spearman	0.0496
ω_2^2	0.750	Pearson	0.0020
ω_2^2	0.818	Spearman	0.0004
ω_2^2	0.699	Kendall	0.0010
ω_2^3	−0.790	Pearson	0.0008
ω_2^3	−0.703	Spearman	0.0050
ω_2^3	−0.479	Kendall	0.0247
ω_2^4	0.748	Pearson	0.0021
ω_2^4	0.795	Spearman	0.0007
ω_2^4	0.601	Kendall	0.0048
ω_3^1	0.671	Spearman	0.0086
ω_3^1	0.454	Kendall	0.0332
ω_3^2	−0.608	Pearson	0.0211
ω_3^2	−0.683	Spearman	0.0071
ω_3^2	−0.454	Kendall	0.0332
ω_3^3	−0.584	Spearman	0.0284

variables in the biomedical problems, including time (Table 6) and clinical features of the novel coronavirus pneumonia caused by SARS-CoV-2.

Declaration of competing interest

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.compbiomed.2021.104247.

References

[1] J. Zhou, P.Y. Zhong, T.H. Zhang, A novel method for alignment-free DNA sequence similarity analysis based on the characterization of complex networks, Evol. Bioinf. Online 12 (2016) 229–235.
[2] A.K. Saw, G. Raj, M. Das, N.C. Talukdar, B.C. Tripathy, S. Nandi, Alignment-free method for DNA sequence clustering using Fuzzy integral similarity, Scientific Reports 9 (2019) 3753.
[3] D. Lichtblau, Alignment-free genomic sequence comparison using FCGR and signal processing, BMC Bionformatics 20 (2019) 742.
[4] L.L. He, R. Dong, R.L. He, S.S.T. Yau, A novel alignment-free method for HIV-1 subtype classification, Infect. Genet. Evol. 77 (2020) 104080.
[5] S. Vinga, J. Almeida, Alignment-free sequence comparison-a review, Bioinformatics 19 (2003) 513–523.
[6] X. Jin, Q. Jiang, Y. Chen, S.J. Lee, R. Nie, S. Yao, D. Zhou, K. He, Similarity/dissimilarity calculation methods of DNA sequences: a survey, J. Mol. Graph. Model. 76 (2017) 342–355.
[7] A. Bielińska, M. Majkowicz, D. Bielińska-Wąż, P. Wąż, Classification studies in various areas of science, in: G. Nikolov, N. Kolkovska, K. Georgiev (Eds.), Numerical Methods and Applications, NMA 2018,, Lecture Notes in Computer Science, vol. 11189, Springer, Cham, 2019, pp. 326–333.
[8] A. Bielińska, M. Majkowicz, P. Wąż, D. Bielińska-Wąż, Mathematical modeling: interdisciplinary similarity studies, in: G. Nikolov, N. Kolkovska, K. Georgiev (Eds.), Numerical Methods and Applications, NMA 2018, Lecture Notes in Computer Science vol. 11189, Springer, Cham, 2019, pp. 334–341.
[9] P. Wąż, D. Bielińska-Wąż, A. Pleskacz, A. Strobel, Identification of stellar spectra using methods of statistical spectroscopy, Acta Phys. Pol. B 39 (2008) 1993–2001.
[10] D. Bielińska-Wąż, P. Wąż, S.C. Basak, Statistical theory of spectra: statistical moments as descriptors in the theory of molecular similarity, Eur. Phys. J. B 50 (2006) 333–338.
[11] D. Bielińska-Wąż, W. Nowak, Ł. Pepłowski, P. Wąż, S.C. Basak, R. Natarajan, Statistical spectroscopy as a tool for the study of molecular spectroscopy, J. Math. Chem. 43 (2008) 1560–1572.
[12] A. Bielińska, D. Bielińska-Wąż, P. Wąż, Classification maps in studies on the retirement threshold, Appl. Sci. 10 (2020) 1282.
[13] P. Wąż, D. Bielińska-Wąż, Asymmetry coefficients as indicators of Chaos, Acta Phys. Pol., A 116 (2009) 987–991.
[14] D. Bielińska-Wąż, S. Subramaniam, Classification studies based on a spectral representation of DNA, J. Theor. Biol. 266 (2010) 667–674.
[15] E. Hamori, J. Ruskin, H Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (1983) 1318–1327.
[16] E. Hamori, Novel DNA sequence representations, Nature 314 (1985) 585–586.
[17] M.A M.A. Gates, Simpler DNA sequence representations, Nature 316 (1985) 219.
[18] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I: methodology and application to globin genes, Curr. Sci. 66 (1994) 309–314.
