J Math Chem (2011) 49:2345–2407
DOI 10.1007/s10910-011-9890-8

ORIGINAL PAPER

Graphical and numerical representations of DNA sequences: statistical aspects of similarity

Dorota Bielińska-Wąż

Received: 18 February 2011 / Accepted: 22 July 2011 / Published online: 28 August 2011

Abstract New approaches aiming at a detailed similarity/dissimilarity analysis of DNA sequences are formulated. Several corrections that enrich the information which may be derived from the alignment methods are proposed. The corrections take into account the distributions along the sequences of the aligned bases (neglected in the standard alignment methods). As a consequence, different aspects of similarity, as for example asymmetry of the gene structure, may be studied either using new similarity measures associated with four-component spectral representation of the DNA sequences or using alignment methods with corrections introduced in this paper. The corrections to the alignment methods and the statistical distribution moment-based descriptors derived from the four-component spectral representation of the DNA sequences are applied to similarity/dissimilarity studies of β-globin gene across species. The studies are supplemented by detailed similarity studies for histones H1 and H4 coding sequences. The data are described according to the latest version of the EMBL database. The work is supplemented by a concise review of the state-of-art graphical representations of DNA sequences.

Keywords Graphical representations of DNA sequences · Descriptors · Similarity/dissimilarity analysis of DNA sequences

1 Introduction

In an article published by Fuchs in Nature in 2002 we read "Future generations may be able to determine whether the sequencing of the human genome in 2001 indeed led to a paradigm shift in biology and biomedicine as some predicted, or whether

D. Bielińska-Wąż (✉)
Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, 87-100 Toruń, Poland
e-mail: dsnake@fizyka.umk.pl

the impact of this event was more gradual instead" [1]. The author observes that "so far, the history of biology has been characterized by a continous shift from the whole organism down to the molecular level, from the descriptive characterization of species over macroscopic observations and morphological and physiological studies to today's molecular dissection of individual genes".

Novel experimental techniques require new computational methods. In order to create good models describing the experimental results, researchers from different areas of science joined computational biology and medical sciences. As a consequence, a new interdisciplinary field adapting methods from many different branches of mathematics, physics, chemistry, and computer science emerged.

A fundamental task coming from sequencing is to understand the code written in the sequence of four letters. A lot has been done to reveal some global characteristics of long DNA sequences. For example Herzel et al. [2] created a model that describes thousands of nearly identical dispersed repetitive sequences present in DNA sequences of higher organisms. The hypothetical model sequences consist of independent equidistributed symbols with randomly interspersed repeats. The model that can be analyzed analytically predicts that the entropy of DNA sequences measuring the information content is much lower than suggested by earlier empirical studies.

A systematic analysis of statistical properties of coding and noncoding DNA sequences has been performed by Mantegna et al. [3]. The authors compared the statistical behavior of coding and noncoding regions in eukaryotic and viral DNA sequences by adapting two tests developed for the analysis of natural languages and symbolic sequences. The authors analyzed some similarities and dissimilarities of statistical properties of coding and noncoding regions. In particular they found that for the three chromosomes they studied, the statistical properties of noncoding regions appear to be closer to those observed in natural languages than those of the coding regions.

Statistical studies aiming at characterization of correlation structures of DNA sequences has been a subject of many studies (for review see [4,5]). In particular Foss [6] using spectral density of individual base positions demonstrated long-range fractal correlations as well as short-range periodicities. Arneodo et al. [7] used the wavelet transform to demonstrate the existence of long-correlations in genes containing introns and noncoding regions. Buldyrev et al. [8] in order to answer the question in computational molecular biology whether long-range correlations are present in both coding and noncoding DNA sequences have used standard Fourier transform analysis and detrended fluctuation analysis. For that purpose, the authors performed analysis of the sequences available in GenBank in 1995. For noncoding sequences, they obtained the presence of long-range correlations. Azbel in his work [9] demonstrated a universality in a DNA statistical structure using an autocorrelation function. However, no long-range correlations have been found in any of the studied DNA sequences. Peng et.al. [10] studied long-range correlations by constructing a map of the nucleotide sequence onto a walk which they referred to as a *DNA walk*. Using such an approach they found long-range correlations in intron-containing genes and in nontranscribed regulatory DNA sequences, but not in complementary DNA sequence or intron-less genes. Visualization technique proposed by Peng et al. is based on a one dimensional DNA walk showing the relative occurrence of purines and pyrimidynes along the

sequence. Silverman and Linsker introduced vectorial representation of the bases in three dimensions [11]. They used the unit vectors of 3D space to construct a Fourier transform. Such Fourier transform graphs representing the sequences were used as measures of DNA periodicity. Another visualization technique based on DNA walk plotted in three-dimensional Cartesian coordinate system has been introduced by Berger [5]. In his work Berger gave also a good review of visualization techniques based on DNA walk and their applications for an analysis of DNA sequences i.e. a study of correlation information, sequence periodicities, and other sequence characteristics. More examples of studies focused on statistical properties of DNA sequences and also on their biological interpretation may be found in [12,13].

Another class of studies is developing methods aiming at detailed sequence comparisons. Most commonly used in computational biology and medical sciences are global and local alignment methods, for example Clustal W [14], Blast [15], Needleman-Wunsch algorithm [16], and T-Coffee [17] (for review see [18,19]).

An alternative to the alignment methods are *alignment-free* methods that can be divided into two groups: numerical similarity/dissimilarity analysis of DNA sequences and similarity/dissimilarity analysis based on graphical representations of DNA sequences. There is a variety of numerical alignment-free methods (for a review up to 2003 see [20]). Recently new numerical alternative methods have been developed, as for example [21–31]. Another group within numerical alignment-free methods are multidimensional graphical representations. Conceptually, they are analogous to the graphical representations but their visualization is difficult (if possible at all). In particular 4D numerical representations [32–34], 5D representation [35], 6D representation [36] have been introduced.

Due to interdisciplinary character of research on DNA, many groups of methods have been developed independently and very often without any knowledge about analogous results obtained in different groups of scientists. In particular, DNA walk has been independently discovered by the scientists working on statistical properties of DNA sequences [5] and by scientists working on graphical representations. Even among researchers working on graphical representations one can find analogous visualization tools discovered independently (see subsequent chapters).

This work is focused on graphical representations of DNA sequences. Biological sequences are often very long, and it is not obvious how to represent them graphically in an easy way that shows the main features of these objects. The size of the plots is restricted by the human abilities of perception. How to restrict the graphs representing the sequences to two-dimensional plots and how to avoid degeneracies has been the subject of numerous studies which resulted in many graphical representations (see subsequent chapters). Graphical representations offer both numerical and visualization tool for similarity/dissimilarity analysis. These methods are still restricted to small groups of users. Computing codes calculating optimal sequence alignment are implemented using dynamic programming and are freely accessible in the internet and that makes them attractive for potential users. However, they are computationally expensive, and methodologically offer too simplistic similarity/dissimilarity analysis. They restrict the multidimensional similarity space of complex objects and show only one aspect of similarity. It becomes more and more popular to replace the alignment methods by alternative ones. In particular, Hönl and Ragan consider numerical

alignment-free methods that can replace multiple-sequence alignment to infer a phylogenetic tree that represents the history of a set of molecular sequences [37]. Graphical representations have been also used for the construction of phylogenetic trees. Since multiple alignment strategy does not work for all types of data, Liao et al. [38] proposed to use the similarity matrix based on their 2D graphical representation of DNA sequences [39] to construct phylogenetic tree. The authors consider mitochondrial sequences belonging to different species. The same graphical representation has been also used by the authors to obtain the phylogenetic relationships of H5N1 avian influenza virus [40] and the phylogenetic relationships of coronaviruses [41]. Another 2D graphical representation [42] has been used by Yu et al. to construct the phylogenetic tree of coronaviruses and lentiviruses [43]. 3D graphical representation has been also used to construct a phylogenetic tree [44]. Wang and Zhang studied molecular phylogeny of H5N1 avian influenza viruses in Asia using 2D and 3D graphical representations of DNA sequences [45]. Graphical representations of DNA sequences have been also generalized for the analysis of similarity/dissimilarity between RNA secondary structures, as for example [46,47]. 2D graphical representation has been used for the characterization of the neuraminidase RNA sequences of H5N1 [48,49] and of H1N1 [50] strains. Also graphical representations of the proteins have been created [51–54].

Graphical representations of the biological sequences (DNA, RNA, proteins) can be applied to all problems that require similarity/dissimilarity analysis. Similarity analysis is not unique to sequences in biology. For instance, the problem of similarity has been developed and applied in computational pharmacology and has resulted in methods such as QSAR, QSPR [55–61] which aim at the prediction of molecular properties. The basic paradigm of quantitative structure-property relationship (QSAR) is that compounds with similar structure have similar properties. This implies a smooth transient behavior in the relation between structure and property/activity, i.e., for any small change in the structure, the magnitude of the physico-chemical property or biological activity changes smoothly rather than in an abrupt, in all-or-none type, way. The molecular similarity measures are based on a large number of descriptors, i.e. of the numerical indices characterizing molecules. The basis for these studies is the development of various kinds of mathematical descriptors [62,63]. In the theory of molecular similarity it is commonly accepted that different descriptors and different similarity measures reveal different aspects of similarity. A pair of complex objects may be similar in one aspect and not similar in another aspect. Using different similarity measures, usually one obtains contradictory results which may be relevant in different contexts. The first QSAR studies on biological sequences using graphical representations of sequences have been already performed [64].

The present work describes the development of fundamental studies related to graphical representations of DNA sequences. First, the corrections to alignment methods are proposed in order to enrich the information related to different aspects of similarity. New similarity measures are created for the alignment distributions. Second, a critical review of graphical representations and their numerical characterization is given.

In the last chapter, new aspects of four-component spectral representation, graphical representation of DNA sequences, recently introduced by the author of this

work [65], are described. It is shown in the last chapter of this work that by using the four-component spectral representation one can recognize the difference in one base between a pair of sequences so it can be used for single nucleotide polymorfism (SNP) analyses which is subject of many investigation, as for example, in a recent work by Bhasi et al. [66]. Another important problem is to identify protein coding regions of genomic sequences [67,68]. First attempts of identifying protein coding genes using graphical representations of DNA sequences based on Z curve [69] or based on trinucleotides [70] have been already performed. It has been shown that the similarity relations are different for exons, and sequences with introns using the four-component spectral representation (see subsequent chapters). Such an observation suggests that the four-component spectral representation that reveals detailed aspects of similarity, as for example the comparisons of asymmetry of the gene structure, can be used to study this problem [71].

2 Corrections to the alignment methods

In this section I introduce corrections that reveal some aspects of similarity which cannot be identified in the standard alignment methods. The similarity space of complex objects is multidimensional. Only simple 1D objects can be classified in a unique way using a single similarity measure. Complex objects may be similar in one aspect and very different in another one. For example, in the case of atoms, if a similarity measure based on their atomic numbers is considered then the periodic table of elements is obtained. However, considering ionization energies as descriptors, the similarity relations between atoms change. A final, single similarity measure is a result of averaging over different aspects of similarity or, a consequence of neglecting most of the aspects of similarity.

The new similarity measures representing different aspects of similarity can be considered separately, or they can be combined in any way to search for their correlations with different biological functions.

2.1 Inadequacy of the alignment methods

The information about similarity of the sequences derived from the alignment methods is rather limited. For example, according to the standard alignment methods, in the following two different cases:

1. A T A T
   A G A G
2. A A T T
   A A G G

the similarity value is the same (50%). The non-zero contributions to the final result come from different positions in the sequences. In the first case, the A bases are spread over the whole sequence (positions 1, 3 give non-zero contributions). In the second case, the A bases are cumulated. In this sense, the alignment methods are degenerate:

different structures give the same result. Then, in the alignment methods these structures are undistinguishable. This is certainly one of the weakest points of the alignment methods. The degree of degeneracy may be very large and increases with the lengths of the sequences. Obviously, the degree of degeneracy in the model example is larger than 2. One can add more than two cases that give the same score as the two model cases. For example, the bases that give non-zero contribution can be also C, T, or G and the positions of the aligned bases can be different. Such details usually have biological consequences but they are not taken into account in the alignment methods.

In order to describe different aspects of similarity in more detail, let us define a discrete alignment distribution n_p for a pair of DNA sequences:

$$n_p = \begin{cases} 1, & \text{if the } p\text{-th positions in the two sequences are occupied by} \\ & \text{two identical bases,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let us introduce a variable x_p running along the sequence

$$x_p = pr, \tag{2}$$

where $p = 1, 2, \ldots K$ is the position in the sequence and r is the resolution that can be selected, depending on the length of the distribution, in a way convenient for the calculations changing the units of lengths.

K is the length of the sequences or subsequences for which the alignment is calculated. Two bases belonging to different sequences, both located on the p-th positions are represented by a pair of numbers, $\{x_p, n_p\}$.

Let us consider multiple alignments. Analogously, as for a pair of sequences according to the standard alignment methods, in the following two cases

	A	T	A	T
1.	A	G	A	G
	A	C	A	C
	A	A	T	T
2.	A	A	G	G
	A	A	C	C

the similarity value is the same (50%). Thus, the alignment method is highly degenerate (different bases on different positions give non-zero contributions and these situations are undistinguishable). As a consequence, additional similarity information should be added for a proper description of the objects. This information is necessary to remove the degeneracy, i.e. to distinguish between different cases. Analogously, as for a pair of sequences, we can define a discrete alignment distribution n_p of several (M) DNA sequences:

$$n_p = \begin{cases} 1, & \text{if the } p\text{-th positions in all } M \text{ sequences are occupied by} \\ & \text{the identical bases,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

M bases belonging to different sequences, located on the p-th positions are represented by $\{x_p, n_p\}$, where x_p is defined in Eq. 2.

2.2 Statistical properties of the discrete alignment distributions

As it is known from the statistics, distributions can be characterized in a convenient way by their moments. Distribution moments are the basic quantities in statistical spectroscopy. The aim of statistical spectroscopy [72–74], is to construct global characteristics of a spectrum. The individual eigenvalues, the experimental energy levels or the intensities of spectral lines are considered as statistical ensembles. Such an approach may be used in many areas of physics to study different kinds of problems. Let me just mention several applications of statistical spectroscopy.

Originally, methods of statistical spectroscopy were used in nuclear physics [75] where the character of the interparticle interactions is not exactly known. Assuming different forms of the Hamiltonian matrix and comparing distributions of the densities of the energy levels derived from this matrix and from the experiment some information about the Hamiltonian may be derived.

Statistical spectroscopy may also be used to study the locations of the individual eigenvalues of the Hamiltonian. In a way this is an inverse problem to the one from which the statistical spectroscopy originated: from global characteristics of the eigenvalues one tries to obtain some information about details of the spectrum. Examples of generating individual energy levels using methods of statistical spectroscopy may be found in the theory of nuclear, atomic, molecular, and solid-state spectra [76–78]. Approximating the eigenvalues by statistical quantities (spectral density distribution moments) is not limited by dimensions of the matrices and that is an advantage comparing to the standard methods based on diagonalization of the Hamiltonian matrix. In particular, we have studied statistical properties of spectra of the Heisenberg Hamiltonian [78]. The distribution of the eigenvalues have been found to be Gaussian-like, well approximated by several-term Gram-Charlier expansions [79]. The exact spectra (obtained by the diagonalizations of the Hamiltonian matrices) have been compared with the ones derived from the moment-generated spectral density distributions. This approximation gives a very good description of the spectrum in its central part however, as one should expect, deteriorates at the extremes. Relations between the exact and the moment-generated spectra are analyzed for several kinds of the lattices as a function of the number of moments. It has been observed that the quality of the statistical description improves with an increase of the dimension of the problem and with a lowering of the symmetry of the lattice.

Another attractive application of the statistical spectroscopy is a description of the shapes of molecular electronic bands [80–83]. Initially, the method of generating of envelopes of the intensities has been introduced for the transitions in crystals [84] and in atoms [85]. Replacing the calculations line by line by the statistical approach with much shorter computing time became also attractive in molecular physics. The shape of a molecular band may be defined as an envelope of the rovibrational lines which constitute the band. The method of determining the shapes of molecular electronic bands consists of several steps. First the expressions for the intensity distribution moments

for the considered system are derived. Then these expressions are used to calculate the moments corresponding to the solution of the pertinent quantum chemical model. Finally, a smooth function for which several lowest moments are equal to the exact ones, is derived. This function is an approximation to the envelope of the electronic band in a molecular spectrum. In particular, I have used this algorithm to derive the intensity spectrum corresponding to the transitions in H_2 molecule using 3-moment trial function [86]. In that paper I have also shown that the quality of the approximation depends on the choice of the trial function rather than on the number of moments taken into account. Adding moments of the order higher than 4 does not improve the results when the Gram–Charlier expansion is taken as the trial function. This process may even be divergent. In some cases a 4-moment Gram–Charlier expansion may give worse results than the 3-moment one. For example, this happens in a spectrum derived from a model based on the harmonic oscillator potential. In the case of H_2 molecule a non-standard 3-moment trial function has to be applied in order to get a high quality approximation of the spectrum (treated as a statistical distribution).

Distributions are commonly, and very conveniently, characterized by their moments. In the present work I describe DNA sequences as distributions and apply the distribution moments to study similarity between these sequences. A similarity measure $\tilde{\Delta}_q$ based on the q-th moment of the discrete alignment distribution n_p is defined as

$$\tilde{\Delta}_q = c \sum_{p=1}^{K} n_p {x_p}^q, \tag{4}$$

where $q = 0, 1, 2, \ldots$.

In this work $r = 1$. Therefore, the values of x_p are equal to p (Eq. 2).

The normalization constant c is defined so that the zeroth moment of the distribution is equal to one ($\tilde{\Delta}_0 = 1$):

$$c = \left(\sum_{p=1}^{K} n_p \right)^{-1}. \tag{5}$$

Comparing sequences, usually one is interested in the quantities that are independent of the lengths of the sequences. For that purpose, moments for which the mean value is equal to 0 ($\Delta_1 = 0$) and the variance is equal to 1 ($\Delta_2 = 1$) can be used as similarity measures:

$$\Delta_q = c \sum_{p=1}^{K} n_p \left[\frac{(x_p - \tilde{\Delta}_1)}{\sqrt{\tilde{\Delta}_2 - (\tilde{\Delta}_1)^2}} \right]^q. \tag{6}$$

Table 1 shows a model example of the alignment distribution n_p for a pair of sequences. The choice of the query sequence has no influence on the results. The length of sequence 1 is 12 and the length of sequence 2 is 15. If $K = 15$ is chosen then for $p > 12$ the distribution is defined as zeros: $n_{13} = n_{14} = n_{15} = 0$. Therefore, the

Table 1 Model example of alignment distributions ($r = 1$)

Seq. 1	A	T	G	A	C	T	T	T	G	C	T	G			
Seq. 2	A	T	G	G	T	G	C	A	C	C	T	G	A	C	T
$K = 15$															
x_p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_p	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0
$K = 12$															
x_p	1	2	3	4	5	6	7	8	9	10	11	12			
n_p	1	1	1	0	0	0	0	0	0	1	1	1			

Table 2 Model example of multiple alignment distribution ($K = 15, M = 3, r = 1$)

Seq. 1	A	T	G	G	T	G	C	A	C	T	T	G	A	C	T
Seq. 2	A	T	G	G	T	G	C	A	C	C	T	G			
Seq. 3	A	T	G	C	T	G	A	C	T	G	C	T			
x_p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_p	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0

contribution to moments for $p > 12$ is zero if $K = 15$. The moments are identical for both values of K. Consequently, the value of K in Eqs. 4–6 may be equal to the length of any of the sequences. Table 2 shows a model example of the multiple alignment distribution for $M = 3$ (Eq. 3). Analogously, the value of K may be set equal to the length of any of the three sequences. Thus, independently of the choice of K the moments remain the same.

2.3 Discrete alignment distribution moments as similarity measures

In this work, $\tilde{\Delta}_q$ and Δ_q are proposed as new similarity measures that can be treated as corrections to the alignment methods. They describe such features of similarity that cannot be identified in the alignment methods. In particular, the two model cases defined at the beginning of this chapter can be distinguished using the new measures.

The new numerical characterization of DNA sequences is exemplified using the β-globin gene of different species. The species and the locations of the sequences in genes as well as the lengths of the sequences, N_1 and N_2, are listed in Table 3. Tables 4, 5, 6, 7, 8, 9, 10, 11 show similarity matrices based on the new measures for the sequences listed in Table 3. Tables 4, 5, 6, 7 correspond to the coding sequences of the first exon, Exon 1^{CDS}, and Tables 8, 9, 10, 11 correspond to the second exon, Exon 2^{CDS}. The similarity matrices are based on different measures: $\tilde{\Delta}_1$ (Tables 4, 8), Δ_3 (Tables 5, 9), Δ_4 (Tables 6, 10), and Δ_5 (Tables 7, 11). The first moment, $\tilde{\Delta}_1$, depends on the length of the alignment distributions. In particular, if the compared sequences are identical then $\tilde{\Delta}_1 = (N+1)/2$ where N is the length of the sequence. This means that the mean of the distribution is located in the middle of the sequence, as expected. For example, if $N = K = 3, r = 1$, then the locations of particular bases along the sequence are

Table 3 Locations of Exon 1^{CDS} and of Exon 2^{CDS} in the β-globin gene and the corresponding lengths of the sequences for different species from the EMBL database

No.	Species	ID/Accession	Exon 1^{CDS}	N_1	Exon 2^{CDS}	N_2
1	Human	U01317	62187-62278	92	62409-62631	223
2	Goat	M15387	279-364	86	493-715	223
3	Opossum	J03643	467-558	92	672-894	223
4	Gallus	V00409	465-556	92	649-871	223
5	Lemur	M15734	154-245	92	376-598	223
6	Mouse	V00722	275-367	93	484-705	222
7	Rabbit	V00882	277-368	92	495-717	223
8	Rat	X06701	310-401	92	517-739	223
9	Gorilla	X61109	4538-4630	93	4761-4982	222
10	Bovine	X00376	278-363	86	492-714	223
11	Chimpanzee	X02345	4189-4293	105	4412-4633	222

Table 4 $\tilde{\Delta}_1$[Exon 1^{CDS}]

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	46.50	46.80	46.22	45.69	50.35	47.82	48.35	47.24	46.50	46.55	46.50
Goat		43.50	45.29	46.39	48.88	45.50	45.48	46.28	46.80	43.33	46.80
Oposs.			46.50	42.97	50.39	47.37	48.31	47.40	46.22	44.14	46.22
Gallus				46.50	49.14	45.75	47.94	46.22	45.69	47.36	45.69
Lemur					46.50	50.62	51.87	50.40	50.35	49.19	50.35
Mouse						47.00	48.69	46.27	48.39	45.30	48.39
Rabbit							46.50	48.49	48.35	45.28	48.35
Rat								46.50	47.24	46.06	47.24
Gorilla									47.00	46.55	47.00
Bovine										43.50	46.55
Chimp.											53.00

described by $x_1 = 1, x_2 = 2$, and $x_3 = 3$. In this case $\tilde{\Delta}_1 = 4/2 = 2$ and $x_2 = 2$ is the middle of the sequence.

