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.

Chemical Physics 513 (2018) 156–164

Contents lists available at ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

Graphical representation methods: How well do they discriminate between homologous gene sequences?

Dwaipayan Sen[a], Proyasha Roy[a], Ashesh Nandy[a,*], Subhash C. Basak[b], Sukhen Das[c]

[a] Centre for Interdisciplinary Research and Education, Jodhpur Park, Kolkata 700068, India
[b] Department of Chemistry & Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA
[c] Physics Department, Jadavpur University, Jadavpur, Kolkata 700032, India

ARTICLE INFO

Keywords:
Graphical representation
Discriminatory power of sequence descriptors
Super-descriptors
2D and 3D methods
Globin genes
Flavivirus envelope genes

ABSTRACT

Graphical representation methods constitute a class of alignment-free techniques for comparative study of biomolecular sequences. In this brief commentary, we study how well some of these methods can discriminate among closely related genes.

1. Introduction

Graphical representations of biomolecular sequences have generated a lot of interest as a tool for alignment-free analysis, evidenced by the large number of research work on the subject. A principle application of these techniques has been in determining the evolutionary relationships analysed between different gene families [1]. The results obtained from the different graphical methods show small differences among them and display conformity with standard phylogenetic studies.

Graphical representations often provide a visual clue to the pattern of distribution of bases along DNA or RNA sequences. The representations are slightly more complicated in the case of protein sequences where one has to contend with 20 basic units, the amino acids, but ingenious schemes from 2D to 20D abstract graphs have been utilised to represent them too [2]. The graphical representation methods remain, to date, among the best to 'see' the base distribution in a DNA or a RNA sequence and follow the variations in a family of genes. This has enabled many applications of the techniques, for example, generation of new plant varieties [3], determination of origins of the SARS-coronavirus [4], identification of conserved regions in influenza virus neuraminidase gene for vaccine design and development [5].

Applications of these methods to viral sequence analyses have been a prime area of interest in view of the need for rapid development of drugs and vaccines against viral diseases. In a number of studies, our group has advocated rational designs of peptide vaccines against Influenza [5], Rotavirus [6], Human Papillomavirus [7] and Zika virus [8]. In particular, we have worked out detailed differences between base distributions within the species of the *Flavivirus* genus and have shown that the Dengue virus which rapidly becomes endemic in a country or region has country/region-specific sequence differences, requiring customized vaccines for better efficacy [9]. Similar graphical studies of viruses have also been carried out by other researchers, including analysis of the Coronavirus by Liao et al. [4].

Such observations raise the question of which among the many graphical representation methods should be employed in order to obtain the most accurate results. An expected tendency would be to consider those methods which discriminate among sequences more clearly than others, i.e., sequence differences should lead to more pronounced differences in their descriptors. In order to address this issue, we had earlier carried out a comparative analysis of several different methods [10] and showed that results of some of the methods correlate strongly with results of other methods, implying that the sequence descriptors arising from the various approaches essentially map the same features; those that do not exhibit high level of correlation could be considered as mapping some attributes unique to those schemes. In this brief commentary, we work out the descriptors for two families of sequences – the globin genes of *Homo sapiens* and the envelope genes of flaviviruses – to determine precisely how much discrimination is evidenced in a selected group of methods.

* Corresponding author.
E-mail address: anandy43@yahoo.com (A. Nandy).

https://doi.org/10.1016/j.chemphys.2018.07.031
Received 25 May 2018; Accepted 21 July 2018
Available online 26 July 2018
0301-0104/

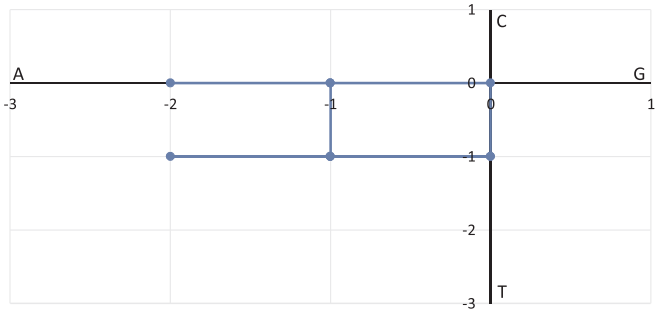

Fig. 2.1. Nandy plot of ATGAAGGCAA.

Table 2.1
The coordinates, the centre of mass of all the points, and the graph radius g_R of the sequence as per Nandy rectangular plot.

Base	Coordinates		Centre of mass		g_R
	X	Y	μ_x	μ_y	
A	-1	0	−0.9	−0.6	1.08
T	−1	−1			
G	0	−1			
A	−1	−1			
A	−2	−1			
G	−1	−1			
G	0	−1			
C	0	0			
A	−1	0			
A	−2	0			

2. Materials and methods

2.1. Description of methods

In our comparative study, we have used six methods based on 2-dimensional (2D) analysis and one 3-dimensional (3D) system (see Ref. [10] for details). The 2D methods are primarily based on the Cartesian coordinate system. In the Nandy plot [11,12], the four nucleic acid bases are assigned to the four axes of a 2D Cartesian coordinate system. A given sequence is plotted based on the distribution of its bases in the corresponding direction; in computations for this note adenine (A) was assigned to the negative x-axis, cytosine (C) to the positive y-axis, guanine (G) to the positive x-axis and thymine (T) to the negative y-axis. The weighted average of the x- and y-coordinates of each point of a sequence of length N represents the centre of mass. The Euclidean distance between the origin and the centre of mass provides a quantitative graph descriptor, termed as the graph radius (g_R). A variant of this method can be seen in the Yau plot [13] in which two quadrants are used with the first quadrant addressing thymine (T) and cytosine (C), and the second quadrant denoting guanine (G) and adenine (A). However, the authors of this method did not prescribe a sequence descriptor. As a result, for the purpose of this commentary, Nandy's graph

Table 2.2
The coordinates, centre of mass of all points and the graph radius as per the Yau et al. plot.