[19] P.M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, Comput. Appl. Biosci. 11 (1995) 503–507.
[20] C.T. Zhang, R. Zhang, H.Y. Ou, The Z curve database: a graphic representation of genome sequences, Bioinformatics 19 (2003) 59–599.
[21] C. Li, J. Wang, On a 3-D representation of DNA primary sequences, comb. Chem. High T. Scr. 7 (2004) 2–27.
[22] B. Liao, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett. 388 (2004) 195–200.
[23] Y. Yao, X. Nan, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, Chem. Phys. Lett. 411 (2005) 248–255.
[24] Z. Cao, B. Liao, R. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, Int. J. Quant. Chem. 108 (2008) 1485–1490.
[25] I. Pesek, J. Žerovnik, A numerical characterization of modified Hamori curve representation of DNA sequences, MATCH Commun, Math. Comput. Chem. 60 (2008) 301–312.
[26] W. Chen, B. Liao, X. Xiang, W. Zhu, An improved binary representation of DNA sequences and its applications, MATCH commun, Math. Comput. Chem. 61 (2009) 767–780.
[27] Z. Cao, R. Li, W. Chen, A 3D graphical representation of DNA sequence based on numerical coding method, Int. J. Quant. Chem. 110 (2010) 975–985.
[28] J.F. Yu, J.H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, MATCH Commun. Math. Comput. Chem. 63 (2010) 493–512.
[29] Y. Li, Y. Qin, X. Zheng, Y. Zhang, Three-unit semicircles curve: a compact 3D graphical representation of DNA sequences based on classifications of nucleotides, Int. J. Quant. Chem. 112 (2012) 2330–2335.
[30] Y. Yang, Y. Zhang, M. Jia, C. Li, L. Meng, Non-degenerate graphical representation of DNA sequences and its applications to phylogenetic analysis, Comb. Chem. High Throughput Screen. 16 (2013) 585–589.
[31] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, Math. Biosci. 214 (2013) 217–224.
[32] D. Bielińska-Wąż, P. Wąż, Spectral-dynamic representation of DNA sequences, J. Biomed. Inf. 72 (2017) 1–7.
[33] D. Bielińska-Wąż, Graphical and numerical representations of DNA sequences: statistical aspects of similarity, J. Math. Chem. 49 (2011) 2345–2407.
[34] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, Int. J. Quant. Chem. 113 (2013) 2413–2446.
[35] C. Raychaudhury, A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, J. Chem. Inf. Comput. Sci. 39 (1999) 243–247.
[36] M. Randić, M. Vračko, A. Nandy, S.C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci. 40 (2000) 1235–1244.
[37] G. Agüero-Chapin, A. Sánchez-Rodríguez, P.I. Hidalgo-Yanes, Y. Pérez-Castillo, R. Molina-Ruiz, K. Marchal, V. Vasconcelos, A. Antunes, An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference, PloS One 6 (2011), e26638.
[38] G. Agüero-Chapin, D. Galpert, R. Molina-Ruiz, E. Ancede-Gallardo, G. Pérez-Machado, G.A. De la Riva, A. Antunes, A. Graph theory-based sequence descriptors as remote homology predictors, Biomolecules 10 (2020) 26.
[39] D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. 442 (2007) 140–144.
[40] P. Wąż, D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (2014) 2141.
[41] V. Aram, A. Iranmanesh, 3D-dynamic representation of DNA sequences, MATCH Commun, Math. Comput. Chem. 67 (2012) 809–816.
[42] X.C. Tang, P.P. Zhou, W.Y. Qiu, On the similarity/dissimilarity of DNA sequences based on 4D graphical representation, Chin. Sci. Bull. 55 (2010) 701–704.
[43] C.J. Tan, S.S. Li, P. Zhu, 4D Graphical representation research of DNA sequences, Int. J. Biomath. (IJB) 8 (2017) 1550004.
[44] B. Liao, R. Li, W. Zhu, X. Xiang, On the similarity of DNA primary sequences based on 5-D representation, J. Math. Chem. 42 (2007) 47–57.

