Journal of Biomedical Informatics 72 (2017) 1–7

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Spectral-dynamic representation of DNA sequences

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

[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

ARTICLE INFO

Article history:
Received 13 February 2017
Revised 3 May 2017
Accepted 1 June 2017
Available online 3 June 2017

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

ABSTRACT

A graphical representation of DNA sequences in which the distribution of a particular base $B = A, C, G, T$ is represented by a set of discrete lines has been formulated. The methodology of this approach has been borrowed from two areas of physics: spectroscopy and dynamics. Consequently, the set of discrete lines is referred to as the B-spectrum. Next, the B-spectrum is transformed to a rigid body composed of material points. In this way a *dynamic representation* of the DNA sequence has been obtained. The centers of mass of these rigid bodies, divided by their moments of inertia, have been taken as the descriptors of the spectra and, thus, of the DNA sequences. The performance of this method on a standard set of data commonly applied by authors introducing new approaches to bioinformatics (the first exons of β-globin genes of different species) proved to be very good.

1. Introduction

The fast increase of data in DNA databases stimulated the development of computational methods aiming at an analysis of this information.

The most commonly used program (with about 5000 citations each year) for the comparison of primary biological sequence information is BLAST (Basic Local Alignment Search Tool) [1].

A decade ago, Randić and coauthors introduced *graphical alignment* of biosequences methods [2,3]. It is worthwhile to point out that compared to computer based programs on protein alignment, these algorithms do not involve any empirical parameters or approximations (in contrast, the BLAST uses empirical parameters).

Protein Alignment Problem has been very recently solved exactly [4–6]. As we read in Ref. [6] a comparison of the exact solution (based on the use of matrices) and BLAST of two proteins having 170 amino acids, BLAST aligned 89 amino acids, while exact solution aligned 95 (i.e. six more).

Alignment-free methods constitute a fast developing branch of bioinformatics. These methods are of interdisciplinary character and researchers representing different areas of natural sciences bring to them new ideas. Therefore, reports on this subject appear in journals which traditionally have been assigned to physics, chemistry, biology, computer science, and many other fields of science, as for example [7–43]. Reviews may be found in Refs. [44–46].

The basic quantities used in these methods are called by the authors *descriptors*, i.e. some numerical values characterizing the biological (DNA, RNA, protein) sequences [47]. The descriptors can be related to some graphical structures (plots) which give graphical representations of the sequences. As a consequence, the same sequences can be compared both graphically and numerically. This kind of approach may be exemplified by developed by us *2D-Dynamic Representation of DNA Sequences* [48–55] and its 3-dimensional generalization [56,57]. We call these methods "dynamic" because the numerical description of the graphs is based on concepts taken from the classical dynamics. In particular, as one of the descriptors of the 2D-dynamic and 3D-dynamic graphs we introduced properly defined, respectively, 2D and 3D moments of inertia [48,56]. In the 3D case the degeneracy, resulting from the overlapping of the 2D-dynamic graphs, has been removed. Our method has also been generalized to three dimensions by Aram and Iranmanesh [58]. As a consequence, two different methods derived from *2D-Dynamic Representations of DNA Sequences* and having the same name (*3D-Dynamic Representation of DNA Sequences*) [56,58] are present in the literature.

The idea of characterizing biological sequences by moments of inertia, introduced by us for the 2D-dynamic graphs [48], has been adopted in several other methods. In particular, Yao et al. applied this idea representing protein sequences by 2D moments of inertia [59] and by 3D moments of inertia [60]. The 3D moments of inertia have also been applied to characterize graphs representing protein sequences by Hou et al. [61]. Recently, we have also introduced 20D moments of inertia as new characteristics of protein sequences [62].

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

http://dx.doi.org/10.1016/j.jbi.2017.06.001
1532-0464/

In the present work 1D moments of inertia are introduced as new characteristics of DNA sequences. These descriptors are constructed from some specific *spectra* representing the sequences. In a way, the present approach is related to the one used in our recent work on molecular spectra in which 1D moments of inertia have been proposed as new descriptors of infrared molecular spectra [63,64]. The aim of the present article is to demonstrate that a similar methodology can be applied to an arbitrary system of discrete objects, *a spectrum*, in particular to the spectrum representing the DNA sequence.

The new method is a branch of the graphical methods called by the authors *spectral representations* of biological sequences, as for example introduced by us *Four-Component Spectral Representation of DNA Sequences* [65,66]. These methods differ from each other by the mathematical definitions of the spectra and by the descriptors [67–72,65,66,73].

A set of descriptors defined for a single sequence constitute a data basis for this object. In the case of molecules a similar approach results in computational techniques known as Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) commonly used in computational pharmacology, toxicology, eco-toxicology [74,75]. We have recently demonstrated that 1D moments of inertia of molecular spectra can be applied in protein QSAR studies to predict environmentally relevant properties of chloronaphthalenes [64]. Analogously to the molecular descriptors, the descriptors of biological sequences have found their applications in QSAR studies, for representing other sources of information like mass spectra of blood serum in clinical proteomics and molecular dynamics trajectories [76–84].