Base	Coordinates		Centre of mass		g_R
	X	Y	μ_x	μ_y	
A	0.5	−0.866	3.482	−2.172	4.104
T	1	0			
G	1.866	−0.5			
A	2.366	−1.366			
A	2.866	−2.232			
G	3.732	−2.732			
G	4.598	−3.232			
C	5.464	−2.732			
A	5.964	−3.598			
A	6.464	−4.464			

Table 2.2.1
Sample data information.

Sequences			Accession no.	Length
Globin gene (*Homo sapiens*)	Alpha1-globin-mRNA	HbA1	NM_000558	429
	Beta-globin-mRNA	HbB	NM_000518	444
	Delta-globin-mRNA	HbD	NM_000519	444
	Gamma-globin-mRNA	HbG1	NM_000559	444
Envelope gene (*Flavivirus*)	Dengue virus type 2	DENV2	JX669476	1485
	West Nile virus	WNV	KC601756	1503
	Yellow fever virus	YFV	JN620362	1479
	Zika virus	ZIKV	KX197192	1512

radius method is used to define the sequence descriptor of the Yau plot for comparative analysis.

The 2D Randic plot [14] and Song-Tang plot [15], employ parallel horizontal lines to denote nucleic acid bases. The key difference

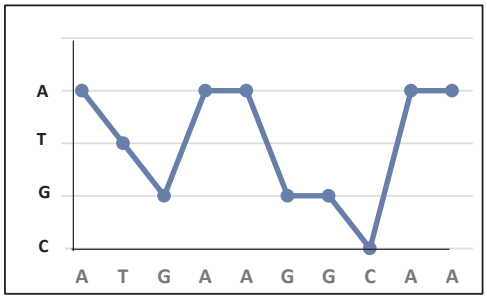

Fig. 2.3. Randic 2D plot for the sequence ATGAAGGCAA.

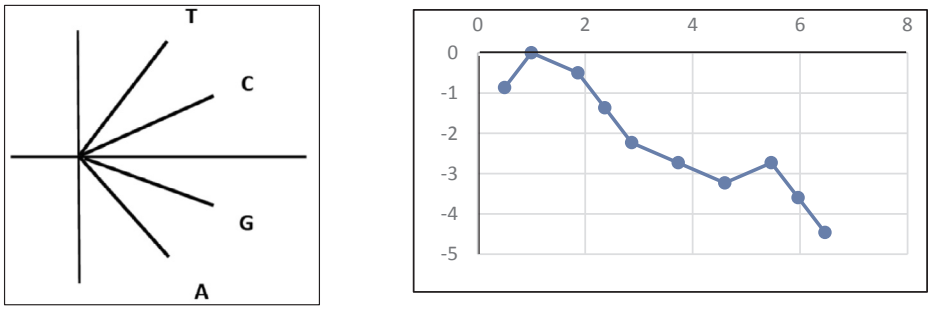

Fig. 2.2. The Yau plot coordinate system and the graph of the sample sequence ATGAAGGCAA.

Table 2.3.1
The D_E matrix in Randic 2D model.

	A	T	G	A	A	G	G	C	A	A
A	0	1.4142	2.8284	3	4	5.3852	6.3246	7.6158	8	9
T		0	1.4142	2.2361	3.1623	4.1231	5.0990	6.3246	7.0711	8.0623
G			0	2.2361	2.8284	3	4	5.0990	6.3246	7.2801
A				0	1	2.8284	3.6056	5	5	6
A					0	2.2361	2.8284	4.2426	4	5
G						0	1	2.2361	3.6056	4.4721
G							0	1.4142	2.8284	3.6056
C								0	3.1623	3.6056
A									0	1
A										0

Table 2.3.2
The D_G matrix in Randic 2D model.

	A	T	G	A	A	G	G	C	A	A
A	0	1	2	3	4	5	6	7	8	9
T		0	1	2	3	4	5	6	7	8
G			0	1	2	3	4	5	6	7
A				0	1	2	3	4	5	6
A					0	1	2	3	4	5
G						0	1	2	3	4
G							0	1	2	3
C								0	1	2
A									0	1
A										0

between the two systems is that, the Randic 2D plot consists of four lines whereas the Song-Tang plot comprises three lines. In the former, each of the four bases are assigned to each of the four lines and in the latter, the central line represents a pair of bases while the two peripheral lines each represent one of the remaining two bases. In both methods, the bases of a given sequence are plotted on the corresponding line along in the direction of the x-axis and then subsequently, joined by straight lines. The sequence descriptors are represented by the leading eigenvalues computed from *M/M* matrices of each graphical method.

The Wang-Zhang [16] and Ji-Li TB curve [17] deviate from the aforementioned systems in that, the Wang-Zhang plot employs a binary system in which the presence and absence of particular bases are represented by 1 and 0, respectively. There are three such systems: non-A, non-C and non-G. If, for example, the non-A configuration is used, A will be denoted with 0 and the remaining bases by 1. In the Ji-Li TB curve, the given sequence is mapped into nodes which are then connected sequentially in order to obtain three curves, R-Y, M-K and W-S. The *L/L* matrices are used to calculate the leading eigenvalues which are taken as the descriptors of sequences for both the analyses.

