MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **77** (2017) 321-332

ISSN 0340 - 6253

2D–Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome

Damian Panas[a], Piotr Wąż[b], Dorota Bielińska–Wąż[a], Ashesh Nandy[c], Subhash C. Basak[d]

[a] *Department of Radiological Informatics and Statistics, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland*

[b] *Department of Nuclear Medicine, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland*

[c] *Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700068, India*

[d] *University of Minnesota Duluth-Natural Resources Research Institute and Department of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*

(Received November 1, 2016)

Abstract

2D-dynamic representation of DNA/RNA sequences has been applied for the characterization of the complete genome sequence of Zika virus. Graphically, *the 2D-dynamic graphs* evolve with time. Numerically, applying descriptors related to the 2D-dynamic graphs, correct classification of the sequences has been obtained. These descriptors have been shown to give an adequate characteristics of the Zika virus genome. The classification diagrams form a new mathematical description of the evolution of the genome sequence of the Zika virus.

1 Introduction

The Zika virus (ZIKV) that was first isolated in Uganda in 1947 had spread as an epidemic across Brazil and Latin America in late 2015 [1,2] and has now been detected in Singapore and the USA where the first local infections have been recently identified [3]. The virus has raised serious concerns for its damaging possibilities that include microcephaly among

children borne by infected mothers [4, 5]. There are no known cures or preventives to date, but the virus is under intense scrutiny [6]. A recent sequence study of the first autochthonous transmission of the virus in Brazil published on the 1st of September 2016 has been accepted by the World Health Organization as a candidate reference strain for studies on the Zika virus [7].

The current limited knowledge about this virus calls for research into various aspects of the viral sequence and its manifestations as well as the exact mechanisms of its actions. We had in a recent paper attempted to characterize the viral genome in terms of its phylogeny, amino acid composition, mutational trends and the like [8]. There we had noted that the African and Asian-American lineages of the Zika virus fall into two distinct clades, which however was not quantified.

The current paper seeks to fill this void by using *2D-dynamic representation* of the viral sequence. We apply this representation of DNA/RNA sequences to characterization of the Zika virus genome using the Trösemeier et al. results (No. 14, Table 1) as well as the sequences considered in our previous work [8] and listed in Table 1. The sequences with undefined bases have been removed from the considerations.

In the theoretical model used in the present work, the DNA/RNA sequence is represented graphically by the so called "2D-dynamic graph". The approach reported here belong to the methodology of graphical representation of DNA/RNA sequences. The bases of the nucleic acids are plotted in a Cartesian coordinate system. Such graphs can be used to formulate various numerical descriptors which are quantitative descriptors of the sequences.

The results show a group of African Zika viral genomes are clearly separated from a cluster of the Asian-American genomes indicating significant mutational changes differentiating the latter group from the former. Such modeling with quantification of differences could be of considerable utility in surveillance programs of the Zika virus.

2 Theory

In the present work we apply the 2D-dynamic representation method introduced and developed by us several years ago [8–14]. In the methods, called in the literature *Graphical representations of DNA/RNA sequences*, the sequence of symbols (A, C, G, T/U corresponding respectively to adenine, cytosine, guanine, thymine/uracil) is represented by

Table 1. Sequence data

No.	Accession no.	Country reference	Year	N
1	HQ234498	Uganda	1947	10269
2	HQ234500	Nigeria	1968	10251
3	KF268948	Central African Rep.	1976	10788
4	KF268949	Central African Rep.	1976-1980	10776
5	KF383119	Senegal	2001	10272
6	KF993678	Canada	2013	10141
7	KJ776791	French Polynesia	2013	10617
8	KU312312	Suriname	2015	10374
9	KU321639	Brazil	2015	10676
10	KU365777	Brazil	2015	10662
11	KU365778	Brazil	2015	10727
12	KU365779	Brazil	2015	10662
13	KU365780	Brazil	2015	10662
14	KX369547	Reference strain	2013	10769

graphs. The advantage of these methods is a possibility of both graphical and numerical comparison of the considered objects [15–28]. For the graphs, different types of descriptors may be created. The descriptors are defined as the values which characterize the graphs numerically (for reviews see [29–31]). The most recent review of graphical methods in bioinformatics one can find in the article titled "Milestones in Graphical Bioinformatics" [31]. The authors of this review article cite 2D-dynamic representation method as one of the "milestones".