The new descriptors of the DNA sequences proposed in this work, composed using the dynamic description of some properly defined abstract spectra, seem to be of importance in the context of their potential use in biomedical sciences. In particular, as it is shown in the present work, they can be applied to hierarchical cluster analysis.

2. Theory

Let us represent the distribution of a particular base in the DNA sequence by a series of lines. The length of each line is equal to 1 and its position in the series corresponds to the position of the base in the sequence. The graphical appearance of such a representation resembles an atomic, molecular, or stellar spectrum composed of a series of sharp spectral lines. Therefore it is referred to as a *spectral representation* of the sequence. In order to make this resemblance closer, the terminology used in spectroscopy has also been introduced.

Thus, the sequence corresponding to the base B, with $\mathrm{B} = \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}$ is called the *spectrum* of B or the B-spectrum, the position of the ith line in the B-spectrum is denoted v_i^{B}, $i = 1, 2, \ldots N_{\mathrm{B}}$, and called *frequency*.

The length of the line (in our case equal to 1 for all lines) is denoted $I_{\mathrm{B}}(v_i^{\mathrm{B}})$, $i = 1, 2, \ldots N_{\mathrm{B}}$, and called the *intensity* of the line.

Since N_{B} is equal to the number of lines in the B-spectrum, i.e. to the number of B bases in the sequence, we have

$$N_{\mathrm{A}} + N_{\mathrm{C}} + N_{\mathrm{G}} + N_{\mathrm{T}} = N, \tag{1}$$

where N is the length of the DNA sequence. For example, a model sequence ATGGTT is represented by the B-spectra composed of the following lines:

$$I_{\mathrm{A}}(v_1^{\mathrm{A}}) = 1, \qquad v_1^{\mathrm{A}} = 1,$$
$$I_{\mathrm{G}}(v_1^{\mathrm{G}}) = I_{\mathrm{G}}(v_2^{\mathrm{G}}) = 1, \qquad v_1^{\mathrm{G}} = 3, \quad v_2^{\mathrm{G}} = 4,$$
$$I_{\mathrm{T}}(v_1^{\mathrm{T}}) = I_{\mathrm{T}}(v_2^{\mathrm{T}}) = I_{\mathrm{T}}(v_3^{\mathrm{T}}) = 1, \qquad v_1^{\mathrm{T}} = 2, \quad v_2^{\mathrm{T}} = 5, \quad v_3^{\mathrm{T}} = 6,$$

and there are no lines in the C-spectrum.

The B-spectra corresponding to a model sequence ATGACTTTGCTGAGT are shown in Fig. 1.

Let us assume that the spectral lines of the B-spectrum are set vertical. Their projections to the horizontal axis give us four sets of points with the following coordinates:

1. $(v_1^{\mathrm{A}}, 0), (v_2^{\mathrm{A}}, 0), \ldots \left(v_{N_{\mathrm{A}}}^{\mathrm{A}}, 0\right)$,
2. $(v_1^{\mathrm{C}}, 0), (v_2^{\mathrm{C}}, 0), \ldots \left(v_{N_{\mathrm{C}}}^{\mathrm{C}}, 0\right)$,
3. $(v_1^{\mathrm{G}}, 0), (v_2^{\mathrm{G}}, 0), \ldots \left(v_{N_{\mathrm{G}}}^{\mathrm{G}}, 0\right)$,
4. $(v_1^{\mathrm{T}}, 0), (v_2^{\mathrm{T}}, 0), \ldots \left(v_{N_{\mathrm{T}}}^{\mathrm{T}}, 0\right)$.

Let us assign to each point $(v_i^{\mathrm{B}}, 0)$ a mass m_i^{B}. In this way we obtain four massive bodies composed of point masses distributed along the horizontal axis. The moments of inertia of these bodies are equal to

$$M_{\mathrm{B}} = \sum_{i=1}^{N_{\mathrm{B}}} m_i^{\mathrm{B}} \left(\tilde{v}_i^{\mathrm{B}}\right)^2, \tag{2}$$

where

$$\tilde{v}_i^{\mathrm{B}} = v_i^{\mathrm{B}} - v_a^{\mathrm{B}}, \tag{3}$$

and $(v_a^{\mathrm{B}}, 0)$ are the coordinates of the centers of mass of the bodies with

$$v_a^{\mathrm{B}} = \frac{1}{N_{\mathrm{B}}} \sum_{i=1}^{N_{\mathrm{B}}} v_i^{\mathrm{B}}. \tag{4}$$

Hereafter, for simplicity, $m_i^{\mathrm{B}} = 1$ has been set for each point. Consequently, the total mass of a spectrum is equal to the total number of points

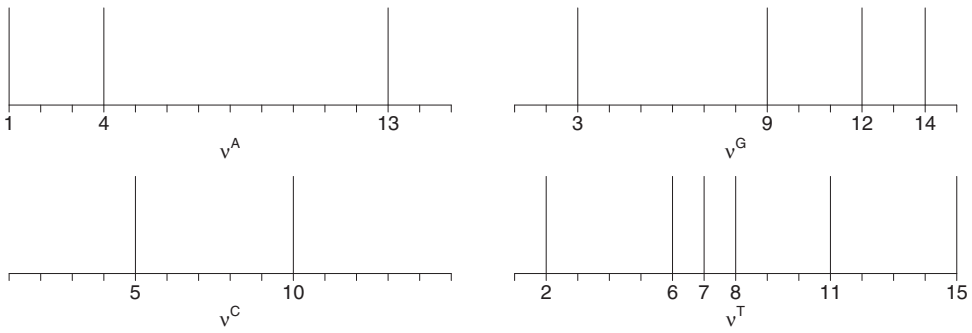

Fig. 1. B-spectra corresponding to a model sequence ATGACTTTGCTGAGT.