For a 3D graphical representation of a sequence, Randic [18] used each of the four corners of a tetrahedron to represent each of the four nucleic acid bases. The sequence descriptor is given by the leading eigenvalues calculated from the ratio of the Euclidean pairwise distance matrix D_E and pairwise graph theoretic distance matrix D_G of all the points on the graph.

The sequence descriptors for the aforementioned 7 graphical methods and the standard deviation of the descriptors from their averages have been calculated. For the Nandy and Yau methods, graph radii have been calculated. Leading eigenvalues from *M/M* matrices have been determined for Randic 3D, Randic 2D and Song-Tang methods as their numerical indices. For Wang-Zhang and Ji-Li plots, *L/L* matrices have been used to compute the leading eigenvalues; the *L/L* matrix for the Ji-Li plot is for the R-Y curve.

2.1.1. *Sample computations in these methods*

To fix our ideas of what these various approaches to graphical representations of nucleotide sequences represent and how the sequence descriptors characteristic to each method are computed, we consider a 10-base oligonucleotide, ATGAAGGCAA, for a simple application of each method. This short sequence constitutes the first ten bases in the hemagglutinin gene of the influenza virus, A/Novosibirsk/02/2009(H1N1), and therefore can be taken as representative of a real-life situation.

2.1.1.1. Nandy 2D rectangular plot. In the Nandy plot [11], the graph of our 10-base sequence drawn according to the given prescription is as shown in Fig. 2.1. Table 2.1 represents the coordinates, the centre of mass of all the points, and the graph radius g_R, the sequence descriptor, of the sequence as described in Section 2.1 [12].

2.1.1.2. Yau et al. plot. In the representation by Yau et al. [13], the four bases being assigned in the first and second quadrant only, the coordinates of the four bases are defined as: $(1/2, -\sqrt{3}/2)$ for A, $(\sqrt{3}/2, -1/2)$ for G, $(\sqrt{3}/2, 1/2)$ for C and $(1/2, \sqrt{3}/2)$ for T. For the purpose of this commentary, as mentioned in Section 2.1, we define a centre of mass and a graph radius as descriptors analogously to the Raychaudhury-Nandy method [12]. The graph and sequence descriptor for our sequence in the Yau et al. [13] plot are given in Fig. 2.2 and Table 2.2.

Table 2.3.3
The *M/M* matrix in Randic 2D model.

	A	T	G	A	A	G	G	C	A	A
A	0	1.4142	1.4142	1	1	1.0770	1.0541	1.088	1	1
T		0	1.4142	1.1180	1.0541	1.0308	1.0198	1.0541	1.0102	1.0078
G			0	2.2361	1.4142	1	1	1.0198	1.0541	1.0400
A				0	1	1.4142	1.2019	1.2500	1	1
A					0	2.2361	1.4142	1.4142	1	1
G						0	1	1.1180	1.2019	1.1180
G							0	1.4142	1.4142	1.2019
C								0	3.1623	1.8028
A									0	1
A										0

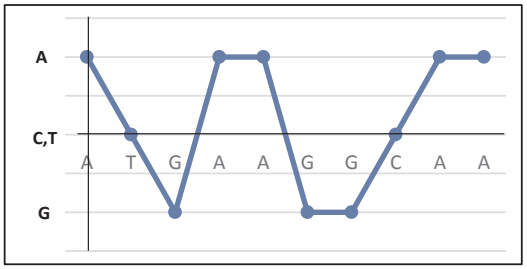

Fig. 2.4. One plot as per Song and Tang's representation of the sequence ATGAAGGCAA.

2.1.1.3. Randic 2D plot. As mentioned briefly in Section 2.1, Randic et al. [14] proposed four horizontal equidistant parallel lines labelled as A, T, G, and C from top to bottom. To plot the nucleic acid sequence, the bases are numbered along the x-axis, and straight lines are drawn from individual points on the four parallel lines according to the bases occurring in the sequence (Fig. 2.3).

An *M/M* matrix method is used to determine a descriptor for the given sequence ATGAAGGCAA. The entries of the *M/M* matrix are determined by dividing the Euclidean distance D_E between two vertices of the zigzag curve by the graph theoretic distance D_G between the two vertices. As an example, if we take the first base A and the eighth base C of our sequence, then the corresponding matrix element is obtained by taking the ratio of Euclidean distance between the A and C bases, i.e., $\sqrt{58}$ (Table 2.3.1), to the number of edges between A and C, which is 7 here (Table 2.3.2). Proceeding in this way, we obtain the *M/M* matrix elements as shown below (Table 2.3.3) from which the leading eigenvalue is calculated and taken as the descriptor of the sequence.

The leading eigenvalue in this case turns out to be 11.2525

2.1.1.4. Song-Tang plot. In the Song and Tang [15] method, counting of the nucleotides is done along the x-axis and the bases are plotted and connected by straight lines as in the case of the Randic 2D plot except that here there are three horizontal lines. In the case of DNA primary sequences, the four bases A, C, G, and T can be grouped as per their characteristics:

- purine R = (A, G) and pyrimidine Y = (C, T)
- amino M = (A, C) and keto K = (G, T)
- weak-H bond W = (A, T) and strong-H bond S = (C, G)

Keeping one of such groups, such as purine, as central line, and cytosine (C) and thymine (T) as the peripheral lines, we can plot a graph. Repeating the same process for the other groups, we get a total of 6 ($^4C_2$) combinations of plots. For each of them, the 2 peripheral lines can also be exchanged. In total, 12 (6 × 2) plots can be obtained.

In one such graph, where cytosine (C) and thymine (T) are assigned to the central line, adenine (A) to the uppermost line and guanine (G) to the lowermost line, the plot for the sequence ATGAAGGCAA is represented as shown in Fig. 2.4.

