CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **86** (4) (2013) 519–529.
http://dx.doi.org/10.5562/cca2338

Original Scientific Article

On Map Representations of DNA†

Milan Randić,[a,*] Boris Horvat,[b,c,d] Gašper Jaklič,[b] Dejan Plavšić,[e] and Tomaž Pisanski[b,c]

[a]*Laboratory for Chemometrics, National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia*
[b]*Institute of Mathematics, Physics and Mechanics, University of Ljubljana, Slovenia*
[c]*University of Primorska, Koper, Slovenia*
[d]*Abelium, d.o.o. Ljubljana, Slovenia*
[e]*NMR Center, Institute Rudjer Bošković, Bijenička cesta 54, Zagreb, Croatia*

RECEIVED JULY 16, 2013; REVISED DECEMBER 14, 2013; ACCEPTED DECEMBER 16, 2013

Abstract. We have constructed graphical (qualitative and visual) representations of DNA sequences as 2D maps and their numerical (quantitative and computational) analysis. The maps are obtained by transforming the four-letter sequences (where letters represent the four nucleic bases) *via* a spiral representation over triangular and square cells grids into a four-color map. The so constructed maps are then represented by distance matrices. We consider the use of several matrix invariants as DNA descriptors for determining the degree of similarity of a selection of DNA sequences. (doi: 10.5562/cca2338)

Keywords: DNA graphical representation, virtual genetic code, numerical representation, four color map, distance matrix of DNA, DNA similarity, DNA descriptors

INTRODUCTION

Graphical representations of DNA were initiated by Hamori and Ruskin[1–3] over 25 years ago as an alternative viewing of DNA sequences. Graphical representations allow one to visually estimate the degree of similarity or the lack of similarities among lengthy DNA sequences composed of four bases represented by letters A, C, G and T, (which stand for adenine, cytosine, guanine, and thymine, respectively). Hamori and Ruskin have associated with the four nucleotides the four directions in the (x, y) plane along the $\pm x$ axis and $\pm y$ axis, adding a move along z-axis for each step in DNA sequence. When the DNA sequence is plotted one obtains a spatial curve representation of DNA. Thus instead of direct comparison of primary DNA sequences they inspected similarity between their graphical representations of DNA constructed in 3D space. A few years later Jeffrey[4] introduced an alternative graphical representation of DNA, in which each nucleotide was represented as a single dot in the interior of a suitably labeled square. Jeffrey's approach was based on Chaos Game representation of long numerical sequences introduced in mathematics a year before by Barnsley.[5] Jeffrey modified the algorithm of the Chaos Game to suit representations of DNA sequences and thus arrived at geometrical patterns that made of large number of points scattered inside a square, which can be viewed as DNA maps.

The pioneering works of Hamori[1–3] and Jeffrey[4] have opened a new direction in comparative study of DNA in Bioinformatics, which offered novel insights on similarities/dissimilarities of DNA sequences. Admittedly novel visual comparisons of DNA were of qualitative nature, which limited their usefulness. However, this all has changed in 2000 when it has been shown that one can arrive at quantitative 2D comparative studies of graphical representations of DNA by numerical characterizations of graphical representation, which have been viewed or associated with various mathematical objects.[6,7] This step signified the major "breakthrough" for graphical representation of biosequences that grew out of early visual representations of DNA, proteins and RNA, which was also extended to quantitative analyses of proteomics maps, which have hitherto been only visually inspected[8–12] and form the basis branch of bioinformatics which of recently has been referred as Graphical Bioinformatics.[13]

Among the dozen alternative graphical representations of DNA that emerged during the following 20−25 years, we would like to mention, in particular, the spectral representations of bio-sequences. Spectral representations of DNA are obtained by associating with the four bases, A, C, G, T, in case of DNA four horizontal

† Dedicated to Professor Douglas Jay Klein on the occasion of his 70th birthday.
* Author to whom correspondence should be addressed. (E-mail: mrandic@msn.com)

lines assigning to them the numerical values from 1−4, and the case of proteins twenty horizontal lines assigning to them the numerical values 1−20 for the 20 natural amino acids. Then one depicts in sequential order, in the case of DNA the four bases, and in the case of proteins the 20 amino acids, as spots over the corresponding horizontal lines in the x, y plane. By connecting the adjacent spots by lines one obtains graphical representations which are reminiscent of spectra in physics and chemistry.

The first spectral representation of DNA appeared in 2003.[14,15] Spectral representations of DNA and proteins have an important advantage over many other graphical representations in that one can shift sequences and subtract them, and in this way one could detect graphically the degree of alignment between DNA and protein sequences.[16]

Comparative Study of Sequences

After Hamori and Ruskin's 3D graphical representation of DNA[1] several researchers considered simplified 2D graphical representations of DNA that amounts to projections of 3D curve on (x,y) plane, with variations on selection of alternative assignments for the A, C, G, T directions along the x and y axes.[17–19] Such simplified graphical representations of DNA are accompanied with some loss of information and do not allow reconstruction of the original DNA sequence, as was possible in the case of the 3D representations of Hamori and Jeffrey. However, 2D graphical representations offered user-friendly DNA patterns, suitable for visual inspection. We should add that despite this significant deficiency, when several years later quantitative approach to characterization of graphical DNA representations was developed it was possible to recover the lost information by considering walks alone.