$$\sum_{i=1}^{N_B} m_i = N_B, \tag{5}$$

i.e. to the total number of lines in the spectrum. It is also convenient to define the *normalized moments of inertia*

$$r_B = \sqrt{\frac{M_B}{N_B}}. \tag{6}$$

The coordinates of the center of mass divided by the normalized moments of inertia,

$$D_B = \frac{\nu_a^B}{r_B}, \tag{7}$$

are proposed as new descriptors of the DNA sequences and standard distances between vectors $D_B^\alpha = [D_A^\alpha, D_C^\alpha, D_G^\alpha, D_T^\alpha]$ and $D_B^\beta = [D_A^\beta, D_C^\beta, D_G^\beta, D_T^\beta]$ in the 4-dimensional space are taken as the similarity measures between a pair of DNA sequences labeled as α and β. In particular, in this work the following distances are considered:

- The Euclidean distance

$$S_{EU}(\alpha,\beta) = S_{EU}(\beta,\alpha) = \left[\sum_B \left|D_B^\alpha - D_B^\beta\right|^2\right]^{1/2}, \tag{8}$$

- The Canberra distance

$$S_{CAN}(\alpha,\beta) = S_{CAN}(\beta,\alpha) = \sum_B \frac{\left|D_B^\alpha - D_B^\beta\right|}{\left|D_B^\alpha\right| + \left|D_B^\beta\right|}, \tag{9}$$

- The Minkowski distance of the 6th order

$$S_{MINK}(\alpha,\beta) = S_{MINK}(\beta,\alpha) = \left[\sum_B \left|D_B^\alpha - D_B^\beta\right|^6\right]^{1/6}, \tag{10}$$

- The Manhattan distance

$$S_{MAN}(\alpha,\beta) = S_{MAN}(\beta,\alpha) = \sum_B \left|D_B^\alpha - D_B^\beta\right|, \tag{11}$$

- The Maximum distance

$$S_{MAX}(\alpha,\beta) = S_{MAX}(\beta,\alpha) = \max_B \left|D_B^\alpha - D_B^\beta\right|. \tag{12}$$

3. Results and discussion

The performance of the new method has been tested using the standard set of data commonly used when introducing new approaches, i.e. the first exons of β-globin gene of different species (Table 1).

Fig. 2 shows the spectral representation of the first exon of β-globin gene of lemur. The locations and the density of lines give an information about the distribution of particular bases along the DNA sequence. It is clearly seen that the number of G bases is larger than of the other ones.

Table 2 shows the descriptors calculated according to Eq. (7) for the considered DNA sequences. They can be compared separately for each B.

As in many other methods, also here, for the first exons of β-globin gene larger degree of similarity is obtained in the case of human-gorilla rather than of human-chimpanzee [19–21,38–40,57]. In the present method

$$\left|D_A^{human} - D_A^{gorilla}\right| < \left|D_A^{human} - D_A^{chimpanzee}\right|,$$
$$\left|D_C^{human} - D_C^{gorilla}\right| < \left|D_C^{human} - D_C^{chimpanzee}\right|,$$
$$\left|D_G^{human} - D_G^{gorilla}\right| \approx \left|D_G^{human} - D_G^{chimpanzee}\right|,$$
$$\left|D_T^{human} - D_T^{gorilla}\right| < \left|D_T^{human} - D_T^{chimpanzee}\right|.$$

Table 1
The first exons of β-globin gene of different species.

k	Species	Gene ID	N
1	Human	U01317	92
2	Gorilla	X61109	93
3	Chimpanzee	X02345	105
4	Bovine	X00376	86
5	Goat	M15387	86
6	Rat	X06701	92
7	Rabbit	V00882	92
8	Mouse	V00722	93
9	Lemur	M15734	92
10	Opossum	J03643	92
11	Gallus	V00409	92

Table 2
Descriptors representing the DNA sequences.

Species	D_A	D_C	D_G	D_T
Human	1.6866	1.5320	1.9797	1.7252
Goat	1.7202	1.6061	1.8827	1.7098
Opossum	1.7335	1.8818	1.8441	1.6278
Gallus	1.6757	1.9309	1.7786	1.5837
Lemur	1.7260	1.7121	1.9806	1.5569
Mouse	1.7007	1.7466	1.8784	1.6961
Rabbit	1.7696	1.6168	1.9209	1.5968
Rat	1.6558	1.7006	1.9136	1.7201
Gorilla	1.6866	1.5320	1.9854	1.7252
Bovine	1.7202	1.5429	1.9045	1.7517
Chimpanzee	1.6650	1.5101	1.9856	1.7271