In this method, the matrix elements and the descriptor are obtained by defining two matrices, *M/M* and *L/L*, as below (Tables 2.4.1, 2.4.2). The leading eigenvalues are computed to form descriptors of the corresponding DNA sequence.

$$(ED)_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$(M/M)_{ij} = (ED)_{ij}/(GD)_{ij} i \neq j$$

$$(M/M)_{ij} = 0$$

$$(PD)_{ji} = (PD)_{ij} = (ED)_{ii+1} + (ED)_{i+1,i+2} + \cdots, +(ED)_{j-1,j} i < j$$

$$(L/L)_{ij} = (ED)_{ij}/(PD)_{ij} i \neq j$$

$$(L/L)_{ij} = 0$$

where *ED*, *GD* and *PD* are the Euclidean distance matrix, graph theoretical distance matrix and path distance matrix, respectively.

The leading eigenvalue in this case turns out to be 10.5856. This is

Table 2.4.2
L/L matrix computed as per the Song-Tang prescription.

	A	T	G	A	A	G	G	C	A	A
A	0	1	1	0.5924	0.6596	0.6488	0.6800	0.6599	0.6596	0.6855
T		0	1	0.6126	0.6800	0.5987	0.6466	0.6451	0.6599	0.6882
G			0	1	0.8740	0.5482	0.6180	0.6466	0.6800	0.7068
A				0	1	0.8740	0.8512	0.7297	0.7078	0.7440
A					0	1	0.8740	0.6800	0.6596	0.7078
G						0	1	0.9262	0.9418	0.9262
G							0	1	1	0.9418
C								0	1	0.9262
A									0	1
A										0

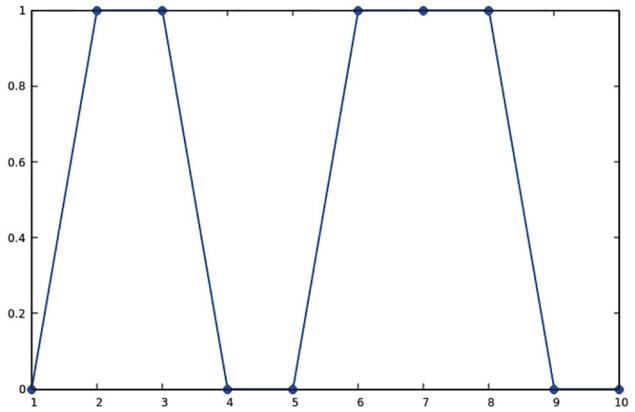

Fig. 2.5.1. Non-A.

Table 2.4.1
M/M matrix computed as per the Song-Tang prescription.

	A	T	G	A	A	G	G	C	A	A
A	0	1.4142	1.4142	1	1	1.0770	1.0541	1.0102	1	1
T		0	1.4142	1.1180	1.0541	1.0308	1.0198	1	1.0102	1.0078
G			0	2.2361	1.4142	1	1	1.0198	1.0541	1.0400
A				0	1	1.4142	1.2019	1.0308	1	1
A					0	2.2361	1.4142	1.0541	1	1
G						0	1	1.1180	1.2019	1.1180
G							0	1.4142	1.4142	1.2019
C								0	1.4142	1.1180
A									0	1
A										0

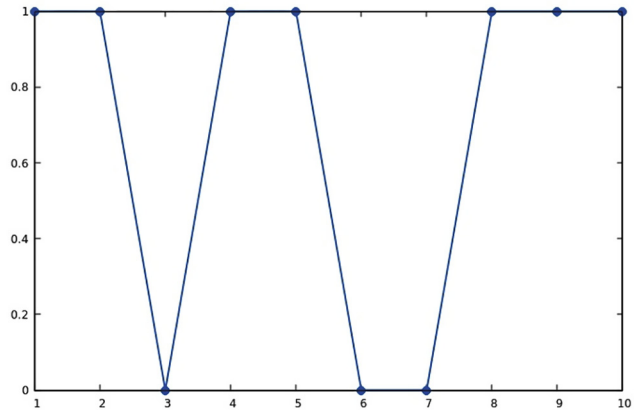

Fig. 2.5.2. Non-C.

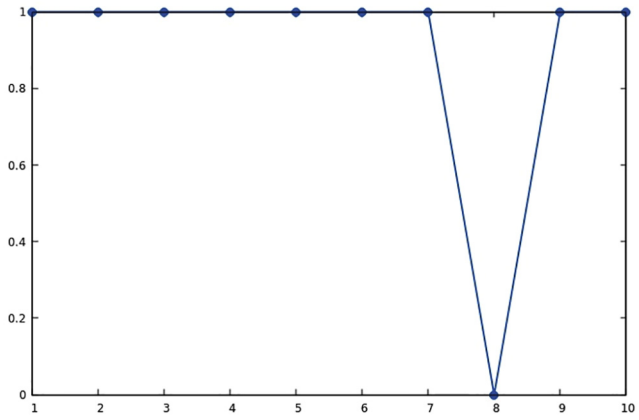

Fig. 2.5.3. Non-G.

Table 2.5
The coordinates and the leading eigenvalues calculated as per the Wang-Zhang model.

Base	Co-ordinates			Descriptor Vector
	Non-A	Non-C	Non-G	
A	0	1	1	Non-A = 0.4665
T	1	1	1	Non-C = 0.3201
G	1	1	0	Non-G = 4.1949
A	0	1	1	
A	0	1	1	
G	1	1	0	
G	1	1	0	
C	1	0	1	
A	0	1	1	
A	0	1	1	

for one characteristic curve. A vector can be constructed to characterise a DNA sequence from these data obtained from different characteristic curves.