In this way numerical matrices representing DNA fully recovered the loss of information. However, any (finite) set of sequence invariants are inherently insufficient to for full reconstruction of DNA. The same is true for representations of molecules by topological indices, physico-chemical, or quantum mechanical descriptors in structure-property and QSAR analyses. Numerous "topological" indices[20–23] that have been successively employed in statistical analyses of chemical data similarly do not generally allow molecular reconstructions.

Therefore, when comparative studies of DNA are based on graph or matrix invariants it is important to have alternative DNA representations as a source for construction of additional DNA descriptors. Generally one can expect that different graphical approaches will encode different structural information and in this way compensate for the inherent partial loss of information accompanying use of graphical and mathematical invariants for representation of biological sequences. In addition use of different descriptors may facilitate to detect artifacts in such studies, which represent noise. This may then help one to identify false-positives and false-negatives, the cases that suggest two DNA sequences to be similar when in fact they are not or the cases that suggest two DNA sequences to be dissimilar when in fact they are similar.

There are other difficulties that have to be considered in comparative study of bio-sequences. Use of widely different descriptors in different studies makes comparisons of results from such sources difficult. What is desirable, but it has not been hitherto available, is to use distinct set of DNA descriptors that are conceptually and structurally related, so that when applied to the same set of data, they will produce different, but structurally related results. In this way one may hope for the same set of DNA sequences to identify and eliminate false positive and false-negative results and allow construction of more reliable similarity dendograms.

In this article we are developing one such graphical approach for DNA, which may help in pointing to false-positive and false-negative results. The approach is based on using related geometrical templates in construction of alternative maps for graphical representation of the same set of DNA sequences. The approach as will be seen also offers user-friendly graphical patterns that facilitate preliminary visual analyses of DNA. The present approach represents an extension and elaboration of the approach that has been presented before by several of the present authors.[24–26] Introduction of 2D DNA maps allows construction of additional numerical sequence descriptors. The complexity of bio-sequences is such that additional novel descriptors in comparative studies of DNA and proteins should be welcome.

Four Color 2D Maps of DNA

Graphical representations of DNA have offered visual inspection of the similarities and the dissimilarities of DNA sequences just as the chemical structural formulas offer visual inspection on the degree of similarities and dissimilarities of chemical structures. But visual comparison can be misleading, thus it is essential to have a quantitative measure of the degree of similarity or lack of similarity of systems considered. During the past dozen years we and our collaborators have been advancing methodologies for quantitative characterizations of DNA,[7,8.16,27–33] RNA,[34,35] proteins,[36–42] and proteomics.[43–52] Some graphical representations are sensitive ton minor and local changes in the sequence composition, which could results in visibly different graphical representations of otherwise similar pairs of DNA or proteins. For example, this is the case with 2D graphical representation of DNA considered by Nandy,[17] Leong and Morgenthaler[18] and Gates.[19]

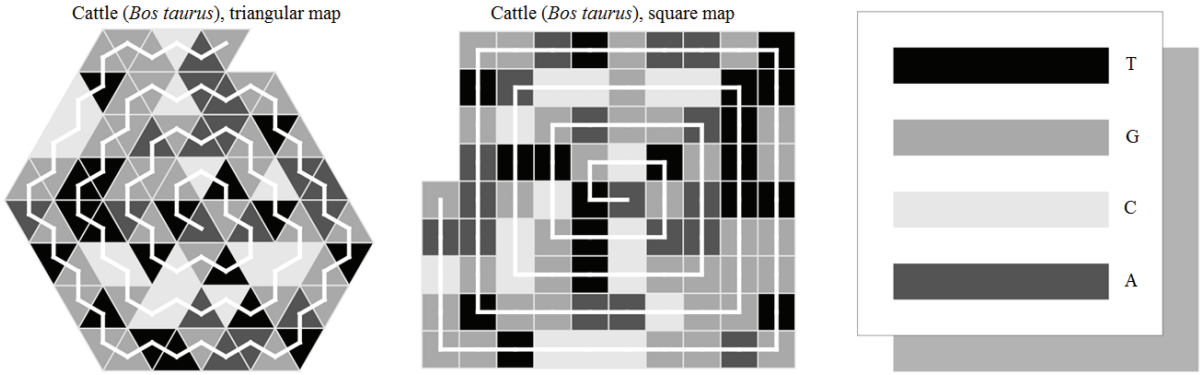

Figure 1. The representations of the first exon of Cattle (*Bos taurus*) *β*-globin gene sequence on triangular and square tessellation grids. The white curves, with the origin in the center of the map, represent the spiral. The four bases are represented by different colors (shades).