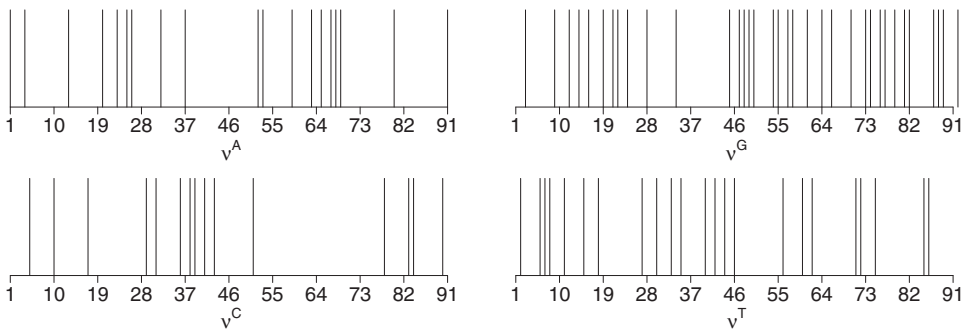

Fig. 2. B-spectra representing the first exon of β-globin gene of lemur.

The descriptors can be considered as components of 4-dimensional vectors in the similarity measures defined in Eqs. (8)–(12). Tables 3–7 show the similarity/dissimilarity matrices using different similarity measures. In all cases the similarity value for human-gorilla is smaller than for human-chimpanzee. Note, that the smaller similarity value in these matrices means that the degree of similarity is larger.

This is clearly seen in Fig. 3, where the similarity values S for human-other species, normalized to $S^{\text{human-gallus}} = 1$, are displayed. The similarity values are slightly different for different similarity measures. As a consequence, the cluster dendrograms which visualize the similarity/dissimilarity matrices may also be different for different measures. Two most distinct dendrograms, corresponding to the Euclidean and to the Canberra measures, are plotted in

Table 3
Similarity/dissimilarity matrix obtained using the Euclidean measure.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.0000										
Goat	0.1276	0.0000									
Opossum	0.3904	0.2905	0.0000								
Gallus	0.4687	0.3663	0.1095	0.0000							
Lemur	0.2497	0.2104	0.2291	0.3032	0.0000						
Mouse	0.2395	0.1426	0.1587	0.2391	0.1780	0.0000					
Rabbit	0.1844	0.1295	0.2799	0.3576	0.1270	0.1824	0.0000				
Rat	0.1838	0.1190	0.2285	0.3004	0.1903	0.0771	0.1877	0.0000			
Gorilla	0.0056	0.1319	0.3924	0.4712	0.2497	0.2420	0.1863	0.1859	0.0000		
Bovine	0.0872	0.0789	0.3661	0.4434	0.2691	0.2137	0.1793	0.1736	0.0921	0.0000	
Chimp.	0.0314	0.1522	0.4156	0.4905	0.2712	0.2640	0.2085	0.2040	0.0308	0.1063	0.0000

Table 4
Similarity/dissimilarity matrix obtained using the Canberra measure.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.0000										
Goat	0.0631	0.0000									
Opossum	0.1807	0.1178	0.0000								
Gallus	0.2147	0.1716	0.0616	0.0000							
Lemur	0.1186	0.1058	0.1073	0.1371	0.0000						
Mouse	0.1044	0.0528	0.0766	0.1191	0.0867	0.0000					
Rabbit	0.1047	0.0617	0.1161	0.1584	0.0691	0.0998	0.0000				
Rat	0.0798	0.0588	0.1195	0.1472	0.0911	0.0430	0.0976	0.0000			
Gorilla	0.0014	0.0645	0.1821	0.2161	0.1196	0.1058	0.1061	0.0813	0.0000		
Bovine	0.0404	0.0379	0.1556	0.2093	0.1321	0.0906	0.0881	0.0792	0.0418	0.0000	
Chimp.	0.0157	0.0788	0.1963	0.2238	0.1338	0.1200	0.1203	0.0826	0.0143	0.0550	0.0000

Table 5
Similarity/dissimilarity matrix obtained using the Minkowski measure.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.0000										
Goat	0.1000	0.0000									
Opossum	0.3500	0.2757	0.0000								
Gallus	0.4001	0.3251	0.0718	0.0000							
Lemur	0.1961	0.1572	0.1767	0.2371	0.0000						
Mouse	0.2150	0.1405	0.1355	0.1866	0.1427	0.0000					
Rabbit	0.1317	0.1131	0.2650	0.3146	0.0964	0.1342	0.0000				
Rat	0.1688	0.0961	0.1819	0.2334	0.1635	0.0520	0.1349	0.0000			
Gorilla	0.0056	0.1050	0.3501	0.4003	0.1961	0.2152	0.1318	0.1688	0.0000		
Bovine	0.0753	0.0641	0.3391	0.3886	0.2069	0.2038	0.1552	0.1579	0.0809	0.0000	
Chimp.	0.0244	0.1123	0.3719	0.4219	0.2126	0.2369	0.1407	0.1907	0.0244	0.0825	0.0000

Table 6
Similarity/dissimilarity matrix obtained using the Manhattan measure.