2.1.1.5. Wang-Zhang plot. In the Wang-Zhang method [16], a binary technique is employed where the presence or absence of bases in the sequence are assigned 0 and 1 according to the specific configuration; for example, 0 for A and 1 for the others if it is a non-A plot. Similar protocol is implemented for the non-C and non-G plots. The graphs for the sequence ATGAAGGCAA are shown in Figs. 2.5.1(Non-A), 2.5.2 (Non-C) 2.5.3 (Non-G).

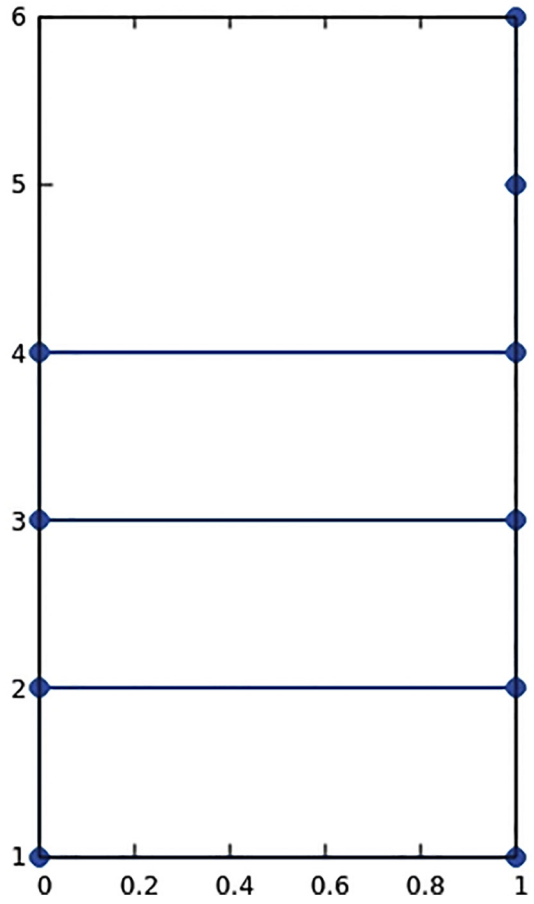

Fig. 2.6.1. R-Y curve.

L/L matrices (Table 2.5) are computed based on the three 2D graphs shown in Fig. 2.5 as a typical example for characterizing the DNA sequence. Next, eigenvalues are computed from the matrices and then the descriptor: $\lambda_{\text{non-N}}$ = maxeig (maximal eigenvalue) + mineig (minimal eigenvalue), where N = A, G, C. Table 2.5 tabulates the leading eigenvalues for our sequence ATGAAGGCAA.

2.1.1.6. Ji-Li method. Ming Ji and Chun Li [17] proposed a 2D graphical representation method where $X = X_1X_2\cdots X_n$ is taken as a DNA primary sequence with n bases and defined a homomorphic map φ_1 by $\varphi_1(X) = \varphi_1(X_1)\varphi_1(X_2)\cdots\varphi_1(X_n)$ where

$$\varphi_1(X_i) = \begin{cases} 1(1,R_i) if X_i \in R \\ 0(0,Y_i) if X_i \in Y \end{cases}$$

and where, $R_i(Y_i)$ is the cumulative occurrence of the numbers of bases $\in R(Y)$ in the first i bases. The DNA sequence is mapped into a series of nodes $P_i's$. Connecting the adjacent nodes, an R-Y curve is obtained and two other maps can also be defined: the M-K and W-S-TB curves for the sequence ATGAAGGCAA. These are shown in Figs. 2.6.1, 2.6.2, 2.6.3.

Here, three matrices namely ED matrix (Table 2.6.1), *M/M* matrix (Table 2.6.2) and *L/L* matrix (Table 2.6.3), are formed as before from which the leading eigenvalues are calculated. These are the descriptors of the corresponding DNA sequence.

2.1.1.7. Randic 3D plot. In the 3D graphical representation of DNA sequence proposed by Randic et al. [18], each of the four bases are assigned to the corners of a tetrahedron where the points are as follows: $(+1, -1, -1)$ for A, $(-1, +1, -1)$ for G, $(-1, -1, +1)$ for C and $(+1, +1, +1)$ for T. The graph is plotted by placing the 1st base say A

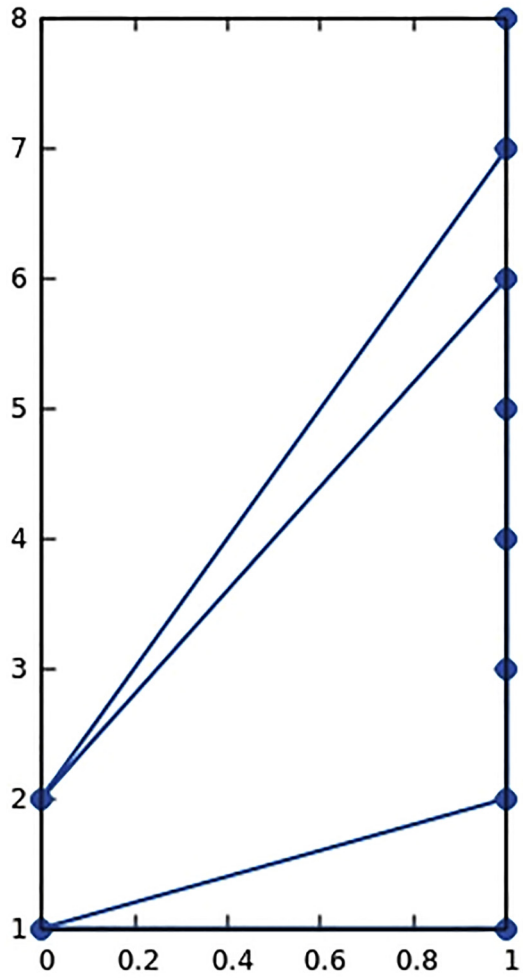

Fig. 2.6.2. M-K curve.