In this method the DNA/RNA sequence is represented by a plot called 2D-dynamic graph. The graph consists of point masses in a 2D space. Each base is represented by a unit vector: A=(-1,0), G=(1,0), C=(0,1), and T/U=(0,-1). The coordinates of the point masses are determined by shifts starting from the point (0,0). At the end of each vector a point mass with $m = 1$ is assigned. The masses are summarized if the ends of vectors meet several times at this point. In this way, we can take into account so called "repetitive walks", i. e. the shifts several times back and forth along the same trace. Masses larger then 1 remove this kind of degeneracy (nonuniqueness). Let us take a model example of a sequence TTCTTAG (Figure 1). The first base in the sequence is T, so (0,0) and (0,-1) are the coordinates of the beginning and of the end of the first vector, respectively, and $m = 1$ is located at (0,-1). Then the algorithm is repeated for the second base in the sequence starting from the end of the previous vector, i. e. the point with coordinates (0,-1) and so on.

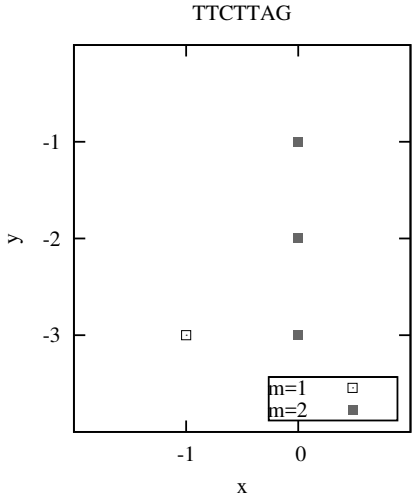

Figure 1. 2D-dynamic graph for a model sequence.

We have proposed several numerical quantities (descriptors) characterizing 2D-dynamic graphs; in particular, the coordinates of the center of mass and the principal moments of inertia are used as descriptors [9].

The coordinates of the center of mass of the 2D-dynamic graph, in the $\{X, Y\}$ coordinate system with (0,0) as the origin, are

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \quad \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \tag{1}$$

where x_i, y_i are the coordinates of mass m_i. Since the total mass of the sequence is $N = \sum_i m_i$, where N is the length of the sequence, Eq. 1 may be rewritten as

$$\mu_x = \frac{1}{N}\sum_i m_i x_i, \quad \mu_y = \frac{1}{N}\sum_i m_i y_i. \tag{2}$$

The moment of inertia tensor of the 2D-dynamic graph is defined in the same way as in the classical dynamics:

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix}, \tag{3}$$

where

$$I_{xy} = I_{yx} = -\sum_i m_i x_i^{\mu} y_i^{\mu}, \tag{4}$$

$$I_{xx} = \sum_i m_i (y_i^{\mu})^2, \tag{5}$$

$$I_{yy} = \sum_i m_i (x_i^{\mu})^2, \tag{6}$$

and x_i^{μ}, y_i^{μ} are the coordinates of mass m_i in the Cartesian coordinate system for which the origin has been selected at the center of mass.

The *principal moments of inertia* are defined as solutions $I = I_{11}, I_{22}$ of the second-order equation

$$\begin{vmatrix} I_{xx} - I & I_{xy} \\ I_{xy} & I_{yy} - I \end{vmatrix} = 0. \tag{7}$$

In a previous work [13] we have shown that the descriptors

$$D_k^{\gamma} = \frac{\mu_{\gamma}}{I_{kk}}; \quad \gamma = x, y; \quad k = 1, 2, \tag{8}$$

correctly characterize the DNA sequence.

3 Results and discussion

The moments of inertia of the 2D-dynamic graphs representing the complete genome sequence of Zika virus are shown in Table 2, and the descriptors, defined in Eq. 8 for these sequences, in Table 3.