Species	Human	Goat	Opposum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.0000										
Goat	0.2202	0.0000									
Opossum	0.6296	0.4094	0.0000								
Gallus	0.7524	0.5994	0.2165	0.0000							
Lemur	0.3888	0.3627	0.3846	0.4979	0.0000						
Mouse	0.3591	0.1779	0.2705	0.4215	0.3013	0.0000					
Rabbit	0.3550	0.2113	0.4089	0.5634	0.2385	0.3405	0.0000				
Rat	0.2706	0.2003	0.4206	0.5216	0.3119	0.1501	0.3282	0.0000			
Gorilla	0.0056	0.2258	0.6353	0.7580	0.3926	0.3648	0.3606	0.2763	0.0000		
Bovine	0.1461	0.1270	0.5364	0.7264	0.4459	0.3049	0.2946	0.2629	0.1518	0.0000	
Chimp.	0.0513	0.2715	0.6809	0.7819	0.4383	0.4104	0.4063	0.2787	0.0456	0.1937	0.0000

Table 7
Similarity/dissimilarity matrix obtained using the Maximum measure.

Species	Human	Goat	Opposum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.0000										
Goat	0.0970	0.0000									
Opossum	0.3498	0.2757	0.0000								
Gallus	0.3989	0.3248	0.0655	0.0000							
Lemur	0.1801	0.1529	0.1696	0.2188	0.0000						
Mouse	0.2146	0.1405	0.1351	0.1843	0.1393	0.0000					
Rabbit	0.1284	0.1130	0.2649	0.3141	0.0953	0.1298	0.0000				
Rat	0.1686	0.0945	0.1811	0.2303	0.1632	0.0460	0.1233	0.0000			
Gorilla	0.0056	0.1027	0.3498	0.3989	0.1801	0.2146	0.1284	0.1686	0.0000		
Bovine	0.0752	0.0632	0.3389	0.3881	0.1948	0.2038	0.1548	0.1578	0.0808	0.0000	
Chimp.	0.0219	0.1029	0.3717	0.4208	0.2021	0.2366	0.1302	0.1906	0.0219	0.0811	0.0000

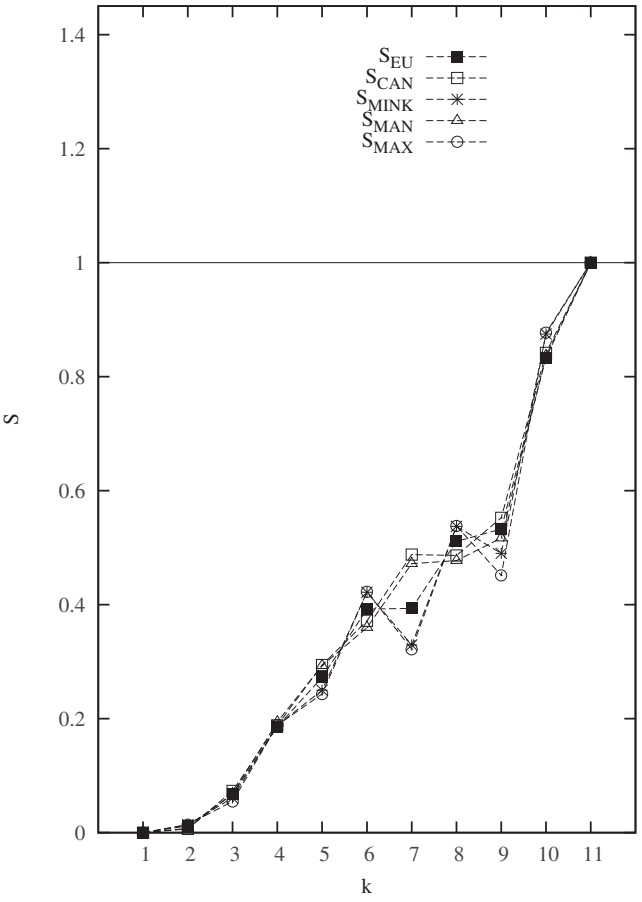

Fig. 3. Similarity values human-other species obtained using different measures (k numbers the species according to Table 1).

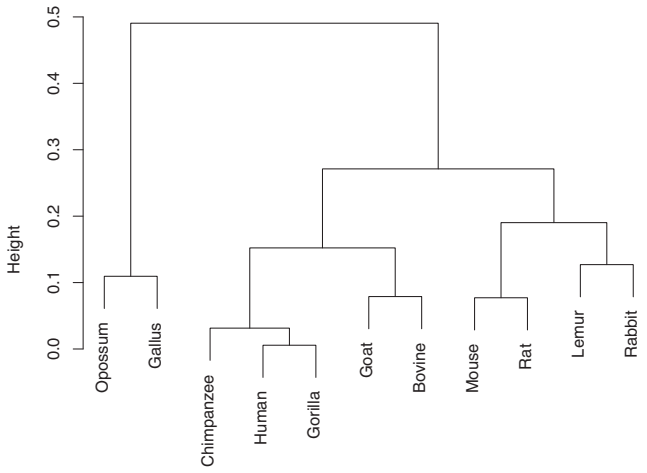

Fig. 4. Cluster dendrogram obtained using the Euclidean measure.