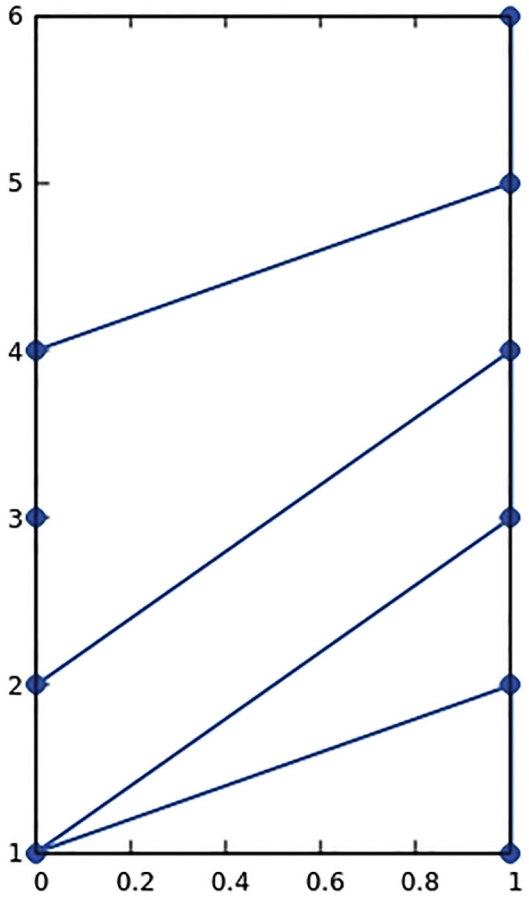

Fig. 2.6.3. W-S curve.

(for our mini sequence ATGAAGGCAA) at the corner position i.e. at $(+1, -1, -1)$ and the next base i.e. T at (2, 0, 0). Proceeding in this manner, we obtain a plot of the sequence ATGAAGGCAA as shown in Fig. 2.7.

In this method, a D_E/D_G matrix (from the Euclidean pairwise distance matrix D_E and a pairwise graph theoretic distance matrix D_G of all the points on the graph) is constructed in which the sequence descriptor is defined as the leading eigenvalue. In the following table each matrix element is calculated as follows: Taking the (1,8) element, i.e., 1st base A to the 8th base C as an example, the Euclidean distance is calculated from the coordinates shown in Table 2.7.1 and is divided by the minimum distance between two consecutive points which is $\sqrt{3}$. The division by the minimum distance is required to normalize the distance scale, so that the Euclidean distance between adjacent vertices equals 1, and not $\sqrt{3}$ (due to taking the side of the cube to be 1). Dividing this value by the number of edges before the 8th base, i.e., 7 here, the value obtained is $=(\sqrt{11}/\sqrt{3})/7 = 0.2736$, the matrix element (Table 2.7.2).

The descriptor value, i.e. the leading eigenvalue, in the Randic 3D model for this sequence is 5.6174.

2.2. Sequence information

Our sample data consist of four members of the *Homo sapiens* globin family and four mosquito-borne human-infecting members of the flavivirus group (Table 2.2.1). Among the globin genes, which are about 444 bases in length, the alpha globin being CG-rich is quite different from the beta globin group which are primarily AT-rich. The flaviviruses are of approximately 1500 bases in length and the Dengue virus is quite different in its adenine richness from the other three flaviviruses [9]. The sequence descriptors should therefore yield sufficiently different numbers to discriminate among the globin genes and between the flaviviruses, which will be compared with the computed standard deviations of the results as percentages of the corresponding descriptor average; higher the percentage, more is the discriminatory power of the method.

3. Results and discussions

Table 3.1 shows the results for the flavivirus group of envelope genes and the globin group of genes. As can be seen, the two oldest of the current crop of graphical representations, viz., Nandy plot and Randic-3D plot, show the best clear discriminatory power. For example, in case of the flaviviral envelope genes, the Nandy plot descriptor ranges from 72.61 for DENV2 to 8.91 for WNV and as low as 4.51 for ZIKV; the Randic 3D plot shows descriptors ranging from 221.99 for DENV2 to 113.57 for WNV and 156.22 for ZIKV. The observation that there is a wide difference between DENV2 and, WNV and ZIKV which reflects the base distributions in their genetic composition was first mooted in a previously published paper in detail [9]. On the other hand, the same three sequences yield descriptor values, respectively, of 1485.51, 1503.43, 1495.21 in the Song-Tang representation and 1487.07, 1505.02, 1496.75 in the Randic 2D representation, showing very little distinction between DENV2, WNV and ZIKV. This is reflected

Table 2.6.1
ED matrix from the Ji-Li R-Y TB curve.

	A	T	G	A	A	G	G	C	A	A
A	0	1	1	2	3	4	5	1.4142	6	7
T		0	1.4142	2.2361	3.1623	4.1231	5.0990	1	6.0828	7.0711
G			0	1	2	3	4	1	5	6
A				0	1	2	3	1.4142	4	5
A					0	1	2	2.2361	3	4
G						0	1	3.1623	2	3
G							0	4.1231	1	2
C								0	5.0990	6.0828
A									0	1
A										0

Table 2.6.2
M/M matrix from the Ji-Li R-Y TB curve.