Table 2. Moments of inertia of the 2D-dynamic graphs

No.	$I_{xx}/10^6$	$I_{yy}/10^6$	$I_{xy}/10^6$	$I_{11}/10^6$	$I_{22}/10^6$
1	5.138	31.07	-1.928	31.22	4.995
2	5.002	44.46	2.446	44.62	4.851
3	4.110	27.87	4.064	28.54	3.434
4	5.040	22.43	5.179	23.85	3.614
5	5.028	34.28	-2.477	34.49	4.820
6	8.873	58.08	-17.98	63.95	3.002
7	10.66	60.07	-16.13	64.87	5.864
8	10.19	59.78	-17.23	65.18	4.789
9	10.20	63.56	-17.58	68.84	4.933
10	10.76	60.96	-18.17	66.85	4.871
11	12.05	57.46	-18.84	64.26	5.251
12	9.889	60.05	-16.74	65.12	4.818
13	10.64	60.84	-18.04	66.65	4.835
14	11.37	60.03	-17.12	65.45	5.955

Table 3. Descriptors characterizing the complete genome sequence of Zika virus

No.	$D_1^x \cdot 10^6$	$D_2^x \cdot 10^5$	$D_1^y \cdot 10^7$	$D_2^y \cdot 10^6$
1	2.408	1.505	-5.668	-3.542
2	2.284	2.101	-5.770	-5.306
3	2.384	1.982	-17.44	-14.50
4	2.443	1.612	-21.60	-14.26
5	2.301	1.646	-4.421	-3.163
6	1.311	2.793	.6524	1.390
7	1.400	1.548	-3.121	-3.452
8	1.399	1.904	-2.267	-3.086
9	1.300	1.814	-3.219	-4.492
10	1.274	1.749	-2.811	-3.858
11	1.291	1.580	-2.764	-3.383
12	1.302	1.759	-2.915	-3.940
13	1.277	1.761	-2.834	-3.906
14	1.294	1.422	-2.454	-2.697

Table 4. Correlations with time

Kind of correlation	N	μ_x	μ_y	$\sqrt{\mu_x^2 + \mu_y^2}$
Spearman	0.276	0.393	0.255	0.391
Kendall	0.205	0.316	0.172	0.316
Pearson	0.243	0.366	0.464	0.177

Table 5. Correlations with time

Kind of correlation	I_{xx}	I_{yy}	I_{xy}	I_{11}	I_{22}
Spearman	0.787	0.787	-0.822	0.857	0.130
Kendall	0.604	0.604	-0.633	0.690	0.086
Pearson	0.846	0.820	-0.857	0.840	0.204

Table 6. Correlations with time

Kind of correlation	D_1^x	D_2^x	D_1^y	D_2^y
Spearman	-0.868	0.064	0.729	0.278
Kendall	-0.719	0.058	0.518	0.172
Pearson	-0.887	0.128	0.563	0.466

Figure 2 shows four 2D-dynamic graphs. As we can see, the time evolution of the complete genome sequence of Zika virus is well represented graphically. Pairs of graphs are similar to each other: HQ234498 Uganda 1947 is similar to KF268948 Central African Republic 1976 and KJ776791 French Polynesia 2013 is similar to KU365777 Brazil 2015. It seems that the structural forms of the graphs evolve with time.