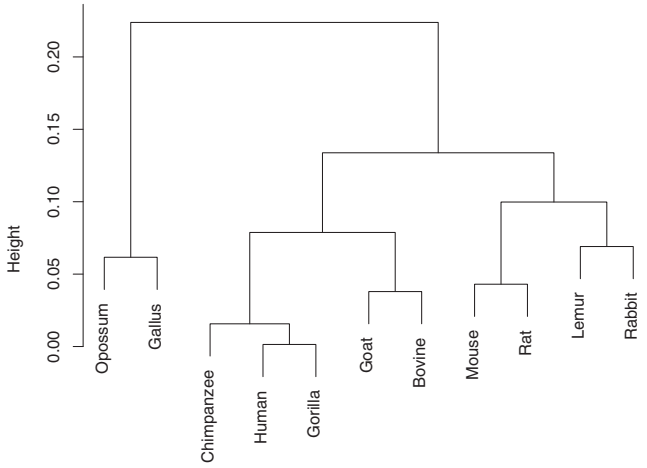

Fig. 5. Cluster dendrogram obtained using the Canberra measure.

Figs. 4 and 5, respectively. The dendrograms have been generated using RKWard version 0.6.1 fronted to the R statistics language with the library stats in the function hclust [85]. As expected, the differences between the dendrograms are quantitative rather than qualitative: they differ by the heights only.

In Fig. 6 the similarity values for human-other species, normalized to $S^{\text{human-gallus}} = 1$ for the first exons of β-globin genes, obtained by different methods using the Euclidean measure are compared with S_{EU} values. As one can see, the dispersion of the results given by different methods is rather large. Different methods may describe different aspects of similarity of the DNA sequences. In general, different descriptors may be important in different situations. They form a basis for the QSAR studies.

Summarizing, the new method is a convenient and intuitive graphical tool for the comparison of DNA sequences. The locations and the density of the lines in the DNA spectrum carry the information about the distributions of the bases along the DNA sequence. Calculating the descriptors is not computationally demanding, also for long sequences. There are no restrictions concerning the lengths of the sequences. The dynamic description of the DNA spectra will enrich the numerical characterization of the considered objects which is particularly important in the QSAR studies analogously as in the case of molecules [63,64].

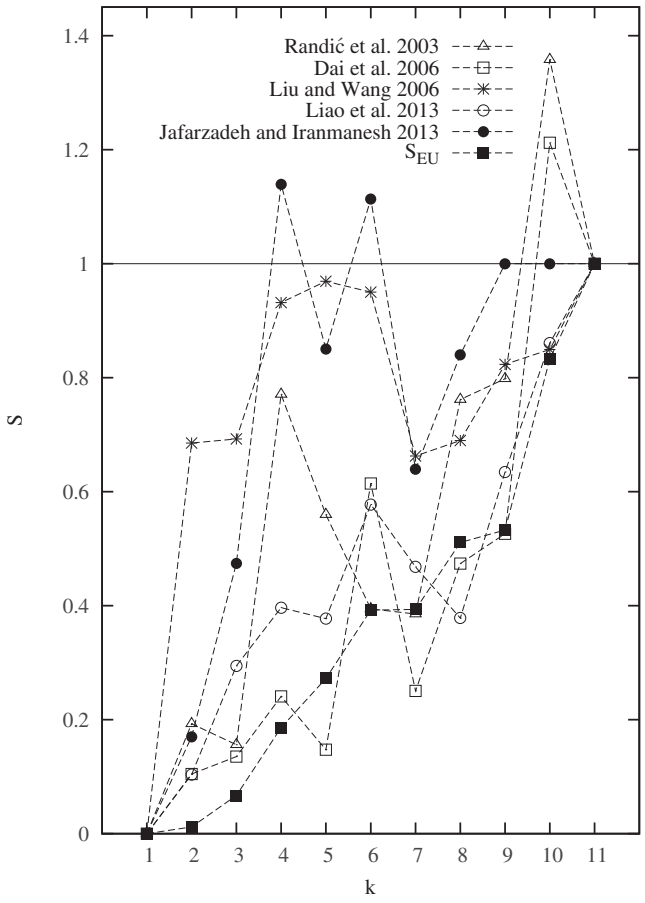

Fig. 6. Similarity values human-other species obtained by different methods using the Euclidean measure (*k* numbers the species according to Table 1).