	A	T	G	A	A	G	G	C	A	A
A	0	1	0.5000	0.6667	0.7500	0.8000	0.8333	0.2020	0.7500	0.7778
T		0	1.4142	1.1180	1.0541	1.0308	1.0198	0.1667	0.8690	0.8839
G			0	1	1	1	1	0.2000	0.8333	0.8571
A				0	1	1	1	0.3536	0.8000	0.8333
A					0	1	1	0.7454	0.7500	0.8000
G						0	1	1.5811	0.6667	0.7500
G							0	4.1231	0.5000	0.6667
C								0	5.0990	3.0414
A									0	1
A										0

Table 2.6.3
L/L matrix from the Ji-Li R-Y TB curve.

	A	T	G	A	A	G	G	C	A	A
A	**0**	1	0.4142	0.5858	0.6796	0.7388	0.7795	0.1342	0.3837	0.4208
T		0	1	0.9262	0.9262	0.9341	0.9418	0.1049	0.4156	0.4522
G			0	1	1	1	1	0.1231	0.3782	0.4219
A				0	1	1	1	0.1985	0.3273	0.3782
A					0	1	1	0.3652	0.2673	0.3273
G						0	1	0.6173	0.1957	0.2673
G							0	1	0.1084	0.1957
C								0	1	0.9973
A									0	1
A										0

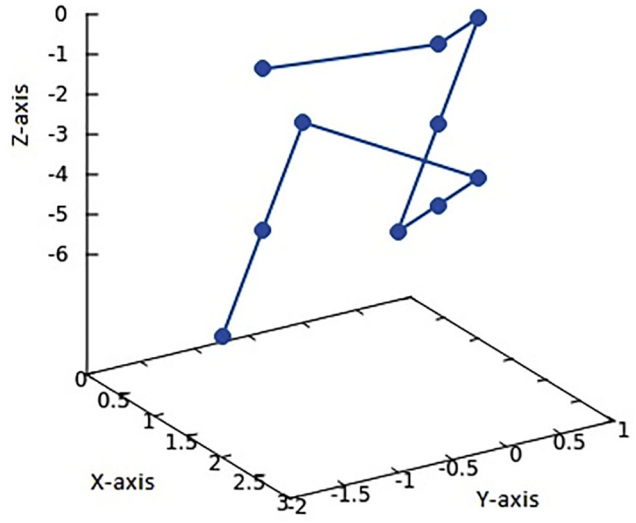

Fig. 2.7. 3D plot as per Randic et al.'s representation of the sequence ATGAA GGCAA.

Table 2.7.1
The coordinates as per Randic 3D plot.

Base	Coordinates		
	X	Y	Z
A	1	−1	−1
T	2	0	0
G	1	1	−1
A	2	0	−2
A	3	−1	−3
G	2	0	−4
G	1	1	−5
C	0	0	−4
A	1	−1	−5
A	2	−2	−6

Table 2.7.2
D_E/D_G matrix in Randic 3D model.

	A	T	G	A	A	G	G	C	A	A
A	0	1	0.5774	0.3333	0.4082	0.3830	0.4303	0.2736	0.2887	0.3333
T		0	1	0.5774	0.6383	0.5774	0.6000	0.4303	0.4286	0.4564
G			0	1	1	0.6383	0.5774	0.3830	0.4303	0.4880
A				0	1	0.5774	0.6383	0.4082	0.3830	0.4303
A					0	1	1	0.6383	0.4082	0.3830
G						0	1	0.5774	0.3333	0.4082
G							0	1	0.5774	0.6383
C								0	1	1
A									0	1
A										0

Table 3.1
Sequence descriptors of envelope gene of flaviviruses and globin genes.

Methods	Sequence Descriptors				Statistics		
	DENV2	WNV	YFV	ZIKV	Average	SD	SD%
Nandy	72.61	8.91	10.92	4.51	24.24	32.36	133.51
Randic 3D	221.99	113.57	144.37	144.93	156.22	46.24	29.60
Wang-Zhang	43.06	40.08	39.01	24.89	36.76	8.10	22.02
Ji-Li	30.05	42.51	37.93	31.94	35.61	5.69	15.99
Yau	508.81	515.83	507.03	521.15	513.21	6.52	1.27
Randic 2D	1487.07	1505.02	1480.92	1514.01	1496.75	15.39	1.03
Song-Tang	1485.51	1503.43	1479.42	1512.47	1495.21	15.37	1.03
	HbA1	HbB	HbD	HbG1	Average	SD	SD%
Nandy	47.31	31.86	29.09	11.13	29.85	14.83	49.68
Randic 3D	101.18	61.56	59.95	46.59	67.32	23.55	34.98
Ji-Li	88.33	78.17	66.41	35.41	67.08	22.93	34.19
Wang-Zhang	22.84	29.72	27.83	24.37	26.19	3.14	12.00
Song-Tang	429.20	444.27	444.3	444.38	440.54	7.56	1.72
Randic 2D	430.86	445.75	445.72	445.99	442.08	7.48	1.69
Yau	158.9	157.40	155.73	155.26	156.82	1.66	1.06

in the standard deviation (SD) as a percentage of the average descriptor values (SD%) for each method; while the Nandy and Randic 3D plot descriptors yield SD% 133.51 and 29.60, respectively, for the envelope genes; in case of the Song-Tang and Randic 2D methods, the SD% are 1.03 and 1.27, respectively. Higher SD% implies greater differences in the sequence descriptor values showing increased discriminatory power among related sequences. This is also borne out by the analysis of the human globin genes (Table 3.1) where, again, the Nandy and Randic 3D plots show the best discriminatory power among the seven methods tested. Our analysis, therefore, indicates that it would be advisable to consider methods like the Nandy plot, Randic 3D plot and others of the same ilk, not analysed in this brief commentary, that have the ability to distinguish among sequence differences for analysis of DNA/RNA sequences of a family of closely similar genes.