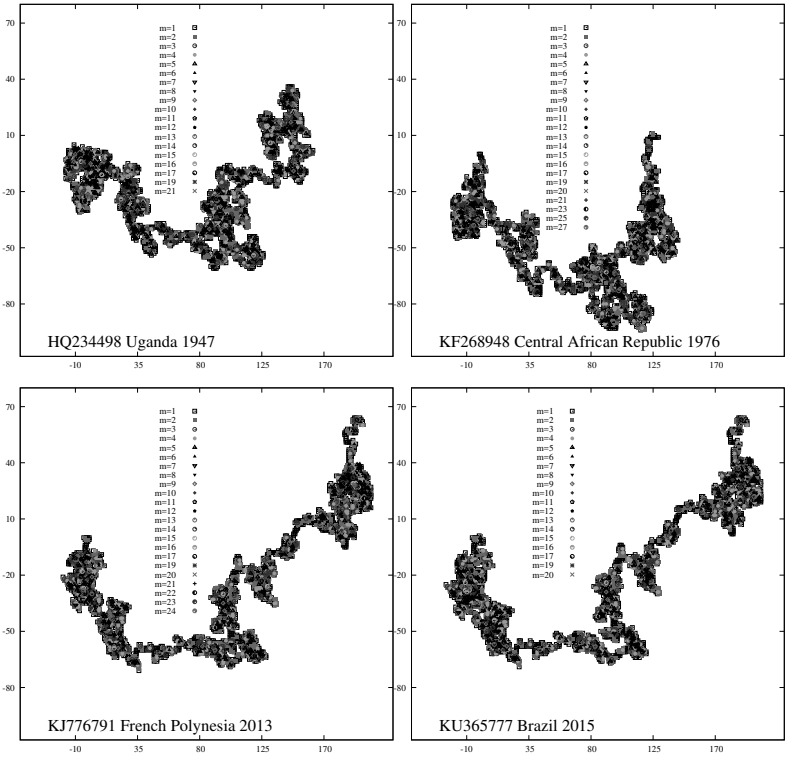

Figure 2. 2D-dynamic graphs representing the complete genome sequence of Zika virus.

This observation is confirmed by the calculations. The correlations with time of the descriptors are shown in Tables 4-6. As time we take the numbers in column 4 of Table 1 (Year). In the case of sequence No. 4 we take year 1978. For all kinds of correlations (Spearman, Kendall, Pearson) we have obtained low correlation with time (correlation coefficients smaller than 0.5) for the following descriptors: N, μ_x, μ_y, the radius $\sqrt{\mu_x^2+\mu_y^2}$, I_{22}, D_2^x, D_2^y. Large correlation with time (all kinds of correlation coefficients larger than 0.5) has been obtained for: I_{xx}, I_{yy}, I_{xy}, I_{11}, D_1^x, D_1^y.

Good visualization of the characterization of the complete genome sequence of Zika virus is given by the classification diagrams. We have obtained classification diagrams based on the moments of inertia and on the descriptors composed of both coordinates of centers of mass and moments of inertia (Eq. 8).

In Figure 3 we see that the moments of inertia I_{xx}, I_{yy}, I_{xy} correctly classify the genomes. The descriptors corresponding to Africa and South America are located in different parts of the diagram and are separated by a plane. Similar results we have obtained for the time (Year in Table 1), I_{11} and I_{22} (Figure 4), for the time, D_1^x and D_1^y (Figure 5), as well as for the time, D_1^x and D_2^x (Figure 6).

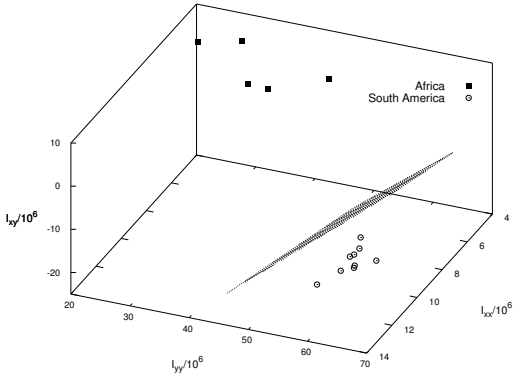

Figure 3. $I_{xx} - I_{yy} - I_{xy}$ classification diagram.

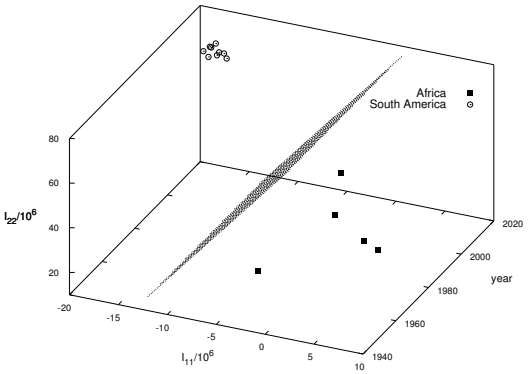

Figure 4. $year - I_{11} - I_{22}$ classification diagram.