One can normalize the similarity matrix based on $\tilde{\Delta}_1$ dividing all its elements by $K + 1$. Then one can easily see whether the mean value is larger or smaller than 1/2. If $\tilde{\Delta}_1$ is equal to 1/2 then the location of the mean value of the distribution is in the middle. If it is greater than 1/2 it is shifted towards the end of the distribution. Since Δ_q are independent of the lengths of the sequences it is convenient to keep at least one similarity measure ($\tilde{\Delta}_1$) that carries the information both about the lengths of the sequences and about the distributions of the aligned bases. Therefore, $\tilde{\Delta}_1$ is not normalized in this work.

An example of the similarity measure independent of the lengths of the sequences is Δ_3 that describes the asymmetry of the aligned distributions. For the symmet-

Table 5 $\Delta_3[\text{Exon } 1^{CDS}]$

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.00	−0.45	0.01	0.02	−0.15	−0.09	−0.10	−0.04	0.00	−0.43	0.00
Goat		0.00	−0.32	−0.38	−0.50	−0.34	−0.37	−0.35	−0.45	0.01	−0.45
Oposs.			0.00	0.21	−0.15	−0.07	−0.11	−0.07	0.01	−0.23	0.01
Gallus				0.00	−0.05	0.03	−0.12	0.01	0.02	−0.46	0.02
Lemur					0.00	−0.15	−0.23	−0.12	−0.15	−0.53	−0.15
Mouse						0.00	−0.11	0.01	−0.10	−0.32	−0.10
Rabbit							0.00	−0.09	−0.10	−0.35	−0.10
Rat								0.00	−0.04	−0.33	−0.04
Gorilla									0.00	−0.43	0.00
Bovine										0.00	−0.43
Chimp.											0.00

Table 6 $\Delta_4[\text{Exon } 1^{CDS}]$

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	1.80	1.94	1.62	1.77	1.84	1.71	1.85	1.78	1.80	1.98	1.80
Goat		1.80	1.85	1.94	2.28	1.98	1.84	1.85	1.94	1.78	1.94
Oposs.			1.80	1.75	1.75	1.61	1.70	1.61	1.62	1.89	1.62
Gallus				1.80	1.86	1.68	1.88	1.71	1.77	1.93	1.77
Lemur					1.80	1.80	2.00	1.83	1.84	2.26	1.84
Mouse						1.80	1.76	1.74	1.71	2.02	1.71
Rabbit							1.80	1.81	1.85	1.88	1.85
Rat								1.80	1.78	1.89	1.78
Gorilla									1.80	1.98	1.80
Bovine										1.80	1.98
Chimp.											1.80

ric distributions, in particular if identical sequences are compared, Δ_3 is equal to zero. It is negative for the left-skewed distributions and positive for right-skewed distributions. One can observe, that the asymmetry of the aligned distributions for Exon 1^{CDS} (Table 5) is different from the one for Exon 2^{CDS} (Table 9). The number of the negative values of Δ_3 is 41 for Exon 1^{CDS} and 21 for Exon 2^{CDS}. For example, in the case of human-mouse sequences, Δ_3 is negative for Exon 1^{CDS} and it is positive for Exon 2^{CDS}.

Another similarity measure independent of the lengths of the sequences is Δ_4. This is the kurtosis parameter, that is the measure of the peakedness of the distribution. Analogously as for the lower order moments, Δ_4 is different for different parts of a gene (Tables 6, 10). For example, the similarity measure based on Δ_4 for gallus-lemur sequences is 1.86 for Exon 1^{CDS} and 1.70 for Exon 2^{CDS}. The similarity relations based on Δ_4 between all the sequences are shown in Fig. 1. The horizontal axis

Table 7 Δ_5[Exon 1^{CDS}]

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.00	−1.66	0.05	0.14	−0.68	−0.35	−0.47	−0.18	0.00	−1.63	0.00
Goat		0.00	−1.12	−1.48	−2.38	−1.32	−1.25	−1.31	−1.66	0.04	−1.66
Oposs.			0.00	0.80	−0.64	−0.23	−0.43	−0.24	0.05	−0.83	0.05
Gallus				0.00	−0.32	0.12	−0.46	0.04	0.14	−1.73	0.14
Lemur					0.00	−0.70	−1.16	−0.61	−0.68	−2.45	−0.68
Mouse						0.00	−0.51	0.07	−0.37	−1.30	−0.37
Rabbit							0.00	−0.46	−0.47	−1.24	−0.47
Rat								0.00	−0.18	−1.29	−0.18
Gorilla									0.00	−1.63	0.00
Bovine										0.00	−1.63
Chimp.											0.00

Table 8 $\tilde{\Delta}_1$[Exon 2^{CDS}]

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	112.00	112.59	114.65	111.68	111.25	99.68	113.16	114.40	101.64	111.05	99.59
Goat		112.00	114.73	112.39	112.32	101.47	111.84	113.82	104.73	111.78	102.57
Oposs.			112.00	116.02	113.49	105.67	115.13	117.58	108.69	115.24	106.76
Gallus				112.00	109.45	104.92	112.63	113.92	107.34	110.51	105.56
Lemur					112.00	100.32	110.55	112.79	105.56	110.28	103.67
Mouse						111.50	102.44	105.71	114.20	102.15	114.78
Rabbit							112.00	115.10	105.16	110.46	103.12
Rat								112.00	107.21	113.96	105.34
Gorilla									111.50	101.55	111.00
Bovine										112.00	99.33
Chimp.											111.50

represents the values listed in Table 6, Δ_4[Exon 1^{CDS}], and the vertical axis represents the values listed in Table 10, Δ_4[Exon 2^{CDS}]. Each point in the figure corresponds to a given pair of species. The points are spread in the whole figure so similarity relations based on Δ_4 for Exon 1^{CDS} and for Exon 2^{CDS} are different from each other—they are not correlated.

All higher-order odd moments are also equal to zero for symmetric distributions. The behavior of Δ_5 is similar to Δ_3 for the same cases. The number of negative values of Δ_5 and the number of negative values of Δ_3 corresponding to Exon 1^{CDS} are the same and equal 41 (Tables 5, 7). In case of Exon 2^{CDS}, the number of negative values of Δ_3 is 19 and the number of negative values of Δ_5 is 21 (Tables 9, 11). This is clearly seen in Figs. 2 and 3(panels b), where the values listed in Table 5 versus values listed in Table 7 (Fig. 2, panel b) and the values listed in Table 9 versus values listed in Table 11 (Fig. 3, panel b) are shown. Analogously as in Fig. 1, a point in Figs. 2 and 3

Table 9 $\Delta_3[\text{Exon } 2^{CDS}]$

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.00	0.00	−0.04	0.01	0.00	0.28	−0.02	−0.06	0.30	0.04	0.30
Goat		0.00	−0.04	−0.04	−0.03	0.22	0.01	−0.07	0.23	0.00	0.21
Oposs.			0.00	−0.09	−0.04	0.11	−0.07	−0.14	0.13	−0.05	0.13
Gallus				0.00	0.01	0.21	−0.02	−0.04	0.23	0.03	0.23
Lemur					0.00	0.28	0.00	−0.05	0.21	0.03	0.19
Mouse						0.00	0.24	0.10	−0.09	0.21	−0.09
Rabbit							0.00	−0.09	0.26	0.05	0.26
Rat								0.00	0.16	−0.06	0.15
Gorilla									0.00	0.35	0.00
Bovine										0.00	0.34
Chimp.											0.00

Table 10 $\Delta_4[\text{Exon } 2^{CDS}]$

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	1.80	1.73	1.64	1.69	1.77	2.08	1.76	1.76	2.04	1.75	2.07
Goat		1.80	1.70	1.68	1.73	2.23	1.73	1.74	2.01	1.79	2.01
Oposs.			1.80	1.70	1.71	1.87	1.70	1.70	1.86	1.67	1.86
Gallus				1.80	1.70	2.04	1.66	1.69	1.94	1.67	1.95
Lemur					1.80	2.16	1.75	1.80	2.00	1.75	2.00
Mouse						1.80	2.13	2.14	1.74	2.18	1.74
Rabbit							1.80	1.76	2.05	1.74	2.07
Rat								1.80	1.92	1.77	1.93
Gorilla									1.80	2.19	1.80
Bovine										1.80	2.20
Chimp.											1.80

Table 11 $\Delta_5[\text{Exon } 2^{CDS}]$

	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	0.00	−0.05	−0.17	0.05	−0.02	1.50	−0.08	−0.25	1.56	0.11	1.62
Goat		0.00	−0.19	−0.11	−0.14	1.38	0.01	−0.25	1.23	0.02	1.21
Oposs.			0.00	−0.32	−0.16	0.71	−0.28	−0.53	0.77	−0.23	0.76
Gallus				0.00	0.09	1.14	−0.05	−0.17	1.12	0.12	1.16
Lemur					0.00	1.53	0.03	−0.21	1.10	0.09	1.06
Mouse						0.00	1.41	0.81	−0.34	1.24	−0.35
Rabbit							0.00	−0.35	1.39	0.15	1.45
Rat								0.00	0.89	−0.25	0.90
Gorilla									0.00	1.89	0.00
Bovine										0.00	1.90
Chimp.											0.00

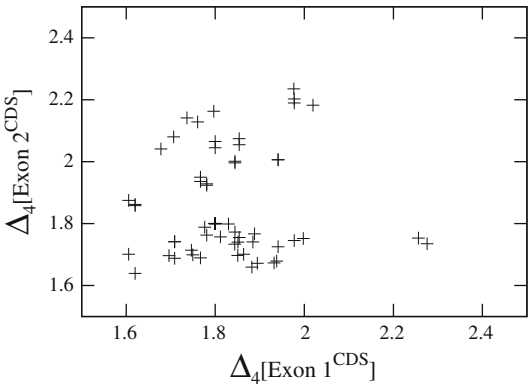

Fig. 1 Δ_4 [Exon 1^{CDS}] – Δ_4[Exon 2^{CDS}] diagram for the sequences listed in Table 3

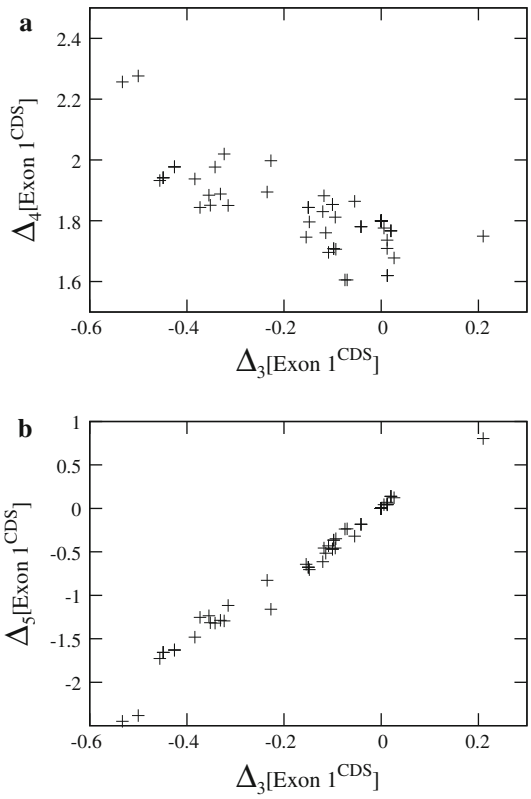

Fig. 2 Δ_3 [Exon 1^{CDS}] – Δ_q[Exon 1^{CDS}] diagrams for the sequences listed in Table 3 ($q = 4$ panel **a**, $q = 5$ panel **b**)

corresponds to a given pair of species. However, in Figs. 2 and 3 the same parts of a gene are represented in both vertical and horizontal axes: Exon 1^{CDS} in Fig. 2 and Exon 2^{CDS} in Fig. 3, while in Fig. 1 the vertical axis corresponds to Exon 2^{CDS} and the horizontal one to Exon 1^{CDS}. The relations between Δ_3 and Δ_5 are approximately linear both for Exon 1^{CDS} (Fig. 2, panel b) and for Exon 2^{CDS} (Fig. 3, panel b). This

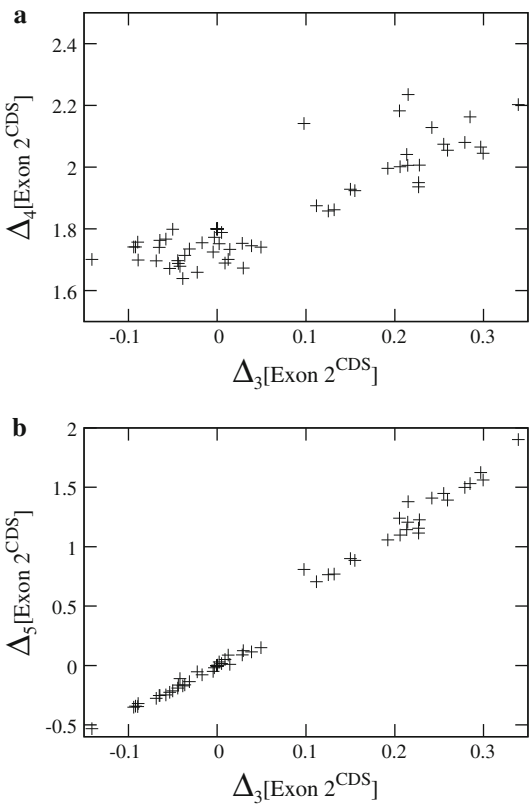

Fig. 3 Δ_3 [Exon 2^{CDS}] − Δ_q[Exon 2^{CDS}] diagrams for the sequences listed in Table 3 ($q = 4$ panel **a**, $q = 5$ panel **b**)

Table 12 Similarity measures of multiple alignment distributions ($M = 11$)

	$\tilde{\Delta}_1$	Δ_3	Δ_4	Δ_5
Exon 1^{CDS}	50.125	−0.744	2.262	−2.859
Exon 2^{CDS}	106.379	0.185	1.688	0.831

means that the information coming from Δ_5 is similar to the one coming from Δ_3. Therefore, the corrections Δ_5 can be neglected. The information coming from Δ_4 is different than the one coming from Δ_3 which is seen in Figs. 2 and 3, panels a, where $\Delta_3 - \Delta_4$ diagrams are shown for the same parts of the gene: Exon 1^{CDS}, Fig. 2 panel a, and Exon 2^{CDS}, Fig. 3 panel a. As we see, the corrections up to Δ_4 are sufficient to introduce the essential similarity information.

New similarity measures for multiple alignment are shown in Table 12 ($M = 11$, the sequences listed in Table 3). The new measures are different for Exon 1^{CDS} and Exon 2^{CDS}. For example Δ_3[Exon 1^{CDS}] is negative and Δ_3[Exon 2^{CDS}] is positive.

Table 13 Model example of four-component multiple alignment distribution ($K = 15, M = 3, r = 1$)

Seq. 1	A	T	G	G	T	G	C	A	C	T	T	G	A	C	T
Seq. 2	A	T	G	G	T	G	C	A	C	C	T	G			
Seq. 3	A	T	G	C	T	G	A	C	T	G	C	T			
x_p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_p^A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n_p^C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n_p^T	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
n_p^G	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0

Table 14 Model example of optimized multiple alignment distribution ($K = 12, M = 3, r = 1$)

Seq. 1	A	T	G	G	T	G	C	A	C	T	T	G
Seq. 2	A	T	G	G	T	G	C	A	C	C	T	G
Seq. 3	A	T	G	C	T	G	A	C	T	–	–	G
x_p	1	2	3	4	5	6	7	8	9	10	11	12
n_p	1	1	1	0	1	1	0	0	0	0	0	1

In order to further enrich the information that can be derived from the alignment methods, one can introduce the four-component alignment distributions, separately for A, C, T, and G bases. A specific γ-component of this distribution, referred to as γ-distribution, is defined as

$$n_p^\gamma = \begin{cases} 1, & \text{if the } p\text{-th positions in all } M \text{ sequences are occupied by} \\ & \text{the base } \gamma, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\gamma = A, C, T, G$ denotes one of the bases. Now, for each of the γ-distributions one can calculate the corrections (the appropriate moments). Such kind of distributions can be created for a pair of sequences ($M = 2$) and also for the multiple alignment studies ($M \geq 3$). A model example of the four-component distribution with $M = 3$ is shown in Table 13.

The same definitions of the distributions (Eqs. 1, 3, 7) can be also used after the maximally scoring alignment of the sequences has been found. A model example of the optimized multiple alignment distribution based on Eq. 3 for $M = 3$ is shown in Table 14. The maximally scoring alignment is obtained if two gaps are introduced in sequence 3.

Obviously, the information is more detailed if four-component optimized distribution is created. Table 15 shows such a distribution (Eq. 7) for the same sequences as shown in Table 14. The application of the new similarity measures to all kinds of the distributions is simple and straightforward. In this way, different aspects of similarity can be revealed.

Table 15 Model example of optimized four-component multiple alignment distribution ($K = 15$, $M = 3$, $r = 1$)

Seq. 1	A	T	G	G	T	G	C	A	C	T	T	G
Seq. 2	A	T	G	G	T	G	C	A	C	C	T	G
Seq. 3	A	T	G	C	T	G	A	C	T	–	–	G
x_p	1	2	3	4	5	6	7	8	9	10	11	12
n_p^A	1	0	0	0	0	0	0	0	0	0	0	0
n_p^C	0	0	0	0	0	0	0	0	0	0	0	0
n_p^T	0	1	0	0	1	0	0	0	0	0	0	0
n_p^G	0	0	1	0	0	1	0	0	0	0	0	1

3 Graphical representations of DNA sequences

An attractive, alternative to the time consuming alignment methods, are graphical representations. They reveal different aspects of similarity, offer both numerical characterization of similarity and the visualization. Also the computing effort is in this case very small. In this section graphical methods are discussed.

In the original approaches, DNA sequences were plotted as either three-dimensional [87] or two-dimensional [88–90] curves. The shapes of the curves were determined by a walk in a space spanned by four vectors that represent the four bases. In the first article on this subject, Hamori proposed a graphical representation method in which the information about the DNA sequence has been mapped into a three-dimensional-space curve. A unit vector of a characteristic direction has been assigned to each of the four nucleotides: adenine A, cytosine C, thymine T, and guanine G. In this approach the shape of the curve (called H-curve) representing the sequence of nucleotides is obtained by joining the vectors in the order of the nucleotides in the sequence. Changing the resolution one can see short-range details or global trends of the distribution of nucleotides. For example, H-curve is shifted in characteristic direction if the sequence is rich in certain nucleotides. It is also easy to recognize the locations of the repeating elements in the sequence. The first mathematical representation Hamori also published in Nature in 1985 under the title "Novel DNA sequence representation" [91]. The same year another article about a new graphical representation titled "Simpler DNA sequence representations" has been also published in Nature by Gates [88]. In this approach, guanine is represented by a unit vector in the positive x-axis direction, complementary cytosine is represented by a negative x-axis unit vector, and adenine and thymine are represented by unit vectors in the positive and negative y-axis directions, respectively. Using such an approach all sequences can be represented in two dimensions in a unique manner, while using the Hamori approach, DNA structure may be viewed from any chosen perspective in two-dimensional plots. Obviously, a chosen perspective of a 3D curve in 2D space gives only a part of the total information about the sequence. However, also in the graphical representation proposed by Gates some information may be lost, as it is shown in a subsequent part of this work.

About 10 years later, Nandy (independently of Gates) published an article "A new graphical representation and analysis of DNA sequence structure: I. Methodology

and application to globin genes" [89]. The idea is very similar to the one presented by Gates. In the scientific correspondence [92], the author explains that he has just brought to his attention that a similar technique was presented by Gates and indicates some advantages of his method: The nontrivial choice of the coordinate system A-G, C-T (purine-pirymidine) instead of the axis system proposed by Gates (C-G, A-T) may give more significant biological information.

One year later (independently of Nandy), Leong and Morgenthaler proposed two new graphical representations of DNA sequences [90]. The first one is a slight modification of the Gates method: they change the unit vectors corresponding to the particular bases. The x-axis represents C and A and y-axis G and T. According to the authors such a change allows to exhibit the distribution of purines (A and G) and pyrimidines (C and T). The authors noticed that some information may be lost if a walk moves several times over the same ground. However, the authors found a good solution to identify in the plot the regions in which the parts of the sequences are hidden: the scale is visible even for long sequences and the numbers that label the bases in the plot are pointed every one hundred. The authors have not proposed any numerical characteristics and a way of indication in the plot of the hidden parts (by labeling or coloring) seems to be a good solution. Leong and Morgenthaler also proposed another, interesting graphical representation: *gap plots* that give the information about the distances between particular bases.

Independently of the graphical representations introduced by Gates [88], Nandy [89], Leong and Morgenthaler [90], similar graphs, also based on vectorial representations of the four bases and constructing 2D DNA walks, have been constructed by Mizraji and Ninio [93] and by Lobry [94]. Lobry also used orthogonal directions but his choice of the unit vectors representing the four bases was different than the ones used in refs. [88–90]. Surprisingly, these two important contributions remained rather unnoticed. The specific choice of the basis vectors done by Mizraji and Ninio seems to solve many problems. The vectors have been chosen so that it is easy to distinguish between coding and noncoding parts of the sequence and the graphs are nondegenerate. Mizraji and Ninio also proposed a graphical representation which shows purine/pirymidyne distributions along the sequence. However, the authors did not propose any numerical representation associated with these graphs.

As a consequence, four similar 2D graphical representations have been created. They differ from each other in the choice of the coordinate systems: x-axis: G-C (Gates), A-G (Nandy), C-A (Leong and Morgenthaler), A-T (Lobry). The most popular became the graphical representation proposed by Nandy, called *Nandy plots*.

However, such a two-dimensional representation may lead to some parts of the sequence being hidden if the walk is performed back and forth along the same trace (so called repetitive walks). Labeling and coloring only approximately localizes the regions in the sequences where the hidden parts are located. A 2D walk does not retain the history of the graph. This is not a linear method: A particular part of the graph may come from different parts of the sequences. The advantage is a small size of the graph representing long sequences and very often the information coming from such a plot may be sufficient. In order to eliminate, or to minimize, the degeneracy caused by the repetitive walks, many different methods have been introduced. For example, Guo et al. [95] introduced a new graphical representation, also based on a walk in 2D

space changing the angles between the basis vectors: the four nucleic acid bases are represented by the vectors: A by $\left(-1, \frac{1}{d}\right)$; T by $\left(\frac{1}{d}, -1\right)$; G by $\left(1, \frac{1}{d}\right)$; and C by $\left(\frac{1}{d}, 1\right)$, where d is a positive integer. The authors have shown that the degree of degeneracy of the new graphs is lower than for Nandy plots, but it is still present and depends on the value of d.