However, given a random collection of gene sequences of a variety of related and non-related sequences, the methods not meaningful in the above analysis could come into effect; for example, between the globin and flavivirus gene sets, the Randic 2D method or the Song-Tang method yield sequence descriptors that are different by at least one order of magnitude (Table 3.1). The non-similar sequence descriptors would tend to provide distinct separate groups where similar sequences, such as the family of flavivirus genes for instance, having almost identical descriptor values in these methods would cluster together, thus providing an easy way of discriminating between divergent sequences. In this context, it is relevant to recall that in the chemical domain it has long been known that no single topological index can discriminate graphs uniquely [19]. Perhaps, it may be worthwhile to introduce a concept of some “super-descriptor” on the lines of topological “super-index” first proposed by Bonchev, Mekenyan and Trinajstic [20] or extracted principal components (PCs) derived from the collection of individual sequence descriptors [21] as some combination of descriptors from various methods for discriminatory power of inter-gene and intra-gene sequences.

4. Conclusion

In graphical representation methods, the base composition and distribution within a nucleic acid sequence are visually projected and further numerically characterized by way of sequence descriptors which provide a quantitative description of the base pattern in the query sequence. Multiple methods are available to compute this. A comparative analysis of seven such methods - Yau method, Wang-Zhang method, Song-Tang method, Randic 3D method, Randic 2D method, Nandy method and Ji-Li method – in this brief commentary have elucidated that not all of them have the full capacity to discriminate between closely related sequences. In a retrospective study of flaviviral envelope gene and human globin genes, where one element of each sample set is known to be characteristically different from the rest, the Nandy and Randic 3D methods have yielded results that allow good discrimination between the sequences, whereas the other methods were unable to reflect similar outcome. However, on a broader scale, for classification of a set of close and distant sequences, the remaining five graphical representation methods are able to discern diverging sequences. We suggest that a kind of “super-descriptor”, which is a blend of elementary descriptors, could be considered to distinguish sufficiently between inter- and intra-gene sequences.

5. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Nandy, M. Harle, S.C. Basak, Mathematical descriptors of DNA sequences: development and applications, ARKIVOC 9 (2006) 211–238.
[2] M. Randic, J. Zupan, A.T. Balaban, D. Vikic-Topic, D. Plavsic, Graphical representation of proteins, Chem. Rev. 111 (2011) 790–862.
[3] I. Wiesner, D. Wiesnerová, 2D random walk representation of *Begonia* × *tuberhybrida* multiallelic loci used for germplasm identification, Biologia Plantarum 54 (2) (2010) 353–356.
[4] B. Liao, Y. Liu, R. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, Chem. Phys. Lett. 421 (2006) 313–318.
[5] A. Ghosh, A. Nandy, P. Nandy, Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase, BMC Struct. Biol. 10 (2010) 6.
[6] A. Ghosh, S. Chattopadhyay, M. Chawla-Sarkar, P. Nandy, A. Nandy, In silico study of rotavirus VP7 surface accessible conserved regions for antiviral drug/vaccine design, PLoS One (2012) 7(7).
[7] S. Dey, A. De, A. Nandy, Rational design of peptide vaccines against multiple types of human papillomavirus, Cancer Inform. 15 (S1) (2016) 1–16.
[8] S. Dey, A. Nandy, S.C. Basak, P. Nandy, S. Das, A bioinformatics approach to designing a Zika virus vaccine, Comput. Biol. Chem. 68 (2017) 143–152.
[9] S. Dey, P. Roy, A. Nandy, S.C. Basak, S. Das, Comparison of base distributions in Dengue, Zika and other flavivirus envelope and NS5 genes, in: Proceedings of the MOL2NET, International Conference on Multidisciplinary Sciences, Sciforum Electronic Conference Series, 3, 2017.
[10] D. Sen, S. Dasgupta, I. Pal, S. Manna, S.C. Basak, A. Nandy, G.D. Grunwald, Intercorrelation of major DNA/RNA sequence descriptors – a Preliminary Study, Curr. Comput.-Aided Drug Des. 12 (3) (2016) 216–228.
[11] 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.
[12] C. Raychaudhury, A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, J. Chem. Inf. Comput. Sci. 39 (1999) 243–247.
[13] S.S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.-K. Ho, DNA sequence representation without degeneracy, Nucl. Acids Res. 31 (2003) 3078–3080.
[14] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.
[15] J. Song, H. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization, J. Biochem. Biophys. Methods 63 (2005) 228–239.
[16] J. Wang, Y. Zhang, Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation, Chem. Phys. Lett. 423 (2006) 50–55.
[17] M. Ji, C. Li, TB curve, a new 2D graphical representation of DNA sequence, J. Math. Chem. 40 (2) (2006).
[18] M. Randic, M. Vracko, 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.
[19] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak, Discrimination of isomeric structures using information theoretic topological indices, J. Comput. Chem. 5 (6) (1984) 581–588.
[20] D. Bonchev, O.V. Mekenyan, N. Trinajstic, Isomer discrimination by topological information approach, J. Comput. Chem. 2 (2) (1981) 127–148.
[21] K. Balasubramanian, S.C. Basak, Characterization of isospectral graphs using graph invariants and derived orthogonal parameters, J. Chem. Inf. Comput. Sci. 38 (3) (1998) 367–373.