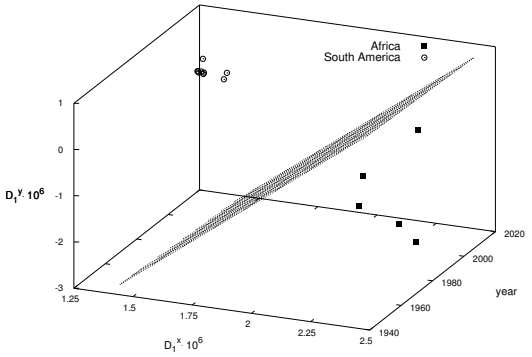

Figure 5. $year - D_1^x - D_1^y$ classification diagram.

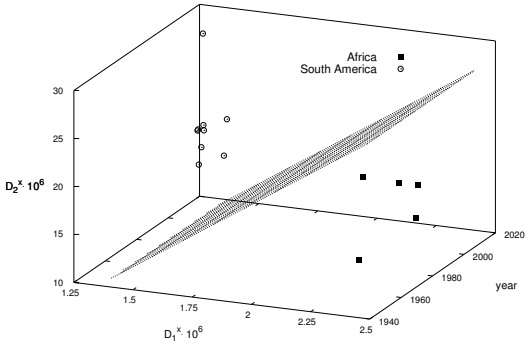

Figure 6. $year - D_1^x - D_2^x$ classification diagram.

We had noted in a recent publication [8] that the African and Asian-American lineages fall into two distinct clades, which is substantiated and quantified here by the clustering in the moments of inertia. It is also to be noted that the dispersion of points in the Asian-American cluster is much wider than for the African group indicating that the mutational changes are far wider in the former than in the other group.

4 Conclusions

Moments of inertia of the 2D-dynamic graphs can be considered as the second-order corrections to the first-order ones (the coordinates of the centers of mass). Using moments of inertia one can obtain a correct characterization of the complete genome sequence of Zika virus (Table 5, Figure 3, Figure 4). The coordinates of the centers of mass can be also considered simultaneously with the moments of inertia for the correct classification of the sequences (Table 6, Figure 5, Figure 6).

Summarizing, the 2D-dynamic representation of the DNA/RNA sequences is a good tool for the characterization of the complete genome sequence of Zika virus. Using this method we have created a new mathematical description of the evolution of the complete genome sequence of the Zika virus.

The results show a clustering of the African genomes clearly separated from a cluster of the Asian-American genomes indicating significant mutational changes differentiating the latter group from the former. Such modeling with quantification of differences could be of considerable utility in surveillance programs of the Zika virus.

References

[1] A. D. Haddow, A. J. Schuh, C. Y. Yasuda, M. R. Kasper, V. Heang, R. Huy, H. Guzman, R. B. Tesh, S. C. Weaver, Genetic characterization of Zika virus strains: Geographic expansion of the Asian lineage, *PLoS Negl. Trop. Dis.* **6** (2012) #e1477.

[2] M. Roa, Zika virus outbreak: Reproductive health and rights in Latin America, *Lancet* **387** (2016) 843–843.

[3] O. Dyer, Outbreak of Zika in Singapore sparks warnings in neighbouring countries, *Brit. Med. J.* **354** (2016) #i4740.

[4] H. Elachola, E. Gozzer, J. Zhuo, Z. A. Memish, A crucial time for public health preparedness: Zika virus and the 2016 Olympics, Umrah, and Hajj, *Lancet* **387** (2016) 630–632.

[5] WHO. "WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain–Barré syndrome". World Health Organization. 1 February 2016. Retrieved 16th February 2016.

[6] J. W. Cross, (July 2016) 2016 Cumulative Selected Peer-Reviewed Publications – Zika. https://www.researchgate.net/publication/305661569_2016_Cumulative_Selected_Peer-Reviewed_Publications_-_Zika

[7] J. Trösemeier, D. Musso, J. Blümel, J. Thézé, O. G. Pybus, S. A. Baylis, Genome sequence of a candidate world health organization reference strain of Zika virus for nucleic acid testing, *Genome Announc.* **4** (2016) #e00917-16.