A further modification of the graphical representation based on a walk in 2D space has been introduced by changing the vectors in such a way that the basis vectors corresponding to pyrimidines (T, C) are located in the first quadrant of the Cartesian coordinate system and to purines (A, G) in the fourth one [38,39]. The unit vectors representing four nucleotides are as follows: $(m, -\sqrt{n})$ for A; $(\sqrt{n}, -m)$ for G; $(\sqrt{n}, m)$ for C; and $(m, \sqrt{n})$ for T, where m is a real number, n is a positive real number and $m \neq \sqrt{n}$. The authors have proved that their method is nondegenerate.

Other examples of modifications of the vectors representing the bases, are new graphical representations proposed by Liu et al. in which the authors introduced polar coordinate system [96] and also H-L curve representation proposed recently by Huang et al. [97,98].

An interesting *2D ladder-like* graphical representation has been also proposed by Li and Wang [99]. Their graphs are based on the division of bases according to their chemical properties. The four bases can be classified into groups: 1. purine R = A, G, pirymidyne Y = C, T; 2. strong H bond S = C, G, weak H bond W = A, T; and 3. amino M = A, C, keto K = G, T. The method is also based on a walk in 2D space with basis vectors (0,1), (1,0) for characteristic sequences (M, K), (R, Y) and (W, S).

Recently, we have proposed another method aimed at some improvement of the original 2D walk method (Nandy plots) [100]. We have called this representation *2D-dynamic graph* because its numerical representation, i.e. the set of descriptors, is analogous to the one used in the dynamics (see subsequent chapter). This method is based on Nandy plots but it removes the degeneracy coming from the repetitive walks. The DNA sequence is represented as a set of material points in 2D space. The distribution of the points in the plane and the way of calculating their masses are shown in the model examples (the left panels of Figs. 4, 5). The method of plotting the graph representing the DNA sequence is based on the 2D walk with the basis vectors identical to the ones of the Nandy plots. The new element is attaching a point mass to the end of each vector. The mass of the point depends on the number of crossings of the graph in this point. The graph still crosses itself but the numbers of crossings are clearly revealed and taken into account in the numerical representation. We start the

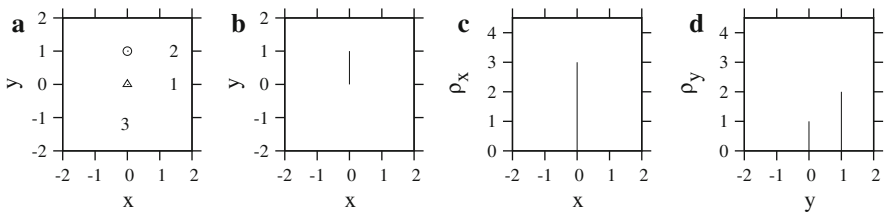

Fig. 4 2D-dynamic graph for a model DNA sequence CTC (panel **a**), the corresponding Nandy plot (panel **b**), and mass-density distributions (ρ_x panel **c**, ρ_y panel **d**). *Circles* correspond to $m = 2$ and *triangles* to $m = 1$. The projected masses are also denoted in panel **a**

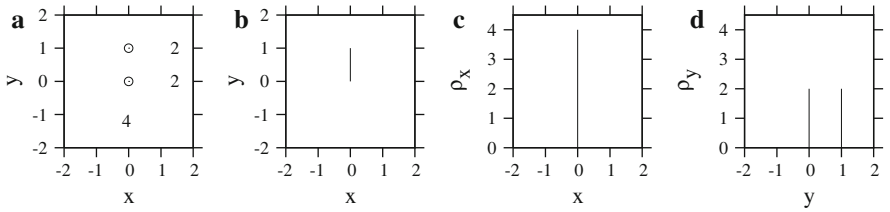

Fig. 5 2D-dynamic graph for a model DNA sequence CTCT (panel **a**), the corresponding Nandy plot (panel **b**), and mass-density distributions (ρ_x panel **c**, ρ_y panel **d**). *Circles* correspond to $m = 2$. The projected masses are also denoted in panel **a**

graph at the point with the coordinates (0,0). Then the shifts are made by unit vectors different for each base: A = (−1,0), G = (1,0), C = (0,1), T(0,−1). Figure 4, panel a, shows the method of plotting the 2D-dynamic graph for a model sequence CTC and Fig. 5, panel a, for CTCT one. The first base in the sequence is C and we make a shift along the vertical axis in the positive direction. At the end of this vector (position (0,1)) we locate the point with the mass equal to 1. The second base in the sequence is T and we make the second shift along the vertical axis in the negative direction starting from the end of the last vector. At the end of the second vector again we locate the point with the mass also 1 (position (0,0)) and so on. If the ends of vectors meet several times at the same point then the mass of this point increases (it is equal to the sum of all masses located in this point). The total mass in the graph is equal to the total number of bases in the sequence (3 in Fig. 4 and 4 in Fig. 5). Different masses are represented by different symbols in the plots. Please note that both sequences CTC and CTCT are represented by the identical Nandy plots (Figs. 4, 5, panels b) since the last shift in Fig. 5 is made along the same trace as the previous one. 2D-dynamic graph removes this degeneracy (the masses of the points (0,0) are different: 1 for CTC and 2 for CTCT). The difference between the two sequences is also revealed in the mass-density distributions which we create for x and y directions [101]. The masses are projected onto two orthogonal directions and then summed for each x and y. In the model examples the results of the projection and of the summation of the masses are shown in Figs. 1 and 2 (panels a). For example, in the x direction, they are 3 and 4 in Figs. 4 and 5, respectively. The mass-density distributions are composed of single lines located at the coordinates corresponding to the projected masses ($x = 0$ for ρ_x and $y = 0$, $y = 1$ for ρ_y). The intensities of the lines correspond to the projected masses. The center panels in Figs. 4 and 5 correspond to mass-density distributions for x direction (ρ_x) and the right ones for y directions (ρ_y). These distributions create another way of visualization of the 2D-dynamic graphs. However, the main reason for the creation of the mass-density distributions is deriving new descriptors related to 2D-dynamic graphs (see the subsequent chapter).

The modifications of the original 2D walk methods resulted also in graphs which became linear-like representations (1D), extending along one direction in 2D space. In such kind of methods only the horizontal axis is associated with the positions of the bases. Therefore these methods are free of the effects of self-overlapping of the graphs. The cost we have to pay for the reduction of the degeneracy, is worse visualization of long sequences.

A combination of a linear-like method with a DNA walk has recently been proposed by Zhang [102]. The author has chosen basis vectors in such a way that the walk is performed along a horizontal axis. One nucleotide is represented by a pair of basis vectors instead of a single vector: (1, 1),(1, 1) corresponds to A, (1, 1), (1, −1) corresponds to T, (1, −1), (1, 1) corresponds to C, and (1, −1), (1, −1) corresponds to G. Since one base is represented by a double vector, the author calls his graphical representation of DNA sequences a *DV-curve*. A recently introduced graphical representation of DNA sequences based on the neighboring dual niclueotides (dinucleotides) [103,104] is another example of a linear representation. The authors plot a dinucleotide (DN) curve representing the distributions of pairs of nucleotides along the sequence.

Several years ago a *four-horizontal-line* graphical representation has been proposed by Randić et al. [105,106]. Instead of considering the four directions along the Cartesian coordinate axes, they draw four horizontal lines separated by unit distances. Each line is associated with one base: A, T, G, and C, from the top. The sequence is written at the bottom of the lowest line, with unit distances between the neighboring bases. The dots (or rectangles) are put on the lines if a particular base appears in the sequence. This graphical representation resembles medieval musical scripts having staff of four lines [107]. For a better visualization the adjacent points are connected by a line, and zigzag-like curve is obtained. The idea proposed by Randić of the visualization of DNA sequence by zigzag curves has been extended by different combinations of labeling the lines and by different number of graphs representing one sequence (characteristic graphs). Usually, the horizontal lines are not plotted.

Another linear graphical representation has been proposed by Li and Wang [108]. The graphical representation is composed of three characteristic graphs, each of them consisting of two horizontal lines. Each line in each graph is assigned to more than one base. Then, the sequence is represented by more than one characteristic graph. The lines in particular graphs are labeled by the following bases: graph 1: M (top line), K (bottom line), graph 2: R (top line), Y (bottom line), graph 3: W (top line), S (bottom line). This means that in graph 1, a dot is put in the top line if the base in the sequence is M (i.e. A or C), and a dot is put in the bottom line for G and T bases. Analogously to Randić, the adjacent dots are connected by a line and again a zigzag curve is obtained.

A similar graphical representation has been proposed by Song and Tang [109]. In this approach, the three classifications are applied to construct six characteristic graphs representing one sequence. Two graphs correspond to one classification. For example in two graphs corresponding to classification purine (R) - pirymidyne (Y), the middle lines correspond to purines and pirymidynes in the first and in the second graph respectively. The other two lines correspond to these bases that are not purines or not pirymidynes, respectively, i.e.: A, Y, G label the lines in the first graph (top, middle, and bottom respectively) and C, R, T label the lines in the second graph.

Another example of an analogous graphical representation has been proposed by Liao and Wang [110]. The sequence is represented by three graphs, and each of them is consisting of two horizontal lines. The lines are labeled as follows: Graph 1: AG top line, CT bottom line; graph 2: AC top line, TG bottom line; graph 3: AT top line, GC bottom line. This means that the dots are put in the top line in graph 1 if the base is A or G, otherwise the dot is put in the bottom line. Three zigzag curves constitute the graphical representation.

In another analogous graphical representation, proposed by Wang and Zhang [111], also consisting of three characteristic graphs, the lines are labeled by the following bases: graph 1: non-A = G, C, T and A, graph 2: non-G = A, C, T and G, graph 3: non-C = A, G, T and C.

A slight modification of this method has been proposed by Yao and Wang [112]. The authors proposed to use *cells* instead of horizontal lines. They considered different shapes of cells, for example a rectangle. Each corner of the rectangle is assigned to a particular base. The cells are placed next to each other. Particular bases in the sequence are put in a proper corner (each base is located in its own cell). The adjacent dots are connected by a line and a zigzag curve representing the sequence is obtained.

The methods described above, based on several horizontal lines, can be also considered as spectral-like representations (lines with some intensities appear in the positions corresponding to the bases in the sequences). This point of view has been expressed in a recent article by Randić [113]. The author presents four-horizontal-line graphs and chaos-game 2D maps [114,115] in the form of spectrum-like graphical representations.

Recently, I have introduced another spectral-like graphical representation called four-component spectral representation [65]. The method is very sensitive. Within this model, differences in only one base can be detected. By using linear graphical representations of DNA sequences the problem of degeneracy can be overcome. However, in technical terms, the visualization of long sequences is rather inconvenient. A good solution for this drawback is introducing a resolution parameter for linear representations as it was done for the four-component spectral representations (for details see Sect. 4).

Another solution is to combine the compact form of the plots characteristic for 2D walks and zigzag curve method, as proposed by Randić et al. [116,117]. In the last approach, the sequence is represented by a zigzag spiral, known in the literature as the *worm curve*. The worm curve represents a path of a robot [116]. It does not intersect itself and uses a little space for the graphical representations of long sequences. Another compact graphical representation, *Four-color map*, has also been proposed by Randić et al. [118]. The map is constructed as a spiral of square cells. The first base is located at the central square of the spiral, and the last base finishes the spiral. Then four different colors are assigned to particular squares representing different bases: red for G, blue for T, green for C, and yellow for A.

The original 3D method proposed by Hamori has been also extended by various authors. In particular, a modified Hamori curve representation of DNA sequences has been recently introduced by Pesek and Zerovnik [119].

Moreover, methods based on a walk in 3D space with different vectors corresponding to particular bases were introduced: vectors located along tetrahedral directions A(1,−1,−1), G(−1,1,−1), C(−1,−1,1), T(1,1,1) [120] or AGC-T curve, where the vectors are chosen as A(1,0,0), G(0,1,0), C(0,0,1), T(1,1,1) [121,122]. Examples of other 3D graphs are representations of one sequence by a set of characteristic 3D curves [123–128]. Another 3D graphical representation, called *Z curve*, combines the properties of several characteristic curves [129]. A single Z curve contains the information about the distributions of purine/pirymidyne, amino/keto and strong H bond/weak H bond.

Recently, new 3D graphical representations based on the frequencies of occurring of pairs of nucleotides (dual nucleotides or dinucleotides) or trinucleotides in DNA sequences have been created. Four nucleotides form 16 dinucleotides and 64 trinucleotides. By assigning different vectors to each pair or to each trinucleotide in 3D space, 3D-curves are obtained. The curves contain the information about neighboring bases and their distributions along the sequence. Dual nucleotides can be also divided into groups according to their chemical properties, as for example purine dinucleotides (AG, GA), pirymidyne dinucleotides (CT, TC), amino ones (AC, CA), keto ones (TG, GT), weak H-bond (AT, TA) and strong H-bond (CG, GC). 3D graphical representation of one sequence by four characteristic curves based on dinucletides has been proposed by Cao et al. [130]. Other 3D graphical representations based on dinucleotides (*PN-curves*) [131], (*DN-curves*) [132], (*D-curves*) [133], or based on trinucleotides (*TN-curves*) [134] have been also proposed.

4 Numerical representations of DNA sequences

Graphical representations constitute a tool allowing visual inspection of the sequences. Moreover, each graph can be characterized by the quantities called in the theory of molecular similarity, *descriptors*. The descriptors representing numerically some properties of the sequences can be used for similarity/dissimilarity analysis of the sequences. The computing time of the calculations of the descriptors is low and the numerical comparison of long sequences becomes attractive. The algorithm of the computation of the descriptors is independent of the visualization tool. Therefore, the graphical representations can be recognized as both numerical and graphical tools separately. However, each descriptor represents some specific properties of the graphs and it is not obvious how to characterize graphical objects by numerical values (for review of methods related to the creation of mathematical descriptors of DNA sequences up to 2006 see [135]).

One of the methods, most commonly used to describe graphs numerically, is transforming the plots to matrices. The method has been initially introduced by Randić et al. for 3D graphical representations [120]. The authors introduced distance matrices, D/D. The numerator in the matrix element (i, j) stands for the Euclidean distance between vertices i and j, and the denominator stands for the graph theoretical distance (the number of arcs separating the two vertices). The authors proposed the leading eigenvalues of the matrices as the descriptors. The normalized leading eigenvalue of a D/D matrix offers a measure of the degree of folding of a chain-like structure or a curve. The authors introduced also higher-order matrix ${}^{k}\mathrm{D}/{}^{k}\mathrm{D}$ that is constructed by taking matrix elements of D/D matrix to power k. In the limit $k \to \infty$, the resulting matrix reduces to a binary matrix ${}^{\infty}\mathrm{D}/{}^{\infty}\mathrm{D}$. As the descriptors the authors also proposed the leading eigenvalues of these matrices. Such kind of descriptors can be viewed as an index of flexibility (or stiffness) of the structure.

The methods of transforming graphs to matrices stimulated introducing new kinds of matrices. Different kinds of matrices associated with the graphs have been introduced by Song and Tang [109]. The authors introduced the Euclidean matrix E, whose (i, j) element is defined as the Euclidean distance between vertices (dots) i and

j of the curve. They also introduced M/M matrix whose elements are defined as a quotient of the Euclidean distance between two vertices of the curve and the number of arcs between the two vertices. The third kind of matrix introduced by these authors is L/L matrix whose elements are defined as a quotient of the Euclidean distance between two vertices of the curve and the sum of geometrical lengths of arcs between the two vertices. As the descriptors the authors chose the leading eigenvalues of M/M and L/L matrices. The authors considered characteristic linear curves and their descriptors characterize the distribution of bases with different chemical structures. The authors also considered higher-order L/L matrices. New kind of matrices has been also proposed by Liao et al. [38]. The authors introduced covariance matrices associated with the graphs.

Usually, the leading eigenvalues of the matrices are taken as descriptors. A discussion of the properties of such kind of descriptors may be found in a recent article by Yuan et al. [136]. Some authors propose to consider more eigenvalues or matrix elements as descriptors of the sequences. Wang and Zhang proposed to take as a descriptor the sum of the maximal and minimal eigenvalues for the matrices associated with their graphical representation, called *three non-base* representation [111]. The authors suggested that the information reflected only by the leading eigenvalue might not be comprehensive enough. Liao et al. [38] took all (two) eigenvalues of the 2×2 covariance matrices. Li and Wang proposed as descriptors normalized matrix norms instead of the eigenvalues [99]. Randić et al. considered as the descriptors average matrix elements of the matrices associated with the four-color map representation of DNA sequences [118]. Liao and Wang proposed as descriptors the average bandwidths [125]. They can be obtained by summing the distance matrix elements along each of the lines parallel to the main diagonal if the matrix is in the canonical form. Qi and Fan took all elements of the matrix as descriptors of the sequences of equal lengths [131]. Pesek and Zerovnik proposed to take as the numerical characterization of the modified Hamori curve a product of first ten and last ten eigenvalues of the descending ordered eigenvalue list of the matrix L/L [119].

Numerical representation of 2D or 3D graphical representations of DNA sequences based on transforming the graphs into matrices and deriving the descriptors from these matrices has been widely used by many authors. These descriptors characterizing a sequence can be used as components of similarity measures between a pair of sequences. Examples of similarity analysis of DNA sequences using this method may be found in [137,105,106,108,116,123,110,112,125,124,109,118,138,126,111].

Numerical representation of a graphical representation can be also performed directly from the coordinates or from the properties of the graphs without transforming the graphs to matrices. Gates plotted each sequence as a graph of the cumulative Manhattan distance (from the origin) against the sequence position [139]. Manhattan or city-block distance considered by Gates is calculated as the arc length between points. For sequences of equal lengths it is convenient to plot the differences of the graphs. As descriptors of the sequences he proposed the means of the Manhattan and Euclidean "fractal" dimensions.

Raychaudhury and Nandy proposed mean x and y coordinate values, and the radius of the graph as descriptors of DNA sequences [140]. Guo and Nandy introduced also

improved mean x and y coordinate values, and the radius of the graph, reducing the degeneracy of the previously defined descriptors of DNA sequences [141]. Yao et al. extended these descriptors to three dimensions defining 3D radius and adding mean z-coordinate as a descriptor [122].

We have extended the set of these 2D descriptors to higher-order moments of the mass-density distributions. The mean x and y coordinate values are equal to the first-order moments ($M_{x,1}$, $M_{y,1}$) of the mass-density distribution, ρ_x and ρ_y respectively. In particular, if in a 2D-dynamic graph we put all masses equal to 1, then the 2D-dynamic graph becomes the Nandy plot and all the moments of the two graphs are identical. Introducing the masses different than 1, the mean x and y coordinate values become the coordinates of the center of mass of the graph and are different than for the Nandy plot. As the new descriptors we proposed moments of the mass-density distributions ρ_x and ρ_y up to the sixth order [101] and up to the eighth order [142]. Higher-order moments give more specific information about the distribution of masses. For example, second-order moments ($M'_{x,2}$, $M'_{y,2}$) give the information about the width of ρ_x and ρ_y. We have shown that the third- ($M''_{x,3}$), fourth- ($M''_{x,4}$), fifth- ($M''_{x,5}$), and sixth-order ($M''_{x,6}$) x-moments of the mass-density distributions representing histone H4 coding sequences have different values for plants than for vertebrates [101]. In the present work, 2D-plots $M''_{x,q} - M''_{x,q'}$ are proposed instead of 1D-plots (descriptors versus labels of the sequences) that were shown in [101]. 2D-plots are shown in Fig. 6. A point in the figure corresponds to a single sequence while a point in Figs. 1, 2, 3 represents a similarity measure between a pair of sequences. The figure consists of 4 plots: part a $M''_{x,4} - M''_{x,5}$, part b $M''_{x,5} - M''_{x,7}$, part c $M''_{x,4} - M''_{x,6}$, and d one $M''_{x,5} - M''_{x,6}$. In all the plots we observe clusterization of evolutionary similar organisms: plants are located in different parts of the plots than the vertebrates.

The differences between histone H4 coding sequences across the species are not big and it is rather difficult to find the descriptors that reveal the clusterization. Please note that y-moments and also x-moments for the order smaller than 4 do not lead to clusterization in this case. In particular, this means that using the Nandy plots for which the descriptors are taken as the mean values (first-order moments) of x and y we cannot get the clusterization. I have also found another set of descriptors (related to the four-component spectral representation) that reveal clusterization for histone H4 and H1 coding sequences (for details see the subsequent chapter).

Analogous (2D visualization) is introduced in the present work for the recently proposed molecular descriptors. Figure 7 shows moment-based classification of the molecules: $M_1 - M'_2$ (top), $M''_3 - M''_4$ (middle), and $M''_5 - M''_6$ (bottom). We have shown that the new molecular descriptors (moments of the intensity distributions) have different values for two kinds of molecules: nitriles and amides. In our recent paper, 1D plots have been presented (descriptors versus labels of the molecules) [143]. Figure 7 shows 2D plots. We observe that the descriptors representing nitriles are located in different parts of the plots than those representing amides. Figures 6 and 7 represent different objects: DNA sequences and molecules, respectively. However, the idea is the same. The clusterization of the descriptors indicates that these descriptors can be a good tool for similarity/dissimilarity analysis. The descriptors cluster (have similar values) for similar objects so they exhibit some properties of the considered objects.

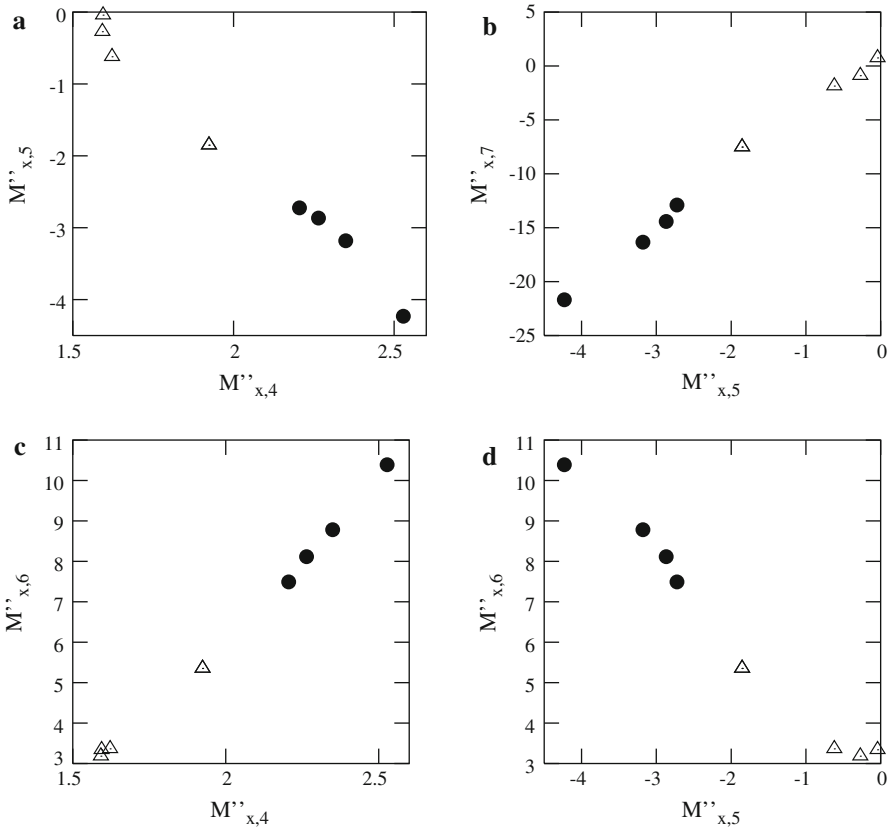

Fig. 6 $M''_{x,q} - M''_{x,q'}$ classification of histone H4 coding sequences. *Dots* correspond to plants and *triangles* to vertebrates

Moreover, some of the plots reveal similar shapes, as for example, middle and bottom parts of Fig. 7. This may suggest correlations between some of the descriptors. However, the shape is similar but not identical. The problem of correlation and extracting the minimal set of moments we studied in ref. [144]. We concluded that a universal set of independent moments does not exist. Usually 4 lowest moments are sufficient to describe the object but also the information coming from higher-order moments cannot be neglected in some cases.