[45] A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16–23.
[46] W. Yang, Q. Cao, L. Qin, X. Wang, Z. Cheng, A. Pan, J. Dai, Q. Sun, F. Zhao, J. Qu, F. Yan, Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): a multi-center study in Wenzhou city, Zhejiang, China, J. Infect. 80 (2020) 388–393.
[47] Y.H. Xu, J.H. Dong, W.M. An, X.Y. Lv, X.P. Yin, J.Z. Zhang, L. Dong, X. Ma, H. J. Zhang, B.L. Gao, Clinical and computed tomographic imaging features of novel coronavirus pneumonia caused by SARS-CoV-2, J. Infect. 80 (2020) 394–400.
[48] P. Wąż, D. Bielińska-Wąż, Non-standard similarity/dissimilarity analysis of DNA sequences, Genomics 104 (2014) 464–471.
[49] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, Y.J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E. C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (2020) 565–574.
[50] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G.F. Gao, W. Tan, China novel coronavirus investigating and research team. A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733.
[51] W.B. Park, N.J. Kwon, S.J. Choi, C.K. Kang, P.G. Choe, J.Y. Kim, J. Yun, G.W. Lee, M.W. Seong, N.J. Kim, J.S. Seo, M.D. Oh, Virus isolation from the first patient with SARS-CoV-2 in korea, J. Kor. Med. Sci. 53 (2020) e84.
[52] Z. Mo, W. Zhu, Y. Sun, Q. Xiang, M. Zheng, M. Chen, Z. Li, One novel representation of DNA sequence based on the global and local position information, Sci. Rep. 8 (2018) 7592.
[53] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202–207.
[54] R. Chi, K. Ding, Novel 4D numerical representation of DNA sequences, Chem. Phys. Lett. 407 (2005) 63–67.
[55] Q. Dai, X. Liu, T. Wang, A novel graphical representation of DNA sequences and its application, J. Mol. Graph. Model. 25 (2006) 340–344.
[56] Y.-Z. Liu, T. Wang, Related matrices of DNA primary sequences based on triplets of nucleic acid bases, Chem. Phys. Lett. 417 (2006) 173–178.
[57] B. Liao, Q. Xiang, L. Cai, Z. Cao, A new graphical coding of DNA sequence and its similarity calculation, Physica A 392 (2013) 4663–4667.
[58] X. Yang, T. Wang, Linear regression model of short k-word: a similarity distance suitable for biological sequences with various lengths, J. Theor. Biol. 337 (2013) 61–70.
[59] Y. Zhang, W. Chen, A new approach to molecular phylogeny of PrimateMitochondrial DNA, MATCH Commun. Math. Comput. Chem. 59 (2008) 625–634.
[60] Y. Huang, T. Wang, New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis, Int. J. Quant. Chem. 112 (2012) 1746–1757.
[61] D. Panas, P. Wąż, D. Bielińska-Wąż, A. Nandy, S.C. Basak, 2D-Dynamic representation of DNA/RNA sequences as a characterization tool of the Zika virus genome, MATCH Commun. Math. Comput. Chem. 77 (2017) 321–332.
[62] D.F. Cuadros, Y.Y. Xiao, Z. Mukandavire, E. Correa-Agudelo, A. Hernandez, H. Kim, N.J. MacKinnon, Spatiotemporal transmission dynamics of the COVID-19 pandemic and its impact on critical healthcare capacity, Health Place 64 (2020) 102404.
[63] M.J. Miller, J.R. Loaiza, A. Takyar, R.H. Gilman, COVID-19 in Latin America: novel transmission dynamics for a global pandemic? PLoS Neglected Trop. Dis. 14 (2020), e0008265.
[64] K. Wu, D. Darcet, Q. Wang, D. Sornette, Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world, Nonlinear Dynam. 101 (2020) 1561–1581.
[65] R.M. Anderson, C. Fraser, A.C. Ghani, C.A. Donnelly, S. Riley, N.M. Ferguson, Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic, Phil. Trans. Roy. Soc. Lond. B 359 (2004) 1091–1105.
[66] A. Nandy, S.C. Basak, Prognosis of possible reassortments in recent H5N2 epidemic influenza in USA: implications for computer-assisted surveillance as well As drug/vaccine design, Curr. Comput. Aided Drug Des. 11 (2015) 110–116.
[67] A. Nandy, S. Manna, S.C. Basak, Computational methodology for peptide vaccine design for Zika virus: a bioinformatics approach, in: Namrata Tomar (Ed.), Immunoinformatics, Methods in Molecular Biology, vol. 2131, Springer, 2020.
[68] Z.L P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C.L. Huang, H.D. Chen, J. Chen, Y. Luo, H. Guo, R.D. Jiang, M.Q. Liu, Y. Chen, X. R. Shen, X. Wang, X.S. Zheng, K. Zhao, Q.J. Chen, F. Deng, L.L. Liu, B. Yan, F. X. Zhan, Y.Y. Wang, G.F. Xiao, Z.L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.
[69] T. Zhang, Q. Wu, Z. Zhang, Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak, Curr. Biol. 30 (2020) 1346–1351, e2.