[8] A. Nandy, S. Dey, S. C. Basak, D. Bielińska–Wąż, P. Wąż, Characterizing the Zika virus genome – A bioinformatics study, *Curr. Comput. Aided Drug Des.* **12** (2016) 87–97.

[9] D. Bielińska–Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, *Chem. Phys. Lett.* **442** (2007) 140–144.

[10] D. Bielińska–Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Distribution moments of 2D-graphs as descriptors of DNA sequences, *Chem. Phys. Lett.* **443** (2007) 408–413.

[11] D. Bielińska–Wąż, P. Wąż, T. Clark, Similarity studies of DNA sequences using genetic methods, *Chem. Phys. Lett.* **445** (2007) 68–73.

[12] D. Bielińska–Wąż, P. Wąż, W. Nowak, A Nandy, S. C. Basak, Similarity and dissimilarity of DNA/RNA sequences, in: T. E. Simos, G. Maroulis (Eds.), *AIP Conference Proceedings*, American Inst. Phys., 2007, pp. 28–30.

[13] P. Wąż, D. Bielińska–Wąż, A. Nandy, Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences, *J. Math. Chem.* **52** (2014) 132–140.

[14] D. Bielińska–Wąż, P. Wąż, 2D-dynamic representation of DNA sequences as a graphical tool in bioinformatics, in: M. D. Todorov (Ed.), *AIP Conference Proceedings*, American Inst. Phys., 2016, pp. #060004-1–#060004-5.

[15] 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.

[16] M. Randić, M. J. Vračko, On the similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **40** (2000) 599–606.

[17] M. Randić, X. Guo, S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **41** (2001) 619–626.

[18] P. He, J. Wang, Numerical characterization of DNA primary sequence, *Int. El. J. Mol. Des.* **1** (2002) 668–674.

[19] 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.

[20] Y. Yao, T. Wang, A class of new 2-D graphical representation of DNA sequences and their application, *Chem. Phys. Lett.* **398** (2004) 318–323.

[21] B. Liao, Y. Zhang, K. Ding, T. J. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *J. Mol. Struct. (Theochem)* **717** (2005) 99–203.

[22] Q. Dai, X. Liu, T. Wang, A novel graphical representation of DNA sequences and its application, *J. Mol. Graph. Model.* **25** (2006) 340–344.

[23] A. Nandy, S. C. Basak, B. D. Gute, Graphical representation and numerical characterization of H5N1 Avian Flu neuraminidase gene sequence, *J. Chem. Inf. Model.* **47** (2007) 945–951.

[24] H. González–Díaz, L.G. Pérez–Montoto, A. Duardo–Sanchez, E. Paniagua, S. Vázquez–Prieto, R. Vilas, M.A. Dea–Ayuela, F. Bolas–Fernández, C. R. Munteanu, J. Dorado, J. Costas, F. M. Ubeira, Generalized lattice graphs for 2D-visualization of biological information, *J. Theor. Biol.* **261** (2009) 136–147.

[25] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L. Pan, Y. Ye, ColorSquare: A colorful square visualization of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **68** (2012) 621–637.

[26] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, *Math Biosci.* **241** (2013) 217–224.

[27] V. Aram, A. Iranmanesh, Z. A. Majid, Spider representation of DNA sequences, *J. Comput. Theor. Nanos.* **11** (2014) 418–420.

[28] Y. Li, Q. Liu, X. Zheng, DUC-Curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment, *Physica A* **456** (2016) 256–270.

[29] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and applications, *Arkivoc* **ix** (2006) 211–238.

[30] D. Bielińska–Wąż, Graphical and numerical representations of DNA sequences: Statistical aspects of similarity, *J. Math. Chem.* **49** (2011) 2345–2407.

[31] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, *Int. J. Quant. Chem.* **113** (2013) 2413–2446.