As the new descriptors of DNA sequences we also proposed the angles between the x axis and the principal axis of inertia of the 2D-dynamic graph (axes for which the tensor of moment of inertia is diagonal) [100]. We also introduced the principal moments of inertia as the descriptors of DNA sequences associated with the 2D-dynamic graph [100]. They are associated with the rotations about the principal axes. The moment of inertia of an object about a given axis describes how difficult is to induce an angular rotation of the object about this axis. If the mass is concentrated close to the axis of rotation, it is easier to accelerate into spinning fast and the moment of inertia is smaller.

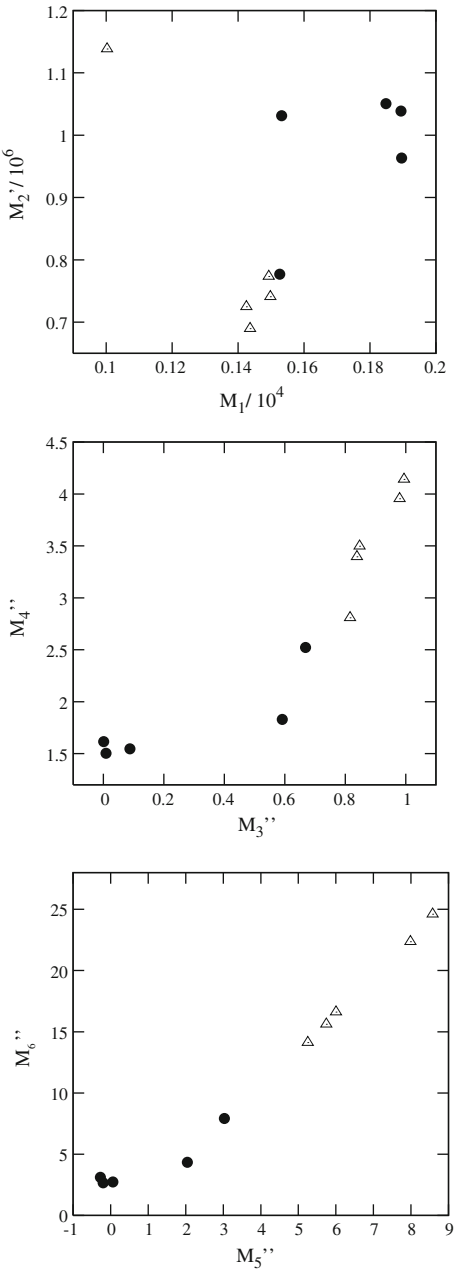

Fig. 7 Moment-based classification of the molecules. *Dots* correspond to nitriles and *triangles* to amides

As a consequence, these descriptors give the information about the concentrations of masses around the axes.

Another kind of new descriptors has been recently proposed by Huang et al. [98]. The authors proposed to take as the descriptors the set of characteristic vectors rep-

resenting all bases in the sequence. Guo and Wang obtained smooth curves from the zigzag curves and took curvatures of the smooth curves as descriptors of the sequences [145]. Yu et al. proposed two kinds of descriptors: a set of coordinates of TN curves, and the probabilities of occurring of particular trinucleotides among all 64 trinucleotides in the sequence [134]. Yu et al. composed 6D vector associated with the D-curve as a descriptor of DNA sequences [133].

Another kind of non-standard descriptors has also been introduced for four-component spectral representation (for the details see the next chapter). The descriptors are the numerical characteristics of the sequences. The next step would be the creation of similarity measures between sequences. In most of the similarity studies the set of descriptors characterizing a sequence is treated as components of a vector. Usually, as the similarity measure the Euclidean distance between the components of the vectors corresponding to a pair of sequences is taken. In particular, for identical sequences, this similarity measure is equal to zero.

Recently, non-standard measures have been introduced. For example Huang et al. defined a measure that changes from 0 to 1 and is equal to 1 for identical sequences [98]. Chen et al. constructed *cosine value* that is a similarity measure of the mean x, y, z coordinates of their graphs [127]. We have used the Manhattan distance normalized by the mean value of the descriptors for the similarity studies of the sequences represented by the 2D-dynamic graphs [101,146]. For identical sequences this measure is equal to zero, as it is assumed for most of similarity studies. Another non-standard similarity measure, also normalized to zero for identical sequences, is introduced in this work for four-component spectral representation (for details see subsequent chapter).

However, in the alignment studies the similarity measure changes from 0 to 100 for identical sequences. Such a measure is also defined for four-component spectral representation ([147], see next chapter).

Another similarity measure, also normalized to 100 for identical sequences, we have used for comparisons of 2D-graphs. This similarity measure introduces non-conventional treatment of graphs and their similarity analysis. We have not calculated the descriptors but the similarity measure has been directly obtained from the graphs.

In our studies we treated the graphs as rigid bodies, as in the classical dynamics. As a similarity measure for a pair of sequences represented by the graphs we took mass overlaps [146]. Using the genetic methods, very efficient in problems of optimization, we found the locations of a pair of graphs for which their mass overlap reaches maximum. In this position the similarity measure is defined as a mass overlap of a pair of graphs. In the process of maximization of the mass overlap we considered shifts and rotations of the graphs.

5 Four-component spectral representation

Recently, I have introduced another graphical representation [65]. In this section, the details and new aspects of this representation are described. Graphically, this representation resembles the molecular spectrum so I call it spectral representation. The DNA sequence is represented by a four-component function (or, graphically, by a four-component spectrum). A single DNA sequence is represented by four abstract

spectra: one for bases A, one for C, one for T and one for G. This means that I decompose each sequence to four components. Each γ-component I call-γ spectrum where $\gamma = A, C, T, G$ denotes one of the bases. Each γ-component is given by a function that is a superposition of the Gaussian functions:

$$I^{\gamma}(x) = \sum_{p=1}^{N} n_p^{\gamma} \exp[-(x - \epsilon_p)^2], \tag{8}$$

where N is the length of the sequence, and

$$n_p^{\gamma} = \begin{cases} 1 & \text{if } \gamma \text{occupies the } p\text{-th position,} \\ 0 & \text{if a base different then} \gamma \text{occupies the} p - \text{th position} \end{cases}$$

is the occupation number of the base γ in the p-th position of the sequence. The deletions, inversions and insertions can easily be described by appropriate changes of the occupation numbers and by insertions of properly constructed subdistributions. In Eq. 8, x is a variable measured along the sequence.

The abstract spectrum $I^{\gamma}(x)$ represents the density of particular bases along the sequence. The p-th base is represented by a single Gaussian function $\exp[-(x-\epsilon_p)^2]$ with the maximum located at

$$x = \epsilon_p = (p-1)r, \quad r > 0. \tag{9}$$

The parameter r is the resolution of $I^{\gamma}(x)$. For the visualization of long sequences it is convenient to take small r. The resolution parameter r determines the differences between the maxima of the Gaussians. The details of spectra are better visible when r is large, i.e. when the neighboring maxima are well separated. With an increasing r the resolution becomes larger. If $r = 1$ then the maximum corresponding to the first base ($p = 1$) is located at $x = \epsilon_1 = 0$ and the maximum corresponding to the last base is located at $x = \epsilon_N = N - 1$. Generally, the locations of the consecutive bases in one of the fourth γ-spectra correspond to $x = 0, r, 2r, \ldots, (N-1)r$, i.e. each single Gaussian function makes the contribution to one of the fourth γ-spectra. If the neighboring γ bases are closely packed then the intensities (I^{γ}) increase. If the sequence does not contain one of γ bases then the contribution to γ-component may be zero and all the contributions are located in one of the three other γ-spectra. Generally, the distributions of particular bases along the sequences are asymmetric and this information is reflected in the form of $I^{\gamma}(x)$.

In principle, x may change from $-\infty$ to $+\infty$. However, in practical terms, $I^{\gamma}(x) = 0$ if $x < -r$ or $x > Nr$. Therefore one can assume that the graphs extend for

$$x \in \langle -r, Nr \rangle.$$

In this way the first and the last bases are considered in the same way as the other ones. However, for the numerical characterization related to this graphical representation the range from $-\infty$ to $+\infty$ is considered.

As the numerical characterization of the four-component spectral representation I propose the properly scaled distribution moments.

Analogously to the definitions of the moments of a discrete distribution (Eqs. 4–6), the q-th moment of the continues distribution $I^{\gamma}(x)$ reads

$$M_q^{\gamma} = c^{\gamma} \int\limits_{R(x)} I^{\gamma}(x) x^q dx, \tag{10}$$

where

$$c^{\gamma} = \left(\int\limits_{R(x)} I^{\gamma}(x) dx \right)^{-1} \tag{11}$$

is the normalization constant and $R(x)$ is the range of x for which the integrand does not vanish. The normalization has been introduced for the numerical characteristics of the sequences. Visualization is independent of the numerical calculations and it is more clear to consider unnormalized plots defined as γ-spectra in Eq. 8. Good descriptors of the distributions are also the centered moments $M_q^{\gamma'}$

$$M_q^{\gamma'} = c^{\gamma} \int\limits_{R(x)} I^{\gamma}(x)(x - M_1^{\gamma})^q dx, \tag{12}$$

for which the first moment is equal to 0, and also $M_q^{\gamma''}$

$$M_q^{\gamma''} = c^{\gamma} \int\limits_{R(x)} I^{\gamma}(x) \left[\frac{x - M_1^{\gamma}}{\sqrt{M_2^{\gamma} - (M_1^{\gamma})^2}} \right]^q dx \tag{13}$$

for which the first moment is equal to 0 and the second one is equal to 1.

Considering several lowest moments it is convenient to perform integrations over the whole range of x (from $-\infty$ to $+\infty$). The integration can be performed analytically and

$$M_q^{\gamma} = c^{\gamma} \sqrt{\pi} \sum_{p=1}^{N} n_p^{\gamma} \epsilon_p Q_p^{(q)}, \tag{14}$$

where

$$c^{\gamma} = \left(\sqrt{\pi} \sum_{p=1}^{N} n_p^{\gamma} \right)^{-1} = \left(\sqrt{\pi} N^{\gamma} \right)^{-1}, \tag{15}$$

$$Q_p^{(1)} = 1, \tag{16}$$

$$Q_p^{(2)} = \epsilon_p + \frac{1}{2\epsilon_p}, \tag{17}$$

$$Q_p^{(3)} = (\epsilon_p)^2 + \frac{3}{2}, \tag{18}$$

$$Q_p^{(4)} = (\epsilon_p)^3 + 3\epsilon_p + \frac{3}{4\epsilon_p}. \tag{19}$$

In the graphical representation defined in Eq. 8, the summations are performed from $p = 1$ to $p = N$ for each γ. However the contributions of many terms are zero. Only the terms for which the occupation number is different than zero give non-zero contribution to the γ-spectrum and their number is N^γ which is the number of γ bases in the sequence and

$$N = \sum_{\gamma=A,C,T,G} N^\gamma. \tag{20}$$

Let us take an example of a model sequence ATAT. The nonvanishing terms that make the contribution to A-spectrum are for $p = 1, 3$. In case of T-spectrum $p = 2, 4$ and for G and C-spectra all the contributions are zeros. As a consequence, the four-component spectrum is

$$\begin{aligned} I^G &= 0, \\ I^C &= 0, \\ I^A &= exp[-(x-\epsilon_1)]^2 + exp[-(x-\epsilon_3)]^2, \\ I^T &= exp[-(x-\epsilon_2)]^2 + exp[-(x-\epsilon_4)]^2. \end{aligned}$$

The descriptors associated with the four-component spectral representation (D_q^γ) have been defined as properly scaled distribution moments [65]. In particular

$$D_1^\gamma = \frac{M_1^\gamma}{r}, \tag{21}$$

$$D_2^\gamma = \frac{M_2^{\gamma\prime}}{r^2}, \tag{22}$$

and

$$D_q^\gamma = M_q^{\gamma\prime\prime} \tag{23}$$

for $q \geq 3$. As it has been shown in the article [65], due to the division by r, D_1^γ and D_2^γ become independent of the resolution.

A convenient tool for visualization are diagrams D_q^γ versus $D_{q'}^\gamma$ [147]. In particular, these diagrams can be used for an identification of genes. In this kind of visualization, different types of classified objects are clustered in different areas of the plots.

As a similarity measure between a pair of sequences labeled by i and j

$$d_q^{\gamma}(i,j) = \frac{\mathrm{Min}\{|D_q^{\gamma}(i)|, |D_q^{\gamma}(j)|\}}{\mathrm{Max}\{|D_q^{\gamma}(i)|, |D_q^{\gamma}(j)|\}} 100\% \tag{24}$$

is proposed, where $q = 1, 2, 3, 4$ [147]. Though q may be easily increased up to higher-orders, as we shall see, the information about similarity sequences is specific enough up to the fourth order. Let us note that d_q^{γ} is consistent with standard measures used in biology: For the identical sequences the similarity value equals 100% and it decreases (approaching 0) if the difference between the two D_q^{γ} increases.

The average information about the similarity of a pair of sequences is contained in the measure

$$d_q^{\mathrm{MEAN}}(i,j) = \sum_{\gamma=A,C,T,G} W^{\gamma}\, d_q^{\gamma}(i,j), \tag{25}$$

where

$$W^{\gamma} = \frac{N^{\gamma}(i) + N^{\gamma}(j)}{N(i) + N(j)} \tag{26}$$

are referred to as the *weights*, $N^{\gamma}(i)$ is the number of γ bases in the i-th sequence, and

$$N(i) = \sum_{\gamma=A,C,T,G} N^{\gamma}(i)$$

is the length of the i-th sequence.

In order to study the problem of convergence of the method with respect to the higher-order moments I consider, for a pair of sequences labeled by i and j, the similarity measure

$$d^n(i,j) = \frac{1}{n}\sum_{q=1}^{n} d_q^{\mathrm{MEAN}}(i,j), \tag{27}$$

where n is the maximum order of moments taken into account.

All definitions may be easily generalized for multiple similarity studies. If J sequences labeled by $\mathbf{i} \equiv \{i_1, i_2, \ldots i_J\}$ are matched then the measures are defined as

$$d_q^{\gamma}(\mathbf{i}) = \frac{\mathrm{Min}\{|D_q^{\gamma}(i_1)|, |D_q^{\gamma}(i_2)|, \ldots |D_q^{\gamma}(i_J)|\}}{\mathrm{Max}\{|D_q^{\gamma}(i_1)|, |D_q^{\gamma}(i_2)|, \ldots |D_q^{\gamma}(i_J)|\}} 100\% \tag{28}$$

and

$$d_q^{\mathrm{MEAN}}(\mathbf{i}) = \sum_{\gamma=A,C,T,G} W^{\gamma}\, d_q^{\gamma}(\mathbf{i}). \tag{29}$$

The weights

$$W^{\gamma} = \frac{N^{\gamma}(i_1) + N^{\gamma}(i_2) + \cdots + N^{\gamma}(i_J)}{N(i_1) + N(i_2) + \cdots + N(i_J)} \tag{30}$$

are equal to the relative numbers of γ bases in all the considered sequences and

$$d^n(\mathbf{i}) = \frac{1}{n}\sum_{q=1}^{n} d_q^{\mathrm{MEAN}}(\mathbf{i}). \tag{31}$$

The measures defined in Eqs. 28, 29 and 31 may change from 0% to 100%, analogously to the ones defined, respectively, in Eqs. 24, 25 and 27.

An alternative similarity measure is defined in this work as

$$s_q^{\gamma}(i,j) = 1 - exp\left[-\left(D_q^{\gamma}(i) - D_q^{\gamma}(j)\right)^2\right]. \tag{32}$$

s_q^{γ} is equal to 0 if the descriptors of the i-th and the j-th sequences are the same ($D_q^{\gamma}(i) = D_q^{\gamma}(j)$) and approaches 1 if the difference between the two descriptors increases. This similarity measure is analogous to the one that we have introduced in the molecular similarity studies [56].

I also introduce a similarity measure between the sequences labeled by i and j that carries the information about several (n) properties

$$S_{\gamma}^{i_1,i_2,\ldots i_n}(i,j) = \sqrt{\frac{1}{n}\left[\left(w_{i_1}s_{i_1}^{\gamma}(i,j)\right)^2 + \left(w_{i_2}s_{i_2}^{\gamma}(i,j)\right)^2 + \cdots + \left(w_{i_n}s_{i_n}^{\gamma}(i,j)\right)^2\right]}, \tag{33}$$

where $i_1 < i_2 < \cdots < i_n$ and $w_{i_1} \ldots w_{i_n}$ denote the weights. $S_{\gamma}^{i_1,i_2,\ldots i_n}(i,j)$ is also normalized to the values belonging to the range from 0 (identical properties) to 1. For example, if we consider similarity of three properties: the width, the asymmetry and the curtosis of the γ-spectrum that are described by s_2^{γ}, s_3^{γ} and s_4^{γ} respectively, then $n = 3, i_1 = 2, i_2 = 3, i_3 = 4$ and the similarity measure is

$$S_{\gamma}^{2,3,4}(i,j) = \sqrt{\frac{1}{3}\left[\left(w_2 s_2^{\gamma}(i,j)\right)^2 + \left(w_3 s_3^{\gamma}(i,j)\right)^2 + \left(w_4 s_4^{\gamma}(i,j)\right)^2\right]}. \tag{34}$$

In this work all the weights in Eqs. 33 and 34 are equal to 1.

The units of descriptors D_{i_k} (Eq. 23) are normalized for $i_k \geq 3$. As a consequence, for example $S_{\gamma}^{3,4}$ is a convenient measure for comparison of sequences of different lengths, if we are interested in the similarity information that is not related to the lengths of the sequences. If the information about the mean value D_1 or about the width D_2 of γ-spectra needs to be compared then $S_{\gamma}^{i_1,i_2,\ldots i_n}$, where i_k are 1 or 2 may be considered.

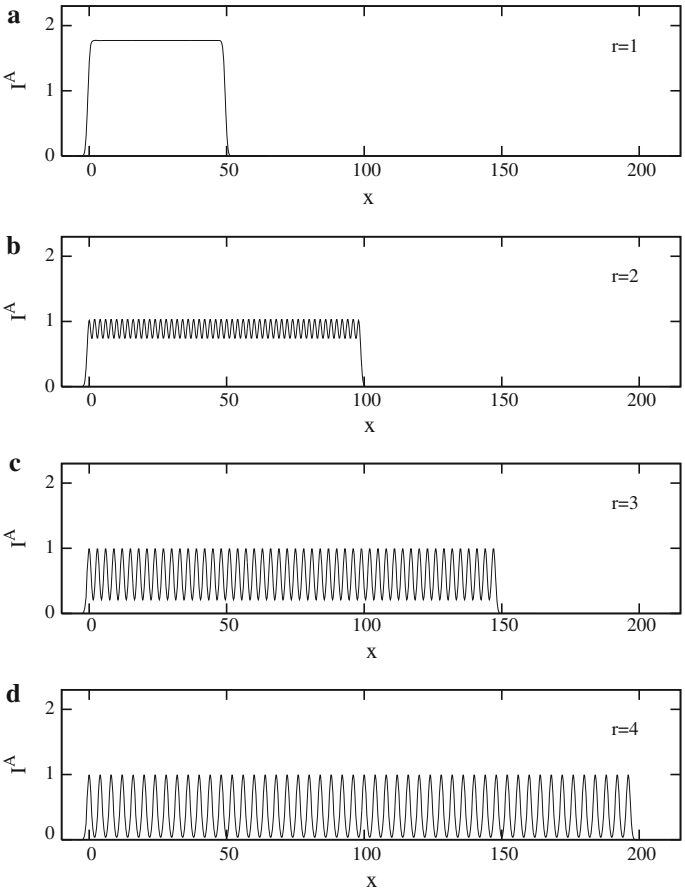

Fig. 8 Spectral representation of a model sequence AAAAA...AA ($N = 50$)

Figure 8 shows the spectral representation defined in Eq. 8 for a model sequence that consists of only A bases i.e. $N^C = N^T = N^G = 0$. The number of A bases is $N = N^A = 50$. In this case C, T, G-spectra are equal to zeros. $I^C(x) = I^T(x) = I^G(x) = 0$ for all x. The four-component spectrum representing this sequence is reduced only to one-component abstract spectrum $I^A(x) = \sum_{p=1}^{50} exp[-(x - \epsilon_p)^2]$. All the panels (a–d) in the figure represent the same model sequence. The difference is the resolution: $r = 1, r = 2, r = 3, r = 4$ in panels a, b, c, d respectively. The particular bases are represented by Gaussians centered at $\epsilon_p = (p - 1)r$, where $p = 1, 2, \ldots 50$. The first base is represented by a Gaussian with the maximum located at $\epsilon_1 = 0$ for all the cases and the last one at $\epsilon_{50} = 49, \epsilon_{50} = 98, \epsilon_{50} = 147, \epsilon_{50} = 196$ for $r = 1, r = 2, r = 3, r = 4$ respectively. For smaller r the bases are located close to each other and as a consequence the neighboring Gaussian functions overlap and we observe the envelope of the spectrum. In particular, if all the bases are the same, the spectrum becomes rectangular (Fig. 8, panel a). Increasing the resolution, the range for which the spectrum is different than zero becomes larger and we have a chance to

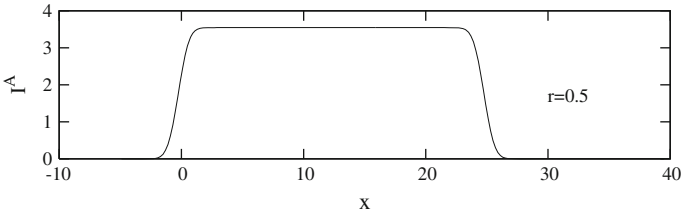

Fig. 9 Spectral representation of a model sequence AAAAA...AA ($N = 50$)

look into details of the spectra. The details are the locations of particular bases along the sequence. For long sequences, the balance between the details of spectra and the range of the plot determined by the location of the last Gaussian $\epsilon_N = (N-1)r$ has to be found. Theoretically, the resolution may change from a small positive value to infinity. However changing the resolution not always results in a change of the information coming from the spectrum. For example, if in the model example the resolution is taken as smaller than 1 then also rectangular representation is obtained. Figure 9 shows I^A spectrum for this model example where $r = 0.5$. The difference between $r = 1$ (Fig. 8, panel a) is the range ($\epsilon_{50} = 24.5$ for $r = 0.5$) and the maximum values of I^A. For smaller resolution the range of the spectrum is smaller and the neighboring maxima are located close to each other. As a consequence of closely located Gaussian functions $\exp[-(x-\epsilon_p)^2]$, the resulting maxima of spectrum I^A are larger (around 3 in Fig. 9 and around 2 in Fig. 8, panel a). However the qualitative information is the same in Fig. 8, panel a, and in Fig. 9.