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[2] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: graphical alignment of DNA sequences, Chem. Phys. Lett. 431 (2006) 375–379.
[3] M. Randić, On a geometry-based approach to protein sequence alignment, J. Math. Chem. 43 (2008) 756–772.
[4] M. Randić, Very efficient search for nucleotide alignments, J. Comput. Chem. 34 (2013) 77–82.
[5] M. Randić, Very efficient search for protein alignment – VESPA, J. Comput. Chem. 33 (2012) 702–707.
[6] M. Randić, T. Pisanski, Proteins alignment: exact versus approximate. An illustration, J. Comput. Chem. 36 (2015) 1069–1074.
[7] 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.
[8] H.I. Jeffrey, Chaos game representation of gene structure, Nucl. Acid Res. 18 (1990) 2163–2170.
[9] 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.
[10] M. Randić, M. Vračko, A. Nandy, S.C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inform. Comput. Sci. 40 (2000). 1325–1244.
[11] 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.
[12] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, SAR QSAR Environ. Res. 15 (2004) 147–157.
[13] M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representation of proteins, Period. Biol. 107 (2005) 403–414.
[14] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. 419 (2006) 528–532.
[15] M. Randić, M. Novič, D. Vikić-Topić, D. Plavšić, Novel numerical and graphical representation of DNA sequences and proteins, SAR QSAR Environ. Res. 17 (2006) 583–595.
[16] M. Randić, J. Zupan, D. Vikić-Topić, Graphical representation of proteins by star-like graphs, J. Mol. Graph. Modell. 26 (2007) 290–305.
[17] M. Randić, J. Zupan, A.T. Balaban, D. Vikić-Topić, D. Plavšić, Graphical representation of proteins, Chem. Rev. 111 (2011) 790–862.
[18] J. Song, K. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization, J. Biochem. Bioph. Meth. 63 (2005) 228–239.
[19] R. Chi, K. Ding, Novel 4D numerical representation of DNA sequences, Chem. Phys. Lett. 407 (2005) 63–67.
[20] Q. Dai, X. Liu, T. Wang, A novel graphical representation of DNA sequences and its application, J. Mol. Graph. Model. 25 (2006) 340–344.
[21] Y. Liu, T. Wang, Related matrices of DNA primary sequences based on triplets of nucleic acid bases, Chem. Phys. Lett. 417 (2006) 173–178.
[22] B. Liao, K. Ding, A 3D graphical representation of DNA sequences and its application, Theor. Comput. Sci. 358 (2006) 56–64.
[23] B. Liao, W. Zhu, Analysis of similarity/dissimilarity of DNA primary sequences based on condensed matrices and information entropies, Curr. Comput. Aided Drug Des. 2 (2006) 95–103.
[24] B. Liao, Y. Liu, R. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, Chem. Phys. Lett. 421 (2006) 313–318.
[25] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, J. Comput. Chem. 27 (2006) 1196–1202.
[26] W. Wang, B. Liao, T. Wang, A graphical method to construct phylogenetic tree, Int. J. Quant. Chem. 106 (2006) 1998–2005.
[27] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, MATCH Commun. Math. Comput. Chem. 56 (2006) 209–216.
[28] B. Liao, X. Shan, W. Zhu, R. Li, Phylogenetic tree construction based on 2D graphical representation, Chem. Phys. Lett. 422 (2006) 282–288.
[29] B. Liao, C. Zeng, F. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, MATCH Commun. Math. Comput. Chem. 59 (2008) 647–652.
[30] W. Chen, B. Liao, Y. Liu, W. Zhu, Z. Su, A numerical representation of DNA sequence and its applications, MATCH Commun. Math. Comput. Chem. 60 (2008) 291–300.
[31] 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.
[32] G. Huang, B. Liao, Y. Li, Z. Liu, H-L curve: a novel 2D graphical representation for DNA sequences, Chem. Phys. Lett. 462 (2008) 129–132.
[33] Z. Liu, B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, MATCH Commun. Math. Comput. Chem. 61 (2009) 541–552.
[34] 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.
[35] Z. Liu, B. Liao, W. Zhu, G. Huang, A 2D graphical representation of DNA sequence based on dual nucleotides and its application, Int. J. Quant. Chem. 109 (2009) 948–958.
[36] G. Huang, B. Liao, R. Li, Similarity studies of DNA sequences based on a new 2D graphical representation, Biophys. Chem. 143 (2009) 55–59.
[37] 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.
[38] 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.
[39] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, Math. Biosci. 241 (2013) 217–224.
[40] 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.
[41] V. Aram, A. Iranmanesh, Z. Majid, Spider representation of DNA sequences, J. Comput. Theor. Nanos. 11 (2014) 418–420.
[42] Y.W. Liu, Y. Peng, A novel technique for analyzing the similarity and dissimilarity of DNA sequences, Genet. Mol. Res. 13 (2014) 570–577.
[43] C. Yin, X.E. Yin, J. Wang, A novel method for comparative analysis of DNA sequences by Ramanujan-Fourier transform, J. Comput. Biol. 21 (2014) 867–879.
[44] D. Bielińska-Wąż, Graphical and numerical representations of DNA sequences: statistical aspects of similarity, J. Math. Chem. 49 (2011) 2345–2407.
[45] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, Int. J. Quant. Chem. 113 (2013) 2413–2446.
[46] O. Bonham-Carter, J. Steele, D. Bastola, Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis, Brief. Bioinform. 15 (2014) 890–905.
[47] A. Nandy, M. Harle, S.C. Basak, Mathematical descriptors of DNA sequences: development and application, Arkivoc ix (2006) 211–238.
[48] D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. 442 (2007) 140–144.