In case of real sequences, there is a natural separation between the neighboring bases. Usually the resolution $r = 1$ and even smaller is sufficient for a good visualization. In Fig. 10, spectral representation of histone H1 coding sequence of Arabidopsis thaliana is shown ($i = 19$, Table 16). The length of the sequence is $N = 822$. The resolution has been taken as $r = 1$. The numbers of particular bases are $N^A = 259$, $N^C = 167$, $N^T = 188$, and $N^G = 208$. The largest number of A bases can be easily seen (large number of lines with large intensities as an effect of overlapping closely located Gaussians representing A bases). The same sequence but with the resolution ten times smaller is shown in Fig. 11. The resolution $r = 0.1$ seems to be sufficient to distinguish between those ranges of x for which the density of bases is larger comparing to ranges that are poor in the considered bases.

A very convenient way of a direct comparison of the difference between a pair of sequences labeled by i and j is plotting the difference I^γ_{ij}. Clearly, for both sequences $I^\gamma(x)$ must be represented with the same resolution in order to compare the distribution of γ bases along the sequence. Figures 12 and 13 show the differences between a pair of sequences. In Fig. 12 the differences with resolution $r = 1$ between the spectra representing histone H4 coding sequence of human ($i = 9$, Table 17) and histone H4 coding sequence of maize ($j = 1$, Table 17) are shown. Positive values of I^γ_{ij} indicate the regions of x for which the base γ is present in I^γ for the ith sequence and is not present in the jth one. The negative values of I^γ_{ij} indicate analogous regions but the γ bases are present in the sequence labeled by j. Such a simple visualization of overlapping of pairs sequences gives a direct information about the differences of

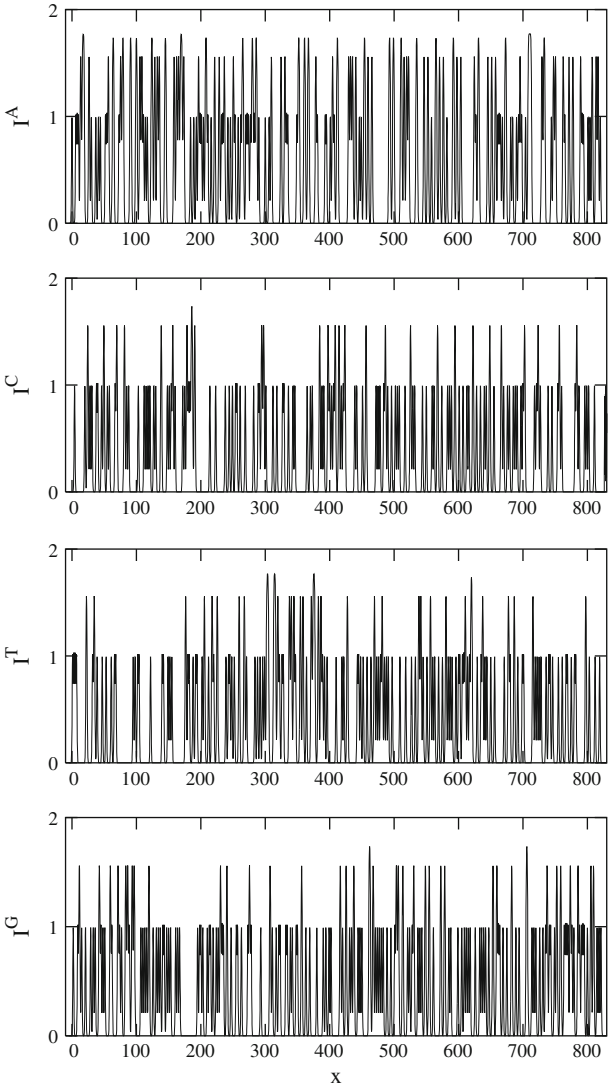

Fig. 10 Spectral representation of histone H1 coding sequence of *Arabidopsis thaliana* AY040059 ($r = 1.0$, $N = 822$, $i = 19$ in Table 16).

the distributions of particular bases along the sequences. In particular, the number of lines in I_{ij}^A, I_{ij}^T, I_{ij}^G is smaller than for I_{ij}^C. This means that the difference of the distributions of C bases along the sequences is the largest comparing to the differences of the distributions of other bases. Moreover comparing the number of lines that are positive to the ones that are negative, for a particular plot, one can easily estimate the differences between the numbers of the particular bases. For example, $N^C = 79$ for the sequence of human and $N^C = 96$ for the sequence of maize so the number of negative lines for I_{ij}^C is larger then the number of the positive ones. Analogously, the

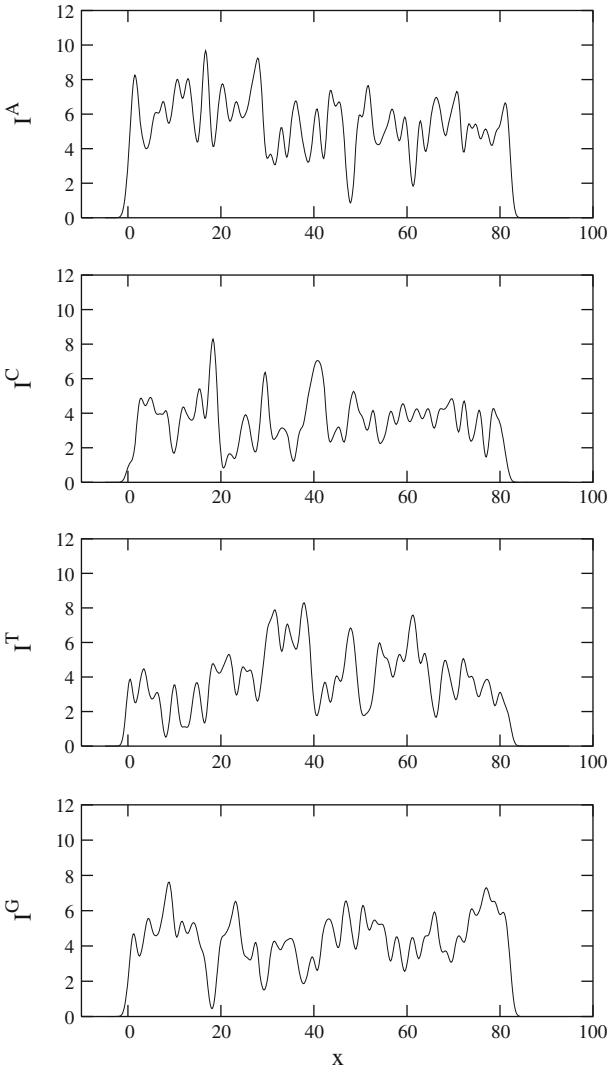

Fig. 11 Spectral representation of histone H1 coding sequence of *Arabidopsis thaliana* AY040059 ($r = 0.1$, $N = 822$ $i = 19$ in Table 16)

number of negative lines for I_{ij}^G can be seen: $N^G = 100$ for the sequence of human and $N^G = 111$ for the sequence of maize. Since the number of A and T bases are larger for the sequence of human then for the sequence of maize, one can observe in I_{ij}^A and I_{ij}^T plots more positive lines than the negative ones.

In Fig. 13 the differences with the resolution $r = 1$ between the spectra representing histone H4 coding sequence of human ($i = 9$, Table 17) and histone H4 coding sequence of mouse ($j = 7$, Table 17) are shown. As a result of the difference between

Table 16 Histone H1 coding sequences from the EMBL database

No.	Species	ID/Accession	N^A	N^C	N^T	N^G	N
1	Homo sapiens (Human)	M60747	214	177	90	185	666
2	Homo sapiens (Human)	M60748	195	199	61	205	660
3	Macaca fascicularis (Crab-eating macaque)	AB179307	49	30	33	65	177
4	Gallus gallus (Chicken)	X01752	165	228	44	223	660
5	Gallus gallus (Chicken)	M17018	170	230	43	220	663
6	Gallus gallus (Chicken)	M17019	179	220	56	223	678
7	Gallus gallus (Chicken)	M17020	175	219	52	214	660
8	Gallus gallus (Chicken)	M17021	178	221	58	218	675
9	Mus musculus (Mouse)	L26164	188	170	91	193	642
10	Mus musculus (Mouse)	Z46227	191	184	96	201	672
11	Mus musculus (Mouse)	Z38128	176	200	74	216	666
12	Mus musculus (Mouse)	X13171	187	177	68	153	585
13	Mus musculus (Mouse)	X72805	172	170	99	186	627
14	Mus musculus (Mouse)	J03482	168	195	65	211	639
15	Mus musculus (Mouse)	M25365	168	196	65	210	639
16	Rattus norvegicus (Rat)	BC061842	187	180	64	154	585
17	Rattus norvegicus (Rat)	X72624	187	180	64	154	585
18	Arabidopsis thaliana (Thale cress)	AY079414	193	99	96	116	504
19	Arabidopsis thaliana (Thale cress)	AY040059	259	167	188	208	822
20	Arabidopsis thaliana (Thale cress)	AY045797	193	99	96	116	504
21	Arabidopsis thaliana (Thale cress)	AF360211	259	167	188	208	822
22	Triticum aestivum (Wheat)	X59872	167	269	49	229	714
23	Triticum aestivum (Wheat)	AF107022	167	267	49	228	711
24	Triticum aestivum (Wheat)	AF107023	168	257	58	231	714
25	Triticum aestivum (Wheat)	AF107024	194	319	52	263	828
26	Triticum aestivum (Wheat)	AF107027	169	269	54	225	717
27	Solanum lycopersicum (Tomato)	AJ224933	284	183	144	205	816
28	Zea mays (Maize)	X57077	187	267	62	225	741
29	Zea mays (Maize)	EU952324	139	214	64	174	591
30	Zea mays (Maize)	EU953635	154	223	60	271	708
31	Zea mays (Maize)	EU954558	168	251	70	252	741
32	Zea mays (Maize)	EU957928	185	271	69	243	768
33	Zea mays (Maize)	EU959342	192	276	69	249	786
34	Zea mays (Maize)	EU960344	125	167	54	218	564
35	Zea mays (Maize)	EU961871	147	211	60	260	678
36	Zea mays (Maize)	EU963944	147	211	63	260	681
37	Zea mays (Maize)	EU964093	194	315	58	255	822

the numbers of A bases one can observe in I_{ij}^A more positive lines than the negative ones: $N^A = 73$ for the sequence of human and $N^A = 65$ for the sequence of mouse. The difference in C bases is also clearly seen. There are more negative than positive

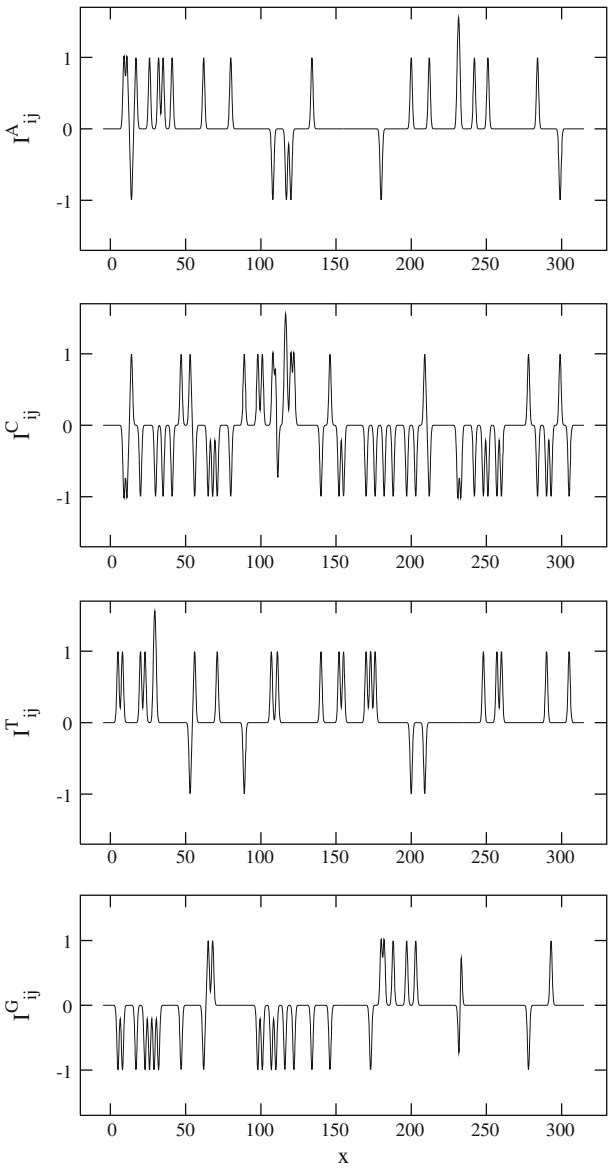

Fig. 12 Differences between the spectra for histone H4 coding sequence of human M60749 and histone H4 coding sequence of maize M13377 ($i = 9$, $j = 1$, Table 17)

lines in I^C_{ij} plot: $N^C = 79$ for the sequence of human and $N^C = 96$ for the sequence of mouse. Generally, comparing Fig. 12 and Fig. 13 one can see that the differences human-maize spectra are larger then the differences human-mouse spectra (the number of lines in Fig. 12 is larger then the number of lines in Fig. 13).

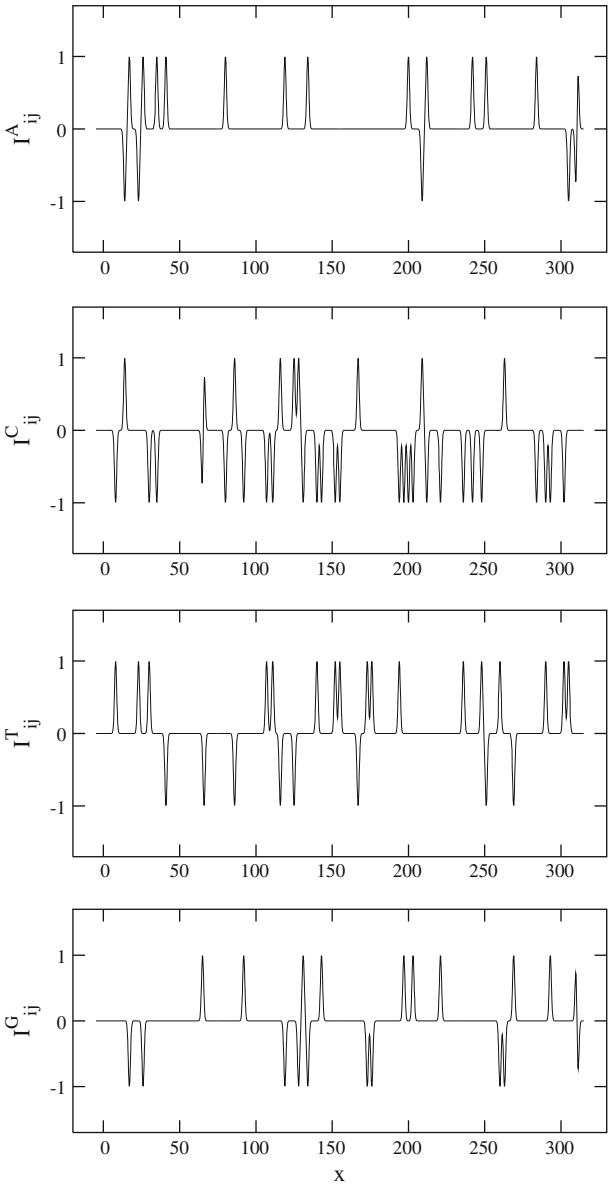

Fig. 13 Differences between the spectra for histone H4 coding sequence of human M60749 and histone H4 coding sequence of mouse V00753 ($i = 9$, $j = 7$, Table 17)

As the descriptors of the four-component spectral representation, I have proposed D_q^{γ}. Figure 14 shows $D_1^G - D_4^G$ diagram for ten sequences listed in Table 17 and for one additional sequence (one point in the figure represents descriptors of one sequence). The additional sequence is histone H4 coding sequence of human (M16707). In many articles the ten sequences were treated as a model set to introduce new graphical and

Table 17 Histone H4 coding sequences from the EMBL database

No.	Species	ID/Accession	N^A	N^C	N^T	N^G	N
1	Zea mays (Maize)	M13377	62	96	43	111	312
2	Zea mays (Maize)	M13370	60	101	41	110	312
3	Zea mays (Maize)	M36659	63	96	42	111	312
4	Gallus gallus (Chicken)	M74533	62	104	38	108	312
5	Gallus gallus (Chicken)	M74534	62	105	37	108	312
6	Triticum aestivum (Wheat)	M12277	62	111	37	102	312
7	Mus musculus (Mouse)	V00753	65	96	51	100	312
8	Rattus norvegicus (Rat)	M27433	68	93	51	100	312
9	Homo sapiens (Human)	M60749	73	79	60	100	312

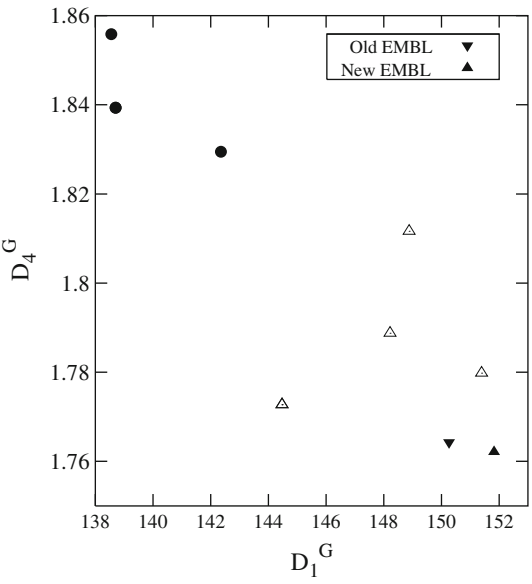

Fig. 14 $D_1^G - D_4^G$ classification diagram for the sequences listed in Table 17. *Dots* correspond to plants and *triangles* to vertebrates

numerical representations. However, there was a mistake in the old version of the EMBL database. Obviously, the length of this coding sequence should be 312 and not 311 as it was in the old version of the EMBL database. The additional base is G, located at the last position of the sequence. The descriptors D_q^V of spectral representation are very sensitive. The difference by only one base can be detected using these descriptors. Moreover, the approximate location of this base can be indicated. The descriptors characterizing the same sequence calculated using the old and new version of the EMBL database have been denoted using different symbols in Fig. 14. Their locations are different in the diagram. It is remarkable that the difference by this very base may be recognized in the plots.

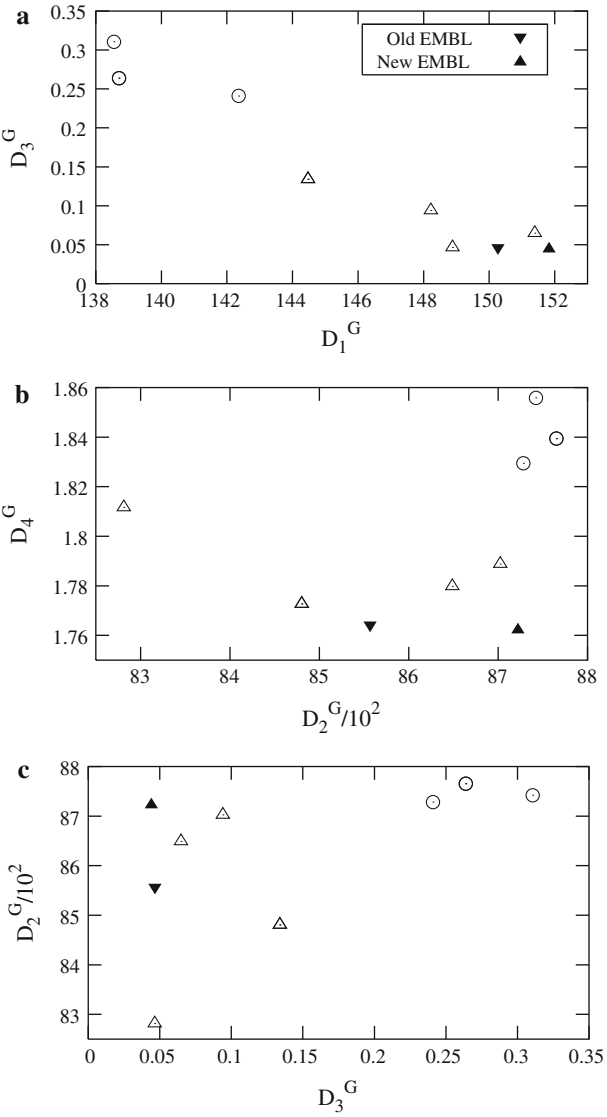

Fig. 15 $D_q^G - D_{q'}^G$ diagrams for the sequences listed in Table 17

Figures 15, 16, 17, 18 show the diagrams also for the sequences listed in Table 17. In particular, Fig. 15 shows diagrams for G-descriptors, Fig. 16 for A-descriptors, Fig. 17 for C-descriptors, and Fig. 18 for T-descriptors. Panels a in the figures correspond to $D_1^\gamma - D_3^\gamma$ diagrams, panels b to $D_2^\gamma - D_4^\gamma$ diagrams, and panels c to $D_3^\gamma - D_2^\gamma$ ones. Since the difference between the old and new version of the EMBL database is only one base G, the A, C, T descriptors are exactly the same for the new and old sequences. The difference in G-descriptors

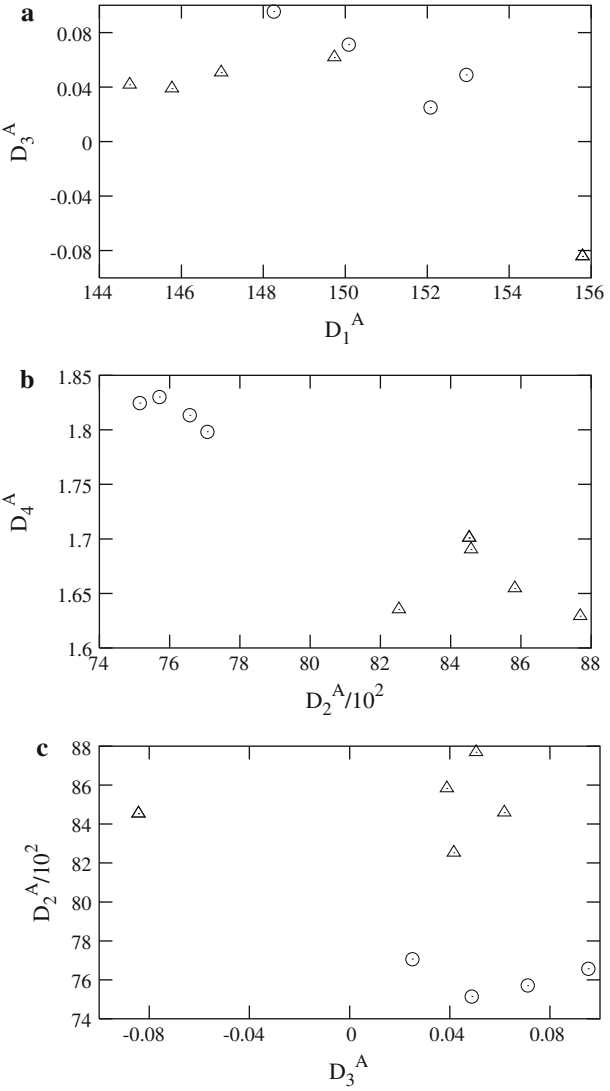

Fig. 16 $D_q^A - D_{q'}^A$ diagrams for the sequences listed in Table 17

indicates the location in the sequence of the base that is different for a pair of sequences. The additional G base in the new sequence causes the shift to larger values of the mean of the distribution (D_1^G becomes larger, Fig. 14, Fig. 15, panel a). The width of the distribution also increases (D_2^G for the new sequence is larger than for the old one, Fig. 15 panels b and c). Higher order descriptors representing asymmetry and kurtosis (D_3^G and D_4^G) also change. The differences between the G-descriptors using the new and old data for histone H4 coding sequence of human (M16707), $D_q^G(no) = D_q^G(new) - D_q^G(old)$ are as fol-

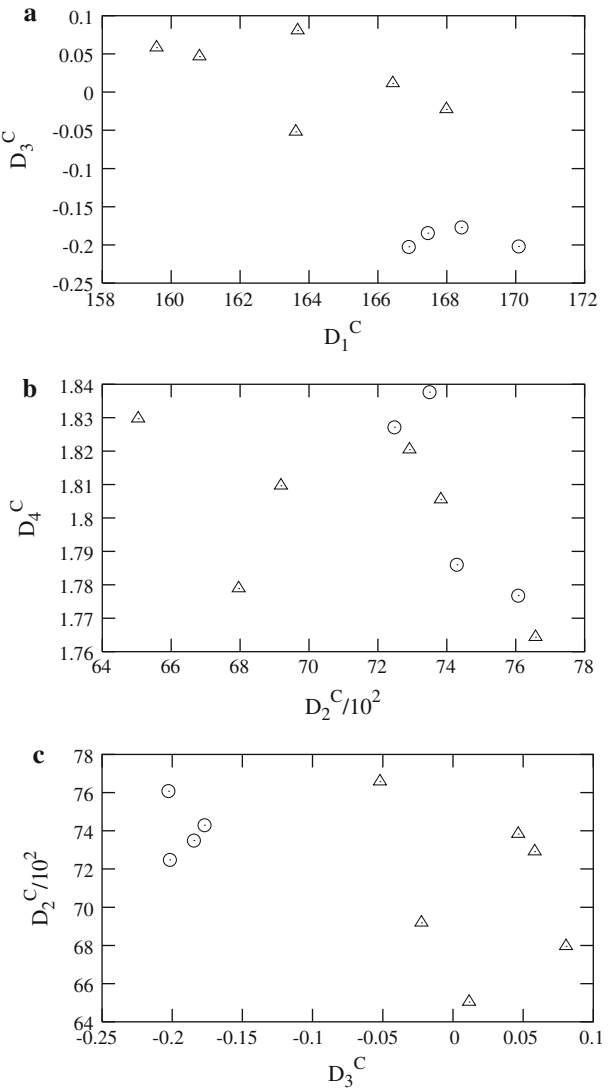

Fig. 17 $D_q^C - D_q'^C$ diagrams for the sequences listed in Table 17

lows: $D_1^G(no) = 1.5605$, $D_2^G(no) = 165.3269$, $D_3^G(no) = -0.0024$, $D_4^G(no) = -0.0022$.

Considering the properties of G and A-spectra (G and A-descriptors shown in Figs. 15, 16, respectively) one can observe clusterization of evolutionary similar organisms: plants and vertebrates that are represented by different symbols in the plots (plants-circles, vertebrates-triangles). Considering the properties of C-spectra (Fig. 17) one can find the properties that are specific for plants and different than for vertebrates and also one can find the properties that are com-

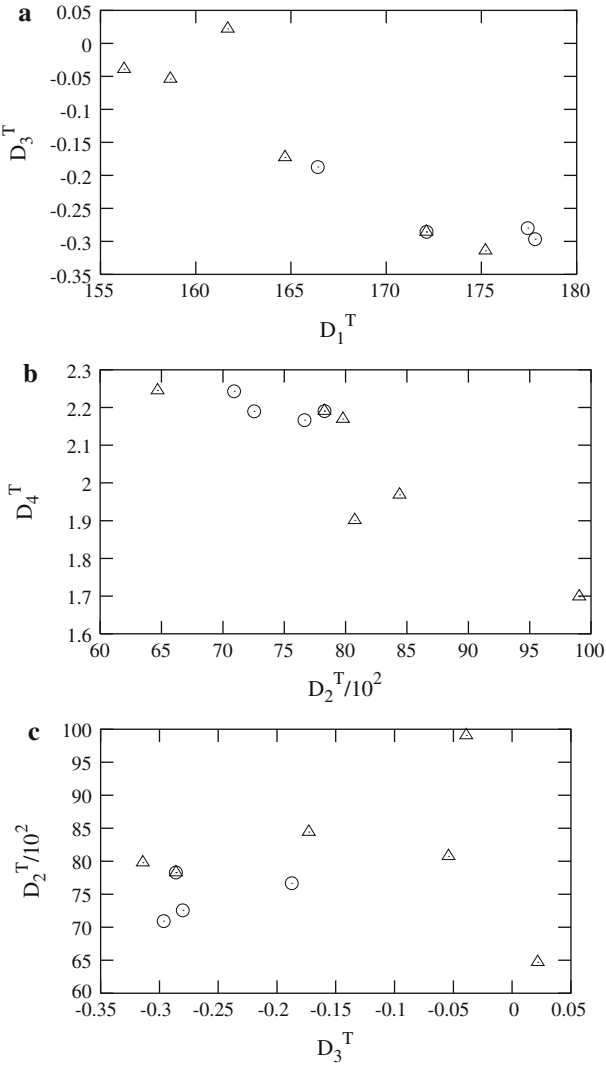

Fig. 18 $D_q^T - D_{q'}^T$ diagrams for the sequences listed in Table 17

mon for plants and vertebrates. For example in panels a and c where $D_1^\gamma - D_3^\gamma$, $D_3^\gamma - D_2^\gamma$ are shown one can observe clusterization. However in panel b, where $D_2^\gamma - D_4^\gamma$ is shown the properties mix for plants and vertebrates. For T-descriptors (Fig. 18) in all the diagrams plants and vertebrates are mixed, they even overlap.

It is interesting to note that most of the similarity measures (both the standard ones and many alternative ones) indicate larger or equal similarity values between histone H1 coding sequences of chicken (labeled by $i = 4, 5$ in Table 17) and plants (labeled in this Table by $j = 1, 2, 3, 6$) than between these of chicken and of vertebrates (labeled

Table 18 Similarity measures between a pair of sequences labeled by i and j $Sim(i, j)$, where i and j are defined in the first column of Table 17

Sim	$Sim(5, 6)$	$Sim(5, 7)$
CL	88	88
d_3^G	43	70
d_3^A	30	46
d_3^C	25	48
d_3^T	100	19

by $j = 7, 8, 9$). However, using new similarity approach it is possible to extract such components of the similarity measures that cluster the sequence of chicken with the ones of vertebrates rather than with the ones of plants [147]. Table 18 shows similarity values obtained using different similarity measures "Sim". Using alignment method (Sim=CL) the similarity value "chicken-plant" CL(5,6) is the same as the similarity value "chicken-vertebrate" CL(5,7). Considering different aspects of similarity, using d_3^γ, one can see that the clusterization of the sequence of chicken with vertebrates is obtained for $\gamma = G, A, C$. However the asymmetry of the gene structure for T bases is identical for the sequence of chicken and of plants ($d_3^T(5, 6) = 100$) and the similarity value is small in case "chicken-vertebrate" ($d_3^T(5, 7) = 19$).

Figures 19 and 20 show the diagrams for the sequences listed in Table 16 (histone H1 coding sequences of different species). In particular, Fig. 19 shows $D_3^\gamma - D_4^\gamma$ diagrams, and Fig. 20 $D_1^\gamma - D_2^\gamma$ ones. Panels a in the figures correspond to A-descriptors, panels b to C-descriptors, panels c to T-descriptors, and panels d to T-descriptors. The lengths of the considered sequences are different. Usually, biologists are interested in comparisons of sequences using measures that are independent of the length of the sequences. For that purpose the normalized descriptors of order 3 and higher (D_q^γ, with $q = 3, 4, \ldots$) can be used, as for example in Fig. 19. However, if the properties dependent on the lengths are of interests then $D_1^\gamma - D_2^\gamma$ diagram gives a good characteristic of the objects. As it is well known, the lengths of the sequences are not related to the complexity of the organisms. The diagrams $D_1^\gamma - D_2^\gamma$ confirm this observation: plants and vertebrates (circles and triangles in the figures) mix for all γ. Approximately, the dependence between D_1^γ and D_2^γ is linear. The most regular linear dependence is for G-descriptors (Fig. 20, panel d). However, using the diagrams for the descriptors independent of the lengths of sequences (Fig. 19), for A and G-descriptors (panels a, d respectively) the clusterization of plants and vertebrates is observed. For C-descriptors, the effect of clusterization is smaller. The effect of clusterization is not observed for T-descriptors. T-descriptors representing sequences of plants and vertebrates even overlap. These observations are the same as in the case of histone H4 coding sequences.

Figures 21, 22, 23, 24, 25 show the relations between the standard calculations Clustal W (CL) and the new measures (Eqs. 24, 25, 27) for the sequences listed in Table 17 (histone H4 coding sequences).

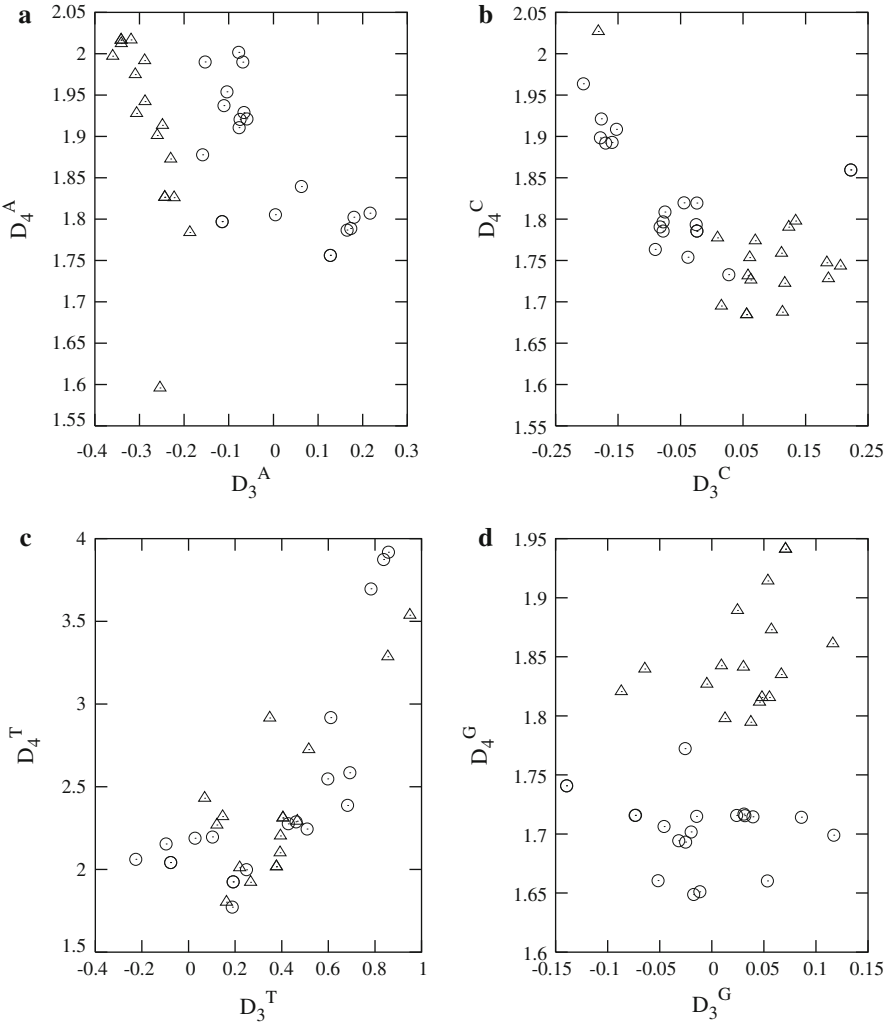

Fig. 19 $D_3^\gamma - D_4^\gamma$ diagrams for the sequences listed in Table 16

Figures 21, 22, 23, 24, 25, 26, 27, 28, 29 are plotted in the same way as it has been done in chapter 2 (Figs. 1, 2, 3). Each point in the plot corresponds to one case: comparison of sequence of species No. i with sequence of species No. j using different methods. For example, the horizontal axis in Fig. 21 corresponds to the similarity matrix between sequences of different species using Clustal W method (CL) and the vertical axes correspond to the similarity matrix between the same sequences using different components of alternative similarity measures d_q^γ. As a consequence each plot represents two similarity matrices, which gives a better visualization of the relations between two different similarity measures. In the figures, the functions $x = y$, where x and y represent, respectively, the horizontal and vertical axes, are plotted

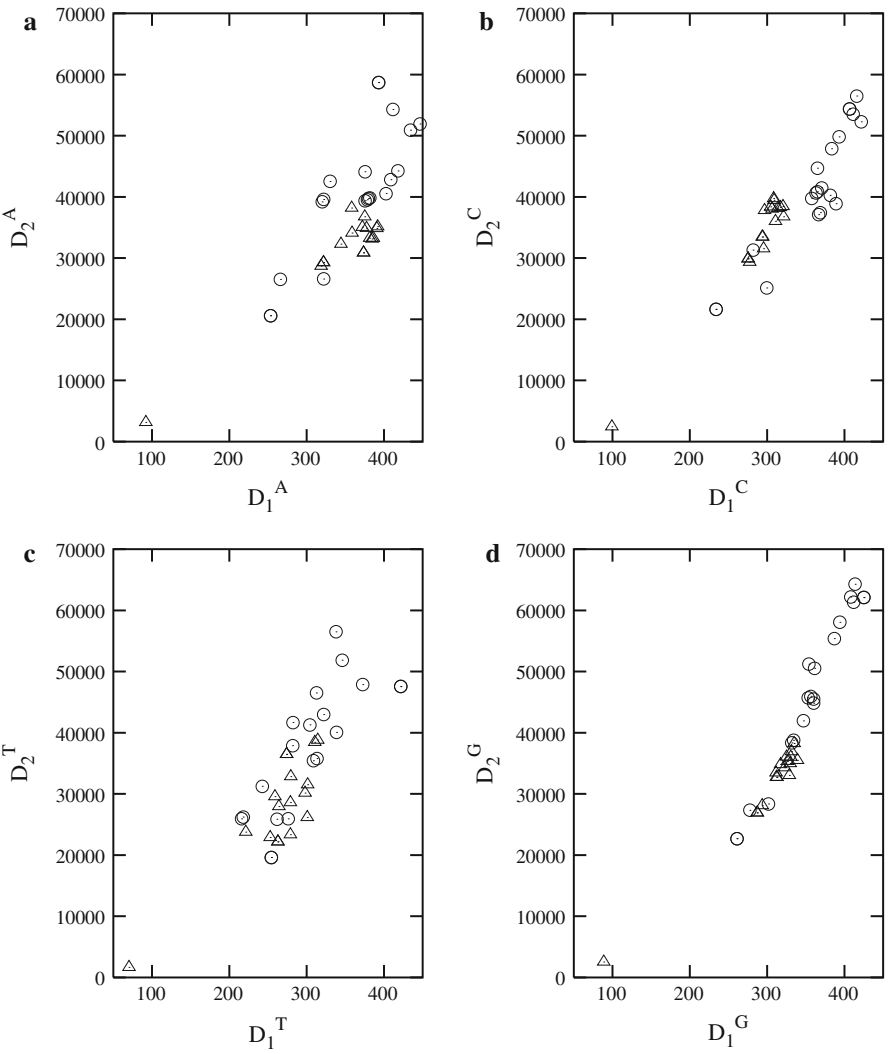

Fig. 20 $D_1^\gamma - D_2^\gamma$ diagrams for the sequences listed in Table 16

(dashed lines). Comparing the distributions of the points around the dashed lines it is easy to recognize these aspects of similarity for which the relations are the same. If the points are concentrated close to the lines then the similarity relations represented by x and y axes are also close to each other.

Figure 21 shows $CL - d_q^A$ diagram, Fig. 22 $CL - d_q^C$ diagram, Fig. 23 $CL - d_q^T$ diagram, and Fig. 24 $CL - d_q^G$ one. The panels a, b, c, d in the figures correspond to $q = 1$, $q = 2$, $q = 3$, and $q = 4$, respectively.

The similarity matrix CL based on Clustal W approach for the considered sequences is given in [146]. Small range of similarity measures indicates small differences

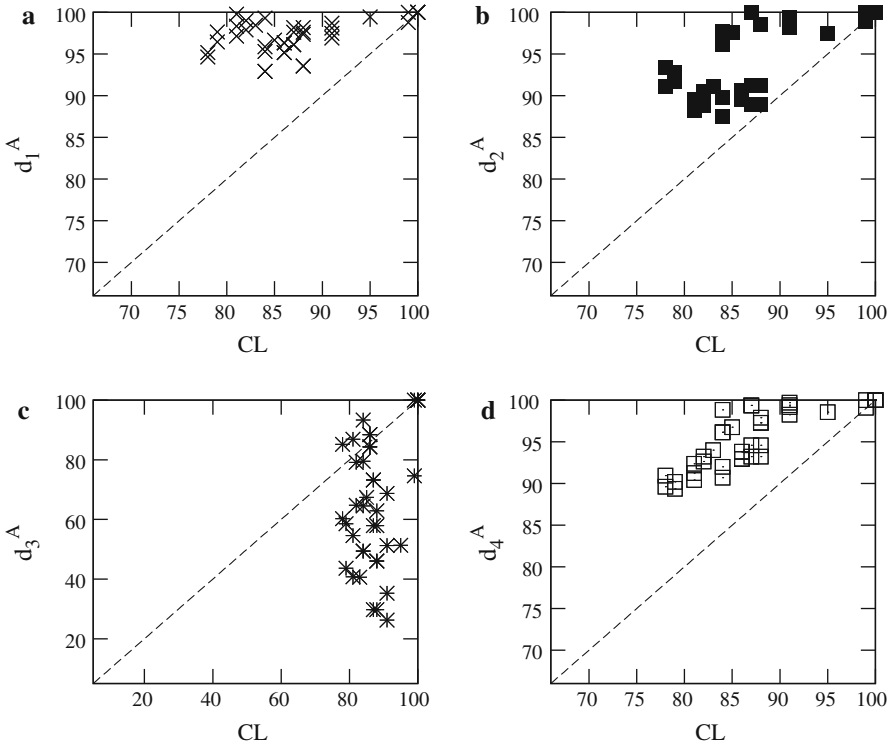

Fig. 21 $CL - d_q^A$ diagrams for the sequences listed in Table 17

between the sequences of different species. The range of values of CL is from 78% to 100%. The ranges of values of d_q^γ for $q = 1$, $q = 2$, and $q = 4$ are smaller than for $q = 3$ for all γ. d_3^γ changes from about 15% to 100% for all γ. The differences between sequences across species using d_4^γ are smaller than using d_3^γ, but different for all γ. The ranges of d_4^γ are the largest for T bases (d_4^T changes from about 75 to 100%) and they are very small for C and G (from about 95 to 100%). For $q = 1$ and $q = 2$ the ranges are the largest for T bases.

The dashed lines in the figures correspond to $CL = d$. The relations of each γ-component of the measure (d_q^γ) with standard measure is different. For the G-components one can observe the cumulating of all the points far from the dashed lines. This means that the information coming from G-measures gives most different results comparing to the standard measure. The points come closer to the dashed lines for A and C measures. For T-measures the points even cross the lines.

Since each γ-component is related in a different way to the standard measure, one may expect that it carries independent similarity information. Averaging the measures over γ, and then averaging over q, d_q^{MEAN} (Eq. 25) and $d^n(i, j)$ (Eq. 27) are obtained. In particular $d^1(i, j) = d_1^{MEAN}(i, j)$. Figure 25 shows the relations of d^n with the standard measure. The convergence of d^n measures to the standard measure CL we have discussed in [147]. In the present paper this effect is shown in detail adding d^1

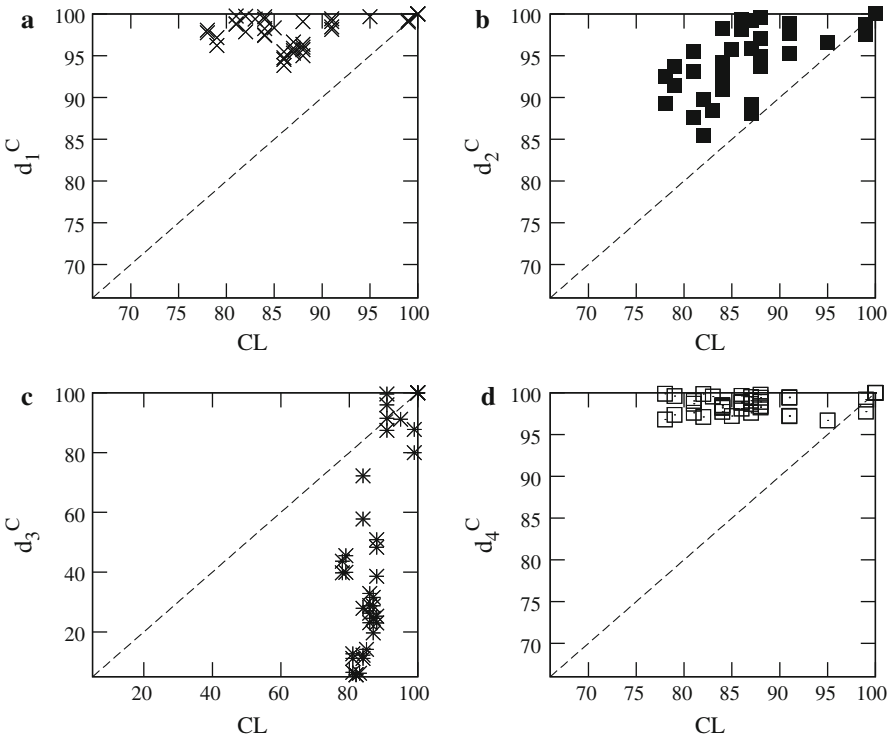

Fig. 22 $CL - d_q^C$ diagrams for the sequences listed in Table 17

term. d^1 is very different from CL (the points are located far away from the dashed line, panel a, Fig. 25). Adding higher-order terms, the points are pushed towards the dashed lines (panels b, c, d Fig. 25).