[49] 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.
[50] D. Bielińska-Wąż, P. Wąż, T. Clark, Similarity studies of DNA sequences using genetic methods, Chem. Phys. Lett. 445 (2007) 68–73.
[51] 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, vol. 2, American Institute of Physics, 2007, pp. 28–30.
[52] 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.
[53] 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, vol. 1773, American Institute of Physics, 2016, pp. 060004-1–060004-5.
[54] 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.
[55] 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.
[56] P. Wąż, D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (2014) 2141.
[57] P. Wąż, D. Bielińska-Wąż, Non-standard similarity/dissimilarity analysis of DNA sequences, Genomics 104 (2014) 464–471.
[58] V. Aram, A. Iranmanesh, 3D-dynamic representation of DNA sequences, MATCH Commun. Math. Comput. Chem. 67 (2012) 809–816.
[59] Y.-H. Yao, Q. Dai, Ch. Li, P.-A. He, X.-Y. Nan, Y.-Z. Zhang, Analysis of similarity/dissimilarity of protein sequences, Proteins-Struct. Funct. Bioinf. 73 (2008) 864–871.
[60] Y.-H. Yao, S. Yan, J. Han, Q. Dai, P.-A. He, A novel descriptor of protein sequences and its application, J. Theor. Biol. 347 (2014) 109–117.
[61] W. Hou, Q. Pan, M. He, A new graphical representation of protein sequences and its applications, Physica A 444 (2016) 996–1002.
[62] A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16–23.
[63] P. Wąż, D. Bielińska-Wąż, Moments of inertia of spectra and distribution moments as molecular descriptors, MATCH Commun. Math. Comput. Chem. 70 (2013) 851–865.
[64] K. Jagiełło, T. Puzyn, P. Wąż, D. Bielińska-Wąż, Moments of inertia of spectra as descriptors for QSAR/QSPR, in: I. Gutman (Ed.), Topics in Chemical Graph Theory, Univ. Kragujevac, Kragujevac, 2014, pp. 151–162.
[65] D. Bielińska-Wąż, Four-component spectral representation of DNA sequences, J. Math. Chem. 47 (2010) 41–51.
[66] D. Bielińska-Wąż, S. Subramaniam, Classification studies based on a spectral representation of DNA, J. Theor. Biol. 266 (2010) 667–674.
[67] Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.
[68] J. Zupan, M. Randić, Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations, J. Chem. Inform. Model. 45 (2005) 309–313.
[69] M. Randić, Spectrum-like graphical representation of DNA based on codons, Acta Chim. Slov. 53 (2006) 477–485.
[70] M. Randić, D. Plavšić, Novel spectral representation of RNA secondary structure without loss of information, Chem. Phys. Lett. 476 (2009) 277–280.
[71] M. Randić, M. Vračko, M. Novič, D. Plavšić, Spectral representation of reduced protein models, SAR QSAR Environ. Res. 20 (2009) 415–427.
[72] Z. Zhang, L. Liu, J. Li, Z. Zhang, Spectral representation of protein sequences, J. Comput. Theor. Nanos. 8 (2011) 1335–1339.
[73] Y. Yao, S. Yan, H. Xu, J. Han, X. Nan, P.A. He, Q. Dai, Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation, Evol. Bioinform. Online 10 (2014) 87–96.
[74] J. Verma, V.M. Khedkar, E.C. Coutinho, 3D-QSAR in drug design – a review, Curr. Top. Med. Chem. 10 (2010) 95–115.
[75] A. Lombardo, O. Schifanella, A. Roncaglioni, E. Benfenati, Quantitative structure-activity relationship (QSAR) in ecotoxicology, in: J. Frard, C. Blaise (Eds.), Encyclopedia of Aquatic Ecotoxicology, Springer, Netherlands, 2013, pp. 945–956.
[76] G. Agüero-Chapín, A. Antunes, F.M. Ubeira, K.C. Chou, H. González-Díaz, Comparative study of topological indices of macro/supramolecular RNA complex networks, J. Chem. Inform. Model. 48 (2008) 2265–2277.
[77] M.A. Dea-Ayuela, Y. Pérez-Castillo, A. Meneses-Marcel, F.M. Ubeira, F. Bolas-Fernández, K.C. Chou, H. González-Díaz, HP-lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence, Bioorg. Med. Chem. 16 (2008) 7770–7776.
[78] S. Vilar, H. González-Díaz, L. Santana, E. Uriarte, QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks, J. Comput. Chem. 29 (2008) 2613–2622.
[79] M. Cruz-Monteagudo, H. González-Díaz, F. Borges, E.R. Dominguez, M.N. Cordeiro, 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. Quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy, Chem. Res. Toxicol. 21 (2008) 619–632.
[80] L.G. Pérez-Montoto, L. Santana, H. González-Díaz, Scoring function for DNA-drug docking of anticancer and antiparasitic compounds based on spectral moments of 2D lattice graphs for molecular dynamics trajectories, Eur. J. Med. Chem. 44 (2009) 4461–4469.
[81] S. Vilar, H. González-Díaz, L. Santana, E. Uriarte, A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer, J. Theor. Biol. 261 (2009) 449–458.
[82] 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.
[83] A. Perez-Bello, C.R. Munteanu, F.M. Ubeira, A.L. De Magalhaes, E. Uriarte, H. González-Díaz, Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices, J. Theor. Biol. 256 (2009) 458–466.
[84] H. González-Díaz, M.A. Dea-Ayuela, L.G. Pérez-Montoto, F.J. Prado-Prado, G. Agüero-Chapín, F. Bolas-Fernández, R.I. Vazquez-Padrón, F.M. Ubeira, QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein, Mol Divers. 14 (2010) 349–369.
[85] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement ward criterion?, J Classif. 31 (2014) 274–295.