Figures 26, 27, 28, 29, 30 show similarity relations for β-globin gene across species using similarity measures defined in Eqs. 32 and 33. These data are the standard ones for alternative methods. Since the sequences in the database are not complete for some species, they are unified in this work and the appropriate locations in the gene are listed in the tables. In particular, the sequences of mouse and of chicken belong to the standard set of data used by many authors. However, several bases are ambiguous for the third exons for the sequences of the two species. As it was already mentioned, the method used in this work is so sensitive that even a difference in a single base can influence the results. Therefore the sequences of mouse and of chicken are omitted from this consideration. Moreover, in gorilla and chimpanzee sequences the stop codons are not available in the database. Therefore for all the species the stop codons are excluded from the calculations. This means that the length of the coding sequence N_{CDS} is three times larger than the corresponding length of the protein sequence for all the species. In this way (excluding the stop codons) all the data used in the calculations are consistent.

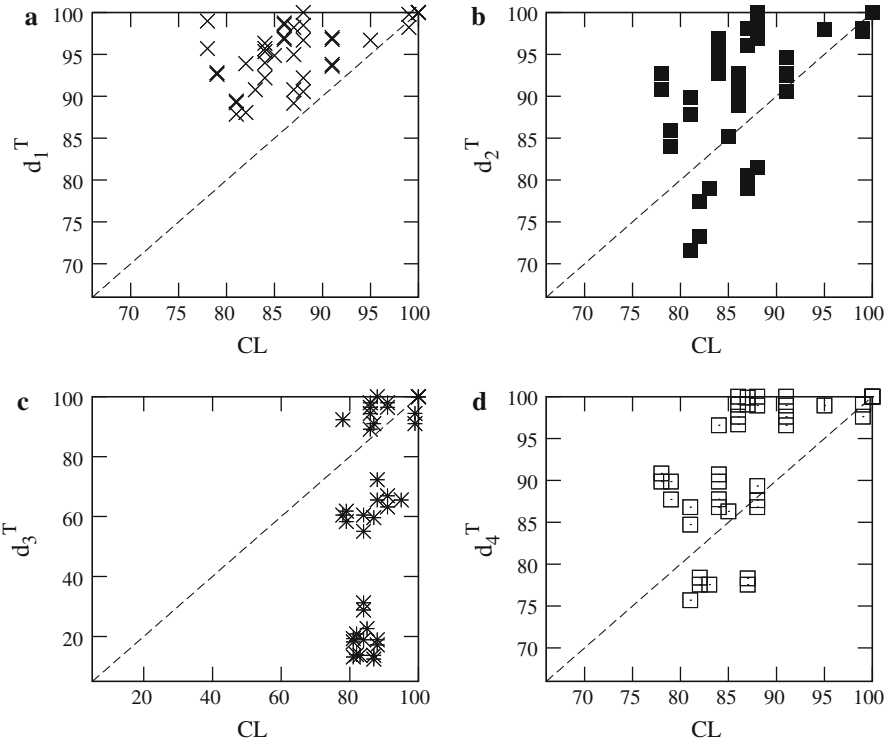

Fig. 23 $CL - d_q^T$ diagrams for the sequences listed in Table 17

The locations in the gene, the numbers of γ bases, N_k^γ, for each k-th exon according to the latest version of the EMBL database are specified in Tables 19, 20, 21, 22, 23. The lengths of the k-th exon, where $k = 1, 2, 3$

$$N_k = \sum_{\gamma=A,C,T,G} N_k^\gamma \tag{35}$$

are also given.

Many authors of alternative methods consider in their studies only the first exon, in fact only the coding part of this exon. In this work, different parts of the gene are considered. The calculations have been performed for several sets of data for different species:

1. Sequences with introns with the length N_{PlusI}, denoted PlusI,
2. Parts of the first exons starting with the start codon (coding sequences of the first exons) with the length N_1, denoted Exon 1^{CDS},
3. The second exons with the length N_2, denoted Exon 2^{CDS},
4. Parts of the third exons excluding the stop codons (coding sequences of the third exons) with the length N_3, denoted Exon 3^{CDS},

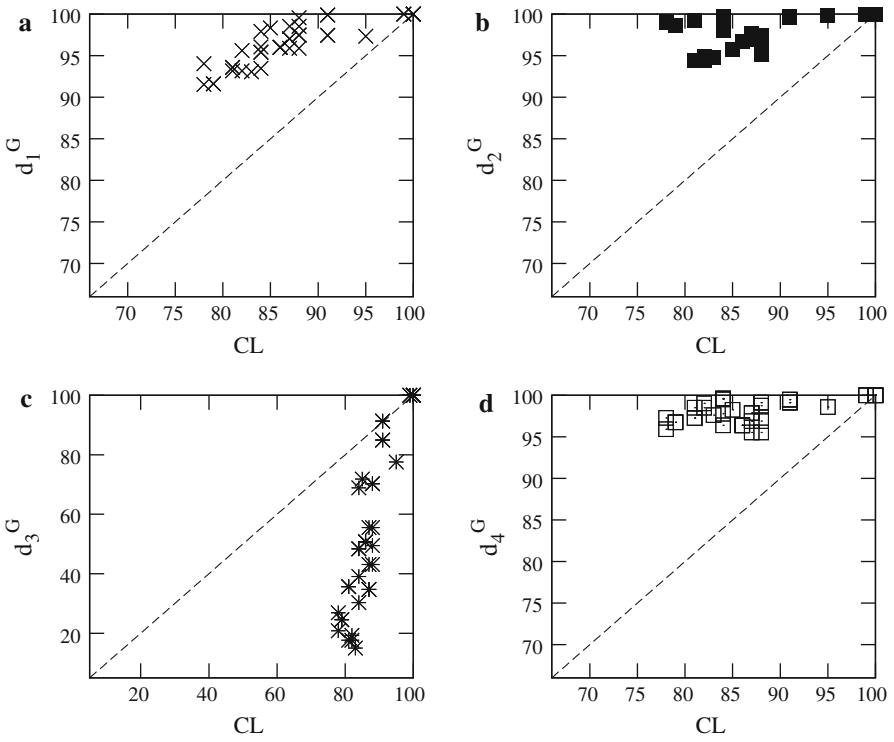

Fig. 24 $CL - d_q^G$ diagrams for the sequences listed in Table 17.

5. The whole first exons which are given in the EMBL database only for three species with the length N_{W1}, denoted Exon 1,
6. The coding sequences with the lengths $\mathrm{N_{CDS}} = \sum_{\mathrm{k}=1}^{3} \mathrm{N_k}$, denoted CDS.

Figure 26 shows $s_3^\gamma - s_4^\gamma$ diagrams for the β-globin coding sequences (CDS) across the species. Panels a, b, c, and d correspond, respectively, to A, C, T, and G bases. The horizontal axes represent the similarity measure based on asymmetry, s_3^γ, and the vertical ones represent the similarity measure based on the kurtosis of the distributions of γ bases along the sequence, s_4^γ. The largest differences between s_3^γ and s_4^γ are for $\gamma = C$. For G distributions most of the points are concentrated very close to the dashed line (similarity relations across species are nearly the same using s_3^G and s_4^G for the coding sequences). Moreover we observe small values of s_4^C, s_4^T, and s_4^G (panels b, c, d) i.e. large similarities between the kurtosis of C, T, and G spectra for different species. The similarity relations between the species based on the comparison of both the asymmetry and the kurtosis $S_\gamma^{3,4}$ are shown in Fig. 27. Panel a shows $S_C^{3,4} - S_A^{3,4}$ diagram, panel b—$S_T^{3,4} - S_A^{3,4}$ diagram, panel c—$S_G^{3,4} - S_A^{3,4}$ diagram, and panel d—$S_G^{3,4} - S_C^{3,4}$ diagram. All the panels correspond to CDS. In all the cases points are

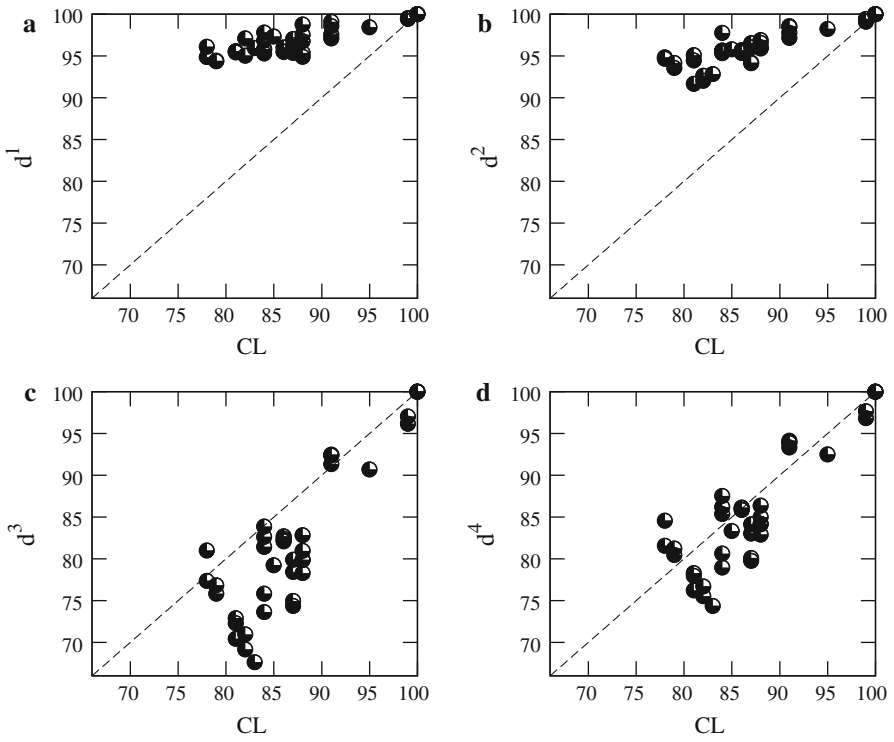

Fig. 25 $CL - d^n$ diagrams for the sequences listed in Table 17

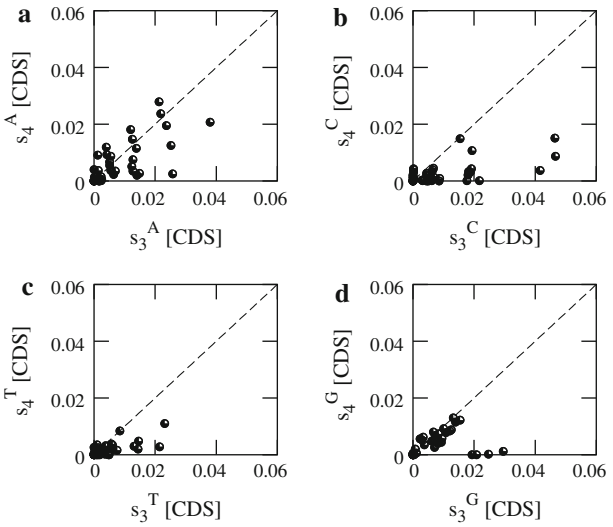

Fig. 26 $d_3^\gamma - d_4^\gamma$ diagrams for β-globin coding sequences across species

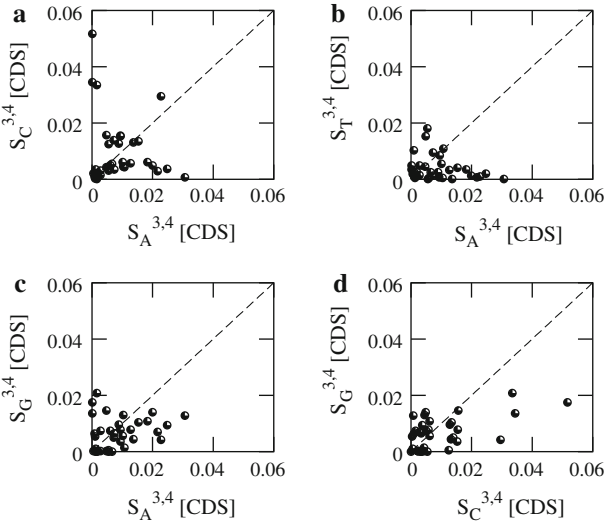

Fig. 27 $S_\gamma^{3,4} - S_{\gamma'}^{3,4}$ diagrams for β-globin coding sequences across species

distributed far from the dashed lines and this indicates that there are no correlations between these measures for different γ. Each $S_\gamma^{3,4}$ carries independent information.

$S_\gamma^{3,4}$ are also shown in Fig. 28. In this figure the measures are compared for different parts of the β-globin gene. The horizontal axes correspond to the sequences with introns, PlusI. The vertical axes correspond to the coding sequences of particular exons: column 1 to Exon 1^{CDS}, column 2 to Exon 2^{CDS} and column 3 to Exon 3^{CDS}. The first row of subfigures correspond to A bases, the second row to C bases, the third row to T bases and the fourth row to G bases.

We observe that the points are concentrated around the dashed lines in the middle column (Exon 2^{CDS}) comparing to the first and to the third columns. Small deviations from the dashed lines mean that the second exon is most representative in the whole sequence, PlusI (the similarity relations across species fulfilled by PlusI and and by Exon 2^{CDS} are closer to each other than the relations fulfilled by PlusI and by the other exons). We have also shown that the similarity relations across species fulfilled by CDS and and by Exon 2^{CDS} are closer to each other than the relations fulfilled by CDS and by the other exons [71].

If we compare the distributions of the points between different bases (rows) one can extract some properties common for particular bases and for some parts of the genes. By a common property we understand close to zero $S_\gamma^{3,4}$ (small values correspond to large similarities). In particular small differences between sequences across species are revealed for G bases for the first and for the second exons (panels j, k) and also for C and for T bases for the second exon (panels e, h). Generally, larger differences are seen for longer sequences. However also for PlusI one can extract properties more common for the species (small ranges of $S_A^{3,4}$[PlusI] and $S_T^{3,4}$[PlusI]—first and third rows).

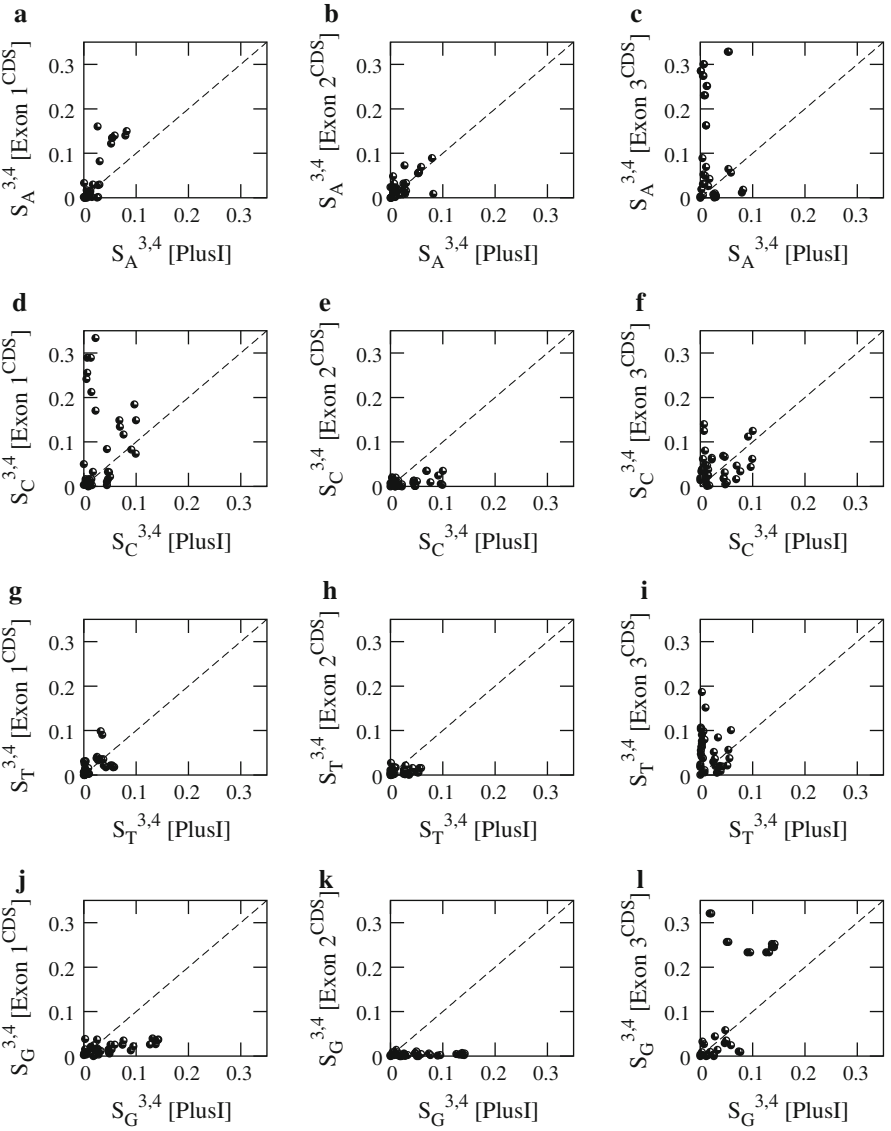

Fig. 28 $S_\gamma^{3,4}$ – $S_\gamma^{3,4}$ diagrams for the sequences listed in Tables 19, 20, 21, 22

Figure 29 shows similarity relations for different exons using standard alignment method Clustal W version 2.0 [148]. As it was mentioned before, the alignment methods do not take into account which bases are aligned. The alignment of all the bases gives the contribution to the final result and, as a consequence, the similarity is large for all the parts. It is not possible to extract detailed properties of similarities. The information coming from these calculations is averaged. Finally, the similarity

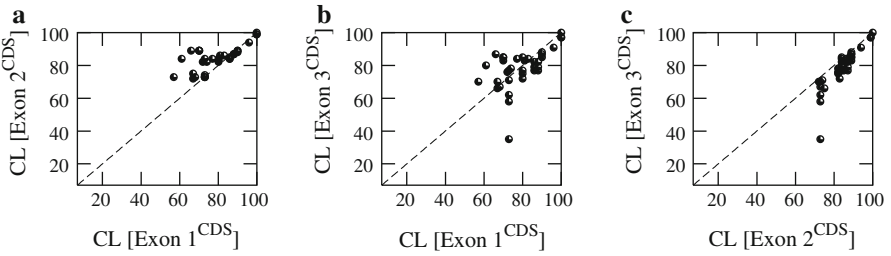

Fig. 29 CL[Exon 1^{CDS}] − CL[Exon k^{CDS}] diagrams, where $k = 1, 2, 3$ for the sequences listed in Tables 20, 21, 22

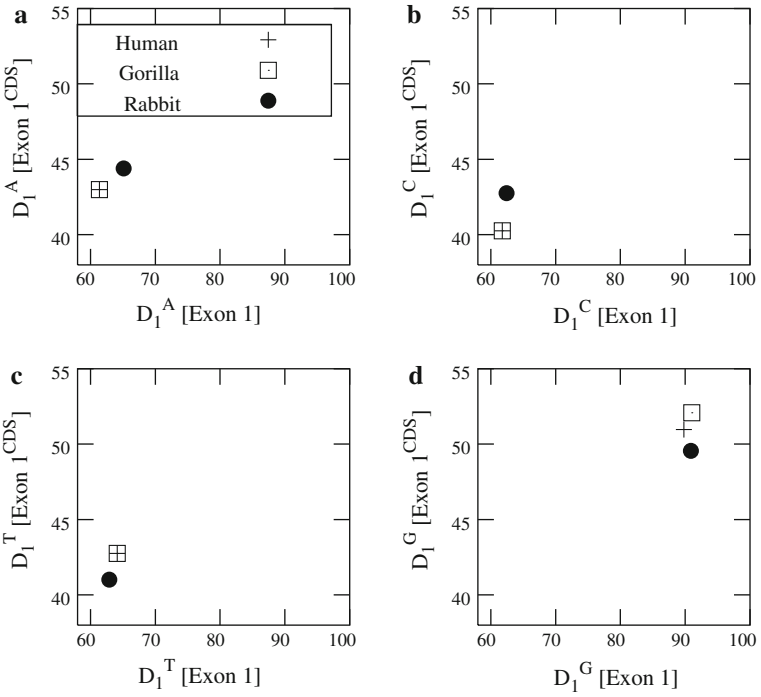

Fig. 30 D_1^{γ}[Exon 1] $-D_1^{\gamma}$[Exon 1^{CDS}] diagrams for the sequences of 3 species listed in Table 23

values for different exons are the same for all the species since most of the points are concentrated close to the dashed lines.

Complete sequences for the first exons are given only for three species (Table 23). The whole sequences of the first exons for human and gorilla differ by only one base. As we see in Fig. 30 this is G base. The descriptors D_1^A, D_1^C, D_1^T are exactly the same for human and gorilla sequences. The difference caused by this single base is recognized by D_1^G (panel d).

Table 19 Locations of sequences with introns (PlusI) in β-globin gene from the EMBL database

No.	Species	ID/Accession	Location in gene	N_{PlusI}	N_{CDS}
1	Homo sapiens (Human)	U01317	62187-63607	1421	441
2	Pan troglodytes (Chimpanzee)	X02345	4189-5531	1343	375
3	Gorilla gorilla (Gorilla)	X61109	4538-5880	1343	363
4	Eulemur macaco (Lemur)	M15734	154-1592	1439	441
5	Rattus norvegicus (Rat)	X06701	310-1502	1193	441
6	Capra hircus (Goat)	M15387	279-1746	1468	435
7	Bos taurus (Bovine)	X00376	278-1738	1461	435
8	Oryctolagus cuniculus (Rabbit)	V00882	277-1416	1140	441
9	Didelphis virginiana (Opossum)	J03643	467-2485	2019	441

Table 20 Locations of the coding sequences of the first exon (Exon 1^{CDS}) in β-globin gene from the EMBL database

No.	Species	Location in gene	N_1	N_1^A	N_1^C	N_1^T	N_1^G
1	Human	62187-62278	92	17	19	20	36
2	Chimpanzee	4189-4293	105	20	20	24	41
3	Gorilla	4538-4630	93	17	19	20	37
4	Lemur	154-245	92	19	15	23	35
5	Rat	310-401	92	20	18	21	33
6	Goat	279-364	86	17	17	17	35
7	Bovine	278-363	86	17	16	18	35
8	Rabbit	277-368	92	18	16	20	38
9	Opossum	467-558	92	21	20	22	29

6 Conclusions

Summarizing, four-component spectral representation has been used for similarity/dissimilarity analysis of histone H4 coding sequences across species (Figs. 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25), of histone H1 coding sequences across species (Figs. 19, 20), and of different parts of β-globin gene across species (Figs. 26, 27, 28, 29, 30). Since many authors use slightly different data for β-globin gene, the locations of different subsequences in this gene and their full description listed in the tables may be helpful for some alternative similarity studies. The numbers of particular bases in all the sequences are also given.

It has been shown that the four-component spectral representation can be used for the classification studies (clusterization of the descriptors representing histones H4 and H1 coding sequences of plants and of vertebrates). Analogous clusterization is also obtained using some descriptors related to 2D-dynamic graphs (Sect. 4). The sensitivity of the four-component spectral representation has also been shown. In particular, a difference between a pair of sequences by only one base can be recognized.

Table 21 Locations of sequences of the second exon (Exon 2^{CDS}) in β-globin gene from the EMBL database

No.	Species	Location in gene	N_2	N_2^A	N_2^C	N_2^T	N_2^G
1	Human	62409-62631	223	44	58	56	65
2	Chimpanzee	4412-4633	222	44	58	56	64
3	Gorilla	4761-4982	222	45	58	56	63
4	Lemur	376-598	223	45	60	57	61
5	Rat	517-739	223	52	57	58	56
6	Goat	493-715	223	50	52	59	62
7	Bovine	492-714	223	49	51	61	62
8	Rabbit	495-717	223	50	55	55	63
9	Opossum	672-894	223	47	54	63	59

Table 22 Locations of the coding sequences of the third exon (Exon 3^{CDS}) in β-globin gene from the EMBL database

No.	Species	Location in gene	N_3	N_3^A	N_3^C	N_3^T	N_3^G
1	Human	63482-63607	126	25	37	29	35
2	Chimpanzee	5484-5531	48	7	14	13	14
3	Gorilla	5833-5880	48	7	13	14	14
4	Lemur	1467-1592	126	20	33	31	42
5	Rat	1377-1502	126	25	36	29	36
6	Goat	1621-1746	126	20	36	30	40
7	Bovine	1613-1738	126	22	32	32	40
8	Rabbit	1291-1416	126	25	33	34	34
9	Opossum	2360-2485	126	25	31	34	36

Table 23 Locations of sequences of the whole first exon (Exon 1) in β-globin gene from the EMBL database

Species	Location in gene	N_{W1}	N_1^A	N_1^C	N_1^T	N_1^G
Human	62137-62278	142	33	35	32	42
Gorilla	4488-4630	143	33	35	32	43
Rabbit	224-368	145	35	31	34	45

Also the approximate location of the difference and the base which is different in the compared sequences can be also determined.

It has been shown that if higher-order terms of similarity measure based on the descriptors of the four-component spectral representation are added and normalized in the same way as in the alignment methods then a convergence to Clustal W results may be obtained. This means that the results obtained with the alignment method may be interpreted as an average of the considered components of the alternative similarity

measures. Calculating an average is always related to some loss of information, i.e. large degree of degeneracy may appear. As we know, this is an inconvenient feature of similarity/dissimilarity analysis. For example, using the alignment methods the two situations

1. AAAA
   AAAA

2. TTTT
   TTTT

cannot be distinguished. Therefore, using the four-component spectral representations one has a chance to decompose the similarity information and remove the degeneracy. Reducing the degeneracy can also be obtained by adding the corrections to the alignment methods related to different aspects of similarity, as it is proposed in Sect. 2 of this work.

It has been shown that each part of β-globin gene demonstrates different similarity relations across species. The relations also change when different aspects of similarity are compared (asymmetry of the gene structure or kurtosis of the distributions). Therefore using different descriptors or different graphical representations the results may be or very often should be contradictory. Different alternative methods describe different aspects of similarity. In particular, most of alternative studies that have been performed for Exon 1^{CDS} of β-globin gene often give contradictory results. For example the similarity value of Exon 1^{CDS} human–goat is larger than human–mouse if the methods described in the works [106,112,126,137,149] are used. The reverse situation i.e. similarity value between the sequences of Exon 1^{CDS} human–goat is smaller than human–mouse if methods taken from [32,33,36,108,110,122,150–152] are applied.

Many authors introducing new graphical representations for beta-globin gene try to avoid considering chimpanzee and gorilla sequences not only because the data are not complete but also because the results are often different than our expectations. We expect the largest similarity for human–chimpanzee sequences. However detailed similarity/dissimilarity analysis of beta-globin gene using four-component spectral representation indicates that this is not true for all parts of the beta-globin gene and for all γ-components of similarity measures. According to the definition of the new measures, $S_\gamma^{3,4}$ becomes smaller if the sequences are more similar. Considering the second exon, I obtain the largest similarity in the case of human–chimpanzee sequences. This means that $S_\gamma^{3,4}$ is the smallest for the two sequences for all γ, and in particular $S_\gamma^{3,4}$=0 for $\gamma = A, C, T$. The difference between the two sequences is only in the distribution of G bases. It is interesting to note that $S_\gamma^{3,4} = 0$ for the second exon, both for C and for T bases, in three cases: human-chimpanzee, human–gorilla and gorilla–chimpanzee sequences. However for other exons, $S_\gamma^{3,4}$ is not always the smallest in the case of human–chimpanzee sequences comparing to human–other species sequences. If the sequence with introns, PlusI, is considered then $S_C^{3,4}$ is the smallest for human–chimpanzee sequences and for $\gamma = A, T, G$, $S_\gamma^{3,4}$ are the smallest for human–gorilla sequences.

Each descriptor may be related to different biological function. Since we are at the beginning of the way of understanding in which contexts particular descriptors may

play the key role, the creation of new alternative methods aiming at similarity/dissimilarity analysis of biological sequences is of particular importance.

References

1. R. Fuchs, Bioinformatics **18**, 505 (2002)
2. H. Herzel, W. Ebeling, A.O. Schmidt, Phys. Rev. E **50**, 5061 (1994)
3. R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **52**, 2939 (1995)
4. W. Li, Comput. Chem. **21**, 257 (1997)
5. J.A. Berger, S.K. Mitra, M. Carli, A. Neri, J. Franklin Inst. **341**, 37 (2004)
6. R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992)
7. A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995)
8. S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **51**, 5084 (1995)
9. M.Y. Azbel, Phys. Rev. Lett. **75**, 168 (1995)
10. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature **356**, 168 (1992)
11. B.D. Silverman, R. Linsker, J. Theor. Biol. **118**, 295 (1986)
12. B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.F. Muzy, A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001)
13. V. Afreixo, C.A.C. Bastos, A.J. Pinho, S.P. Garcia, P.J.S.G. Ferreira, Bioinformatics **25**, 3064 (2009)
14. R. Chenna, H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, J.D. Thompson, Nucleic Acids Res. **31**, 3497 (2003)
15. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. **215**, 403 (1990)
16. S.B. Needleman, C.D. Wunsch, J. Mol. Biol. **48**, 443 (1970)
17. C. Notredame, D.G. Higgins, J. Heringa, J. Mol. Biol. **302**, 205 (2000)
18. R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 1998)
19. M.S. Waterman, *Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics* (Chapman and Hall/CRC, Boca Raton, FL, 1995)
20. S. Vinga, J. Almeida, Bioinformatics **19**, 513 (2003)
21. T.D. Pham, J. Zuegg, Bioinformatics **20**, 3455 (2004)
22. G. Jaklič, T. Pisanski, M. Randić, J. Comput. Biol. **13**, 1558 (2006)
23. B.-H. Zhang, H.-S. Wang, L. Xu, Chemometr. Intell. Lab. Syst. **87**, 194 (2007)
24. W. Chen, B. Liao, Y. Liu, W. Zhu, Z. Su, MATCH Commun. Math. Comput. Chem. **60**, 291 (2008)
25. Y. Zhang, MATCH Commun. Math. Comput. Chem. **60**, 313 (2008)
26. C. Li, X. Yu, N. Helal, Chem. Phys. Lett. **459**, 172 (2008)
27. C. Li, J. Wang, J. Math. Chem. **43**, 26 (2008)
28. W. Chen, Y. Zhang, MATCH Commun. Math. Comput. Chem. **61**, 533 (2009)
29. W. Chen, Y. Zhang, MATCH Commun. Math. Comput. Chem. **61**, 781 (2009)
30. J. Feng, Y. Hu, P. Wan, A. Zhang, W. Zhao, J. Theor. Biol. **266**, 703 (2010)
31. F. Bai, J. Zhang, J. Zheng, Appl. Math. Lett. **24**, 232 (2011)
32. M. Randić, A.T. Balaban, J. Chem. Inf. Comput. Sci. **43**, 532 (2003)
33. R. Chi, K. Ding, Chem. Phys. Lett. **407**, 63 (2005)
34. X.C. Tang, P.P. Zhou, W.Y. Qiu, Chinese Sci. Bull. **55**, 701 (2010)
35. B. Liao, R. Li, W. Zhu, X. Xiang, J. Math. Chem. **42**, 47 (2007)
36. B. Liao, T. Wang, J. Chem. Inf. Comput. Sci. **44**, 1666 (2004)
37. M. Hönl, M.A. Ragan, Syst. Biol. **56**, 206 (2007)
38. B. Liao, M. Tan, K. Ding, Chem. Phys. Lett. **414**, 296 (2005)
39. B. Liao, Chem. Phys. Lett. **401**, 196 (2005)

40. B. Liao, X. Shan, W. Zhu, R. Li, Chem. Phys. Lett. **422**, 282 (2006)
41. B. Liao, X. Xiang, W. Zhu, J. Comput. Chem. **27**, 1196 (2006)
42. S.S. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y. Ho, Nucleic Acid Res. **31**, 3078 (2003)
43. C. Yu, Q. Liang, C. Yin, R.L. He, S.-T. Yau, DNA Res. **17**, 155 (2010)
44. W. Wang, B. Liao, T. Wang, W. Zhu, Int. J. Quantum Chem. **106**, 1998 (2006)
45. H. Wang, Y. Zhang, Int. J. Quantum Chem. **110**, 1964 (2010)
46. F. Bai, W. Zhu, T. Wang, Chem. Phys. Lett. **408**, 258 (2005)
47. M. Randić, D. Plavšić, Chem. Phys. Lett. **476**, 277 (2009)
48. A. Nandy, S.C. Basak, B.D. Gute, J. Chem. Inf. Model. **47**, 945 (2007)
49. A. Ghosh, A. Nandy, P. Nandy, B.D. Gute, S.C. Basak, J. Chem. Inf. Model. **49**, 2627 (2009)
50. A. Ghosh, A. Nandy, P. Nandy, BMC Struct. Biol. **10**, 22 (2010)
51. M. Randić, Chem. Phys. Lett. **440**, 291 (2007)
52. Y. Li, G. Huang, B. Liao, Z. Liu, MATCH Commun. Math. Comput. Chem. **61**, 519 (2009)
53. Y.-H. Yao, Q. Dai, L. Li, X.-Y. Nan, P.-A. He, Y.-Z. Zhang, J. Comput. Chem. **31**, 1045 (2010)
54. P.-A. He, Y.-P. Zhang, Y.-H Yao, Y.-F. Tang, X.-Y. Nan, J. Comput. Chem. **31**, 2136 (2010)
55. A. Bender, R.C. Glen, Org. Biomol. Chem. **2**, 3204 (2004)
56. D. Bielińska-Wąż, P. Wąż, S.C. Basak, Eur. Phys. J. B **50**, 333 (2006)
57. R. Carbó-Dorca, P.G. Mezey (eds.), *Advances in Molecular Similarity, vol. 2* (JAI Press, Stamford, 1998), p. 297
58. D.J. Livingstone, T. Clark, M.G. Ford, B.D. Hudson, D.C. Whitley, SAR QSAR Environ. Res. **19**, 285 (2008)
59. J. Devillers, A.T. Balaban (eds.), *Topological Indices and Related Descriptors in QSAR and QSPR* (Gordon and Breach Science Publishers, The Netherlands, 1999), p. 811
60. S.C. Basak, B.D. Gute, D. Mills, D.M. Hawkins, J. Mol. Struct. (Theochem) **622**, 127 (2003)
61. S.C. Basak, D. Mills, J. Math. Chem. **49**, 185 (2011)
62. D. Bielińska-Wąż, P. Wąż, S.C. Basak, R. Natarajan, in *Symmetry, Spectroscopy and SCHUR*, ed. by R.C. King et al. (Nicolaus University Press, Toruń, 2006), pp. 27–32
63. D. Bielińska-Wąż, P. Wąż, S.C. Basak, J. Math. Chem. **42**, 1003 (2007)
64. G. Aguero-Chapin, H. González-Díaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. González-Díaz, FEBS Lett. **580**, 723 (2006)
65. D. Bielińska-Wąż, J. Math. Chem. **47**, 41 (2010)
66. K. Bhasi, L. Zhang, D. Brazeau, A. Zhang, M. Ramanathan, Bioinformatics **22**, 1569 (2006)
67. C. Yin, S.-T Yau, J. Theor. Biol. **247**, 687 (2007)
68. M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M.F. Lin, M. Kellis, K. Lindblad-Toh, E.S. Lander, PNAS **104**, 19428 (2007)
69. C.T. Zhang, J. Wang, Nucleic Acids Res. **28**, 2804 (2000)
70. J.-F. Yu, X. Sun, J. Comput. Chem. **31**, 2126 (2010)
71. D. Bielińska-Wąż, S. Subramaniam, A new view on similarity of DNA sequences (in preparation).
72. K.K. Mon, J.B. French, Ann. Phys. NY **95**, 90 (1975)
73. T.A. Brody, J. Flores, J.B. French, P.A. Mello, A. Pandey, S.S.M. Wong, Rev. Mod. Phys. **53**, 385 (1981)
74. J.B. French, V.K. Kota, Annual Review of Nuclear and Particle Science, ed. J.D. Jackson, H.E. Gove, R.F. Schwitters (Palo Alto, CA, 1982), p. 35
75. H.A. Bethe, Phys. Rev. **50**, 332 (1936)
76. K.F. Ratcliff, Phys. Rev. C **3**, 117 (1971)
77. M. Bancewicz, G.H.F. Diercksen, J. Karwowski, Phys. Rev. A **40**, 5507 (1989)
78. D. Bielińska-Wąż, N. Flocke, J. Karwowski, Phys. Rev. B **59**, 2676 (1999)
79. M.G. Kendall, *The Advanced Theory of Statistics, vol. 1* (Charles Griffin, London, 1943)
80. V.S. Ivanov, V.B. Sovkov, Opt. Spectrosc. **74**, 30 (1993)
81. V.S. Ivanov, V.B. Sovkov, Opt. Spectrosc. **74**, 52 (1993)
82. D. Bielińska-Wąż, J. Karwowski, Phys. Rev. A **52**, 1067 (1995)
83. D. Bielińska-Wąż, J. Karwowski, J. Quant. Spec. Rad. Transf. **59**, 39 (1998)
84. M. Lax, J. Chem. Phys. **20**, 1752 (1952)
85. C. Bauche-Arnoult, J. Bauche, M. Klapisch, Phys. Rev. A **31**, 2248 (1985)
86. D. Bielińska-Wąż, Symmetry and Structural Properties of Condensed Matter, ed. T. Lulek et al. (World Scientific, Singapore 1999), pp. 212–221.
87. E. Hamori, Nature **314**, 585 (1985)

88. M.A. Gates, Nature **316**, 219 (1985)
89. A. Nandy, Curr. Sci. **66**, 309 (1994)
90. P.M. Leong, S. Morgenthaler, Comput. Appl. Biosci. **11**, 503 (1995)
91. E. Hamori, J. Ruskin, J. Biol. Chem. **258**, 1318 (1983)
92. A. Nandy, Curr. Sci. **66**, 821 (1994)
93. E. Mizraji, L. Ninio, Biochemie **67**, 445 (1985)
94. J.R. Lobry, Biochemie **78**, 323 (1996)
95. X. Guo, M. Randić, S.C. Basak, Chem. Phys. Lett. **350**, 106 (2001)
96. Y. Liu, X. Guo, L. Pan, S. Wang, J. Chem. Inf. Comput. Sci. **42**, 529 (2002)
97. G. Huang, B. Liao, Y. Li, Z. Liu, Chem. Phys. Lett. **462**, 129 (2008)
98. G. Huang, B. Liao, Y. Li, Y. Yu, Biophys. Chem. **143**, 55 (2009)
99. C. Li, J. Wang, Internet Electron. J. Mol. Des. **1**, 000 (2003)
100. D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, Chem. Phys. Lett. **442**, 140 (2007)
101. D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Chem. Phys. Lett. **443**, 408 (2007)
102. Z.-J. Zhang, Bioinformatics **25**, 1112 (2009)
103. Z. Liu, B. Liao, W. Zhu, G. Huang, Int. J. Quantum Chem. **109**, 948 (2009)
104. Z. Liu, B. Liao, W. Zhu, MATCH Commun. Math. Comput. Chem. **61**, 541 (2009)
105. M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. **368**, 1 (2003)
106. M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. **371**, 202 (2003)
107. P.A. Scholes, *The Oxford Companion to Music, 10th ed* (Oxford University Press, Oxford, 1986)
108. C. Li, J. Wang, Comb. Chem. High Throughput Screen. **6**, 795 (2003)
109. J. Song, H. Tang, J. Biochem. Biophys. Methods **63**, 228 (2005)
110. B. Liao, T. Wang, J. Comput. Chem. **25**, 1364 (2004)
111. J. Wang, Y. Zhang, Chem. Phys. Lett. **423**, 50 (2006)
112. Y. Yao, T. Wang, Chem. Phys. Lett. **398**, 318 (2004)
113. M. Randić, Chem. Phys. Lett. **456**, 84 (2008)
114. H.J. Jeffrey, Nucleic Acids Res. **18**, 2163 (1990)
115. H.J. Jeffrey, Comput. Graphics **16**, 25 (1992)
116. M. Randić, M. Vračko, J. Zupan, M. Novič, Chem. Phys. Lett. **373**, 558 (2003)
117. M. Randić, Chem. Phys. Lett. **386**, 468 (2004)
118. M. Randić, N. Lerš, D. Plavšić, S.C. Basak, A.T. Balaban, Chem. Phys. Lett. **407**, 205 (2005)
119. I. Pesek, J. Zerovnik, MATCH Commun. Math. Comput. Chem. **60**, 301 (2008)
120. M. Randić, M. Vračko, A. Nandy, S.C. Basak, J. Chem. Inf. Comp. Sci. **40**, 1235 (2000)
121. C. Li, J. Wang, Comb. Chem. High Throughput Screen. **7**, 23 (2004)
122. Y. Yao, X. Nan, T. Wang, Chem. Phys. Lett. **411**, 248 (2005)
123. C. Yuan, B. Liao, T. Wang, Chem. Phys. Lett. **379**, 412 (2003)
124. B. Liao, T. Wang, J. Mol. Struct. Theochem **681**, 209 (2004)
125. B. Liao, T. Wang, Chem. Phys. Lett. **388**, 195 (2004)
126. B. Liao, Y. Zhang, K. Ding, T.J. Wang, Mol. Struct. **717**, 199 (2005)
127. W. Chen, B. Liao, X. Xiang, W. Zhu, MATCH Commun. Math. Comput. Chem. **61**, 767 (2009)
128. Z. Cao, R. Li, W. Chen, Int. J. Quantum. Chem **110**, 975 (2010)
129. C.-T. Zhang, R. Zhang, H.-Y Ou, Bioinformatics **19**, 593 (2003)
130. Z. Cao, B. Liao, R. Li, Int. J. Quantum. Chem. **108**, 1485 (2008)
131. Z.-H. Qi, T.-R. Fan, Chem. Phys. Lett. **442**, 434 (2007)
132. X.-Q. Qi, J. Wen, Z.-H. Qi, J. Theor. Biol. **249**, 681 (2007)
133. J.-F. Yu, J.-H. Wang, X. Sun, MATCH Commun. Math. Comput. Chem. **63**, 493 (2010)
134. J.-F. Yu, X. Sun, J.-H. Wang, J. Theor. Biol. **261**, 459 (2009)
135. A. Nandy, M. Harle, S.C. Basak, Arkivoc ix (2006) 211.
136. C. Yuan, L. Liu, T. Wang, C. Li, J. Math. Chem. **43**, 1177 (2008)
137. P. He, J. Wang, Internet Electron. J. Mol. Des. **1**, 668 (2002)
138. B. Liao, M. Tan, K. Ding, Chem. Phys. Lett. **414**, 296 (2005)
139. M.A. Gates, J. Theor. Biol. **119**, 319 (1986)
140. C. Raychaudhury, A. Nandy, J. Chem. Inf. Comput. Sci. **39**, 243 (1999)
141. X. Guo, A. Nandy, Chem. Phys. Lett. **369**, 361 (2003)
142. D. Bielińska-Wąż, P. Wąż, W. Nowak, A. Nandy, S.C. Basak, American Institute of Physics (AIP) Conference Proceedings 963 (New York 2007), pp. 28–30.

143. D. Bielińska-Wąż, W. Nowak, Ł. Pepłowski, P. Wąż, S.C. Basak, R. Natarajan, J. Math. Chem. **43**, 1560 (2008)
144. D. Bielińska-Wąż, P. Wąż, J. Math. Chem. **43**, 1287 (2008)
145. Y. Guo, T. Wang, J. Mol. Struct. Theochem. **853**, 62 (2008)
146. D. Bielińska-Wąż, P. Wąż, T. Clark, Chem. Phys. Lett. **445**, 68 (2007)
147. D. Bielińska-Wąż, S. Subramaniam, J. Theor. Biol. **266**, 667 (2010)
148. M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Bioinformatics **23**, 2947 (2007)
149. P. He, J. Wang, J. Chem. Inf. Comput. Sci. **42**, 1080 (2002)
150. M. Randić, M.J. Vračko, J. Chem. Inf. Comput. Sci. **40**, 599 (2000)
151. M. Randić, X. Guo, S.C. Basak, J. Chem. Inf. Comput. Sci. **41**, 619 (2001)
152. Y.-z. Liu, T.-m. Wang, Chem. Phys. Lett. **417**, 173 (2006)