The basic idea behind the "Four Color" representation of DNA is to transform a 1D object (DNA sequence) into a 2D object (DNA map). For illustration consider a short DNA sequence of the first exon of *β*-globin gene of cattle (*Bos taurus*), listed at the top of Table 1. By writing the DNA sequence in a spiral form, either using regular triangular or square tessellation in a 2D plane, one obtains four letter labeled triangles or

Table 1. The coding sequences of the first exon of *β*-globin gene of nine species

Cattle (*Bos taurus*)				
ATGCTGACTG	CTGAGGAGAA	GGCTGCCGTC	ACCGCCTTTT	GGGGCAAGGT
GAAAGTGGAT	GAAGTTGGTG	GTGAGGCCCT	GGGCAG	
Chicken (*Gallus gallus*)				
ATGGTGCACT	GGACTGCTGA	GGAGAAGCAG	CTCATCACCG	GCCTCTGGGG
CAAGGTCAAT	GTGGCCGAAT	GTGGGGCCGA	AGCCCTGGCC	AG
Goat (*Capra hircus*)				
ATGCTGACTG	CTGAGGAGAA	GGCTGCCGTC	ACCGGCTTCT	GGGGCAAGGT
GAAAGTGGAT	GAAGTTGGTG	CTGAGGCCCT	GGGCAG	
Gorilla (*Gorilla gorilla*)				
ATGGTGCACC	TGACTCCTGA	GGAGAAGTCT	GCCGTTACTG	CCCTGTGGGG
CAAGGTGAAC	GTGGATGAAG	TTGGTGGTGA	GGCCCTGGGC	AGG
Human (*Homo sapiens*)				
ATGGTGCACC	TGACTCCTGA	GGAGAAGTCT	GCCGTTACTG	CCCTGTGGGG
CAAGGTGAAC	GTGGATGAAG	TTGGTGGTGA	GGCCCTGGGC	AG
Lemur (*Eulemur macaco*)				
ATGACTTTGC	TGAGTGCTGA	GGAGAATGCT	CATGTCACCT	CTCTGTGGGG
CAAGGTGGAT	GTAGAGAAAG	TTGGTGGCGA	GGCCTTGGGC	AG
Opossum (*Didelphis virginiana*)				
ATGGTGCACT	TGACTTCTGA	GGAGAAGAAC	TGCATCACTA	CCATCTGGTC
TAAGGTGCAG	GTTGACCAGA	CTGGTGGTGA	GGCCCTTGGC	AG
Rabbit (*Oryctolagus cuniculus*)				
ATGGTGCATC	TGTCCAGTGA	GGAGAAGTCT	GCGGTCACTG	CCCTGTGGGG
CAAGGTGAAT	GTGGAAGAAG	TTGGTGGTGA	GGCCCTGGGC	
Rat (*Rattus norvegicus*)				
ATGGTGCACC	TAACTGATGC	TGAGAAGGCT	ACTGTTAGTG	GCCTGTGGGG
AAAGGTGAAC	CCTGATAATG	TTGGCGCTGA	GGCCCTGGGC	AG

square cells (Figure 1). When adjacent cells having the same label are fused into a single region one obtains a map in which regions having the same nucleotide base can be colored, each color for one nucleotide. This straightforward construction transforms a 1D (sequential or linear) object into a 2D geometrical object (map).

Once a map has been constructed, one searches for numerical characterizations of the so constructed geometrical map. Clearly if two DNA sequences are very similar, then also the derived 2D maps will be similar. Alternatively, if two 2D maps are widely different then the corresponding DNA sequences will also be very different. Because of loss of information accompanying use of mathematical invariants as descriptors, as a rule, not only it is possible that different sequences may have an identical map, but similar maps may have many the same sets of invariants. Thus similar maps need not correspond to similar sequences. This is one of reasons for considering alternative graphical representations of DNA sequences and construction of different sets of map descriptors. Because graphical representations of DNA and proteins are based on non-aligned sequences it would be desirable to have graphical representations of DNA and proteins that are not very sensitive to relative shifts of two sequences. The four-color 2D map representations appear to be less sensitive to minor substitutions, insertions or deletions in sequences.

On Sequence Alignment Problem

One of the central problems of Bioinformatics is DNA and protein alignments. They have allowed one to arrive at the degree of similarity between different DNA and proteins, based on the number and length of gaps in pair-wise comparisons. Graphical Bioinformatics[32,49] allows one to arrive at measures of similarity-dissimilarity of DNA and proteins without considering DNA or protein alignment problem. Though graphical representations of DNA and proteins allow quantitative comparisons of different sequences without prior alignment we should add that spectral representations of DNA and proteins can also be used for searching for DNA and protein alignments. Thus, this central problem of bioinformatics of determining the degree of sequence alignment is not beyond reach of graphical approaches in bioinformatics. It appears that the potential of graphical representations of bio-sequences for alignment and comparisons has been hitherto overlooked.

Recall that the problem of sequence alignment, which has preoccupied computer scientists for about five decades, resulted in numerous computer packages that continue to be used by DNA and protein scientists. Among the better known and more widely used are: the dynamic program for global alignment, of Needleman and Wunsch, reported in 1970,[53] a general method applicable to search for similarities in the amino acid sequences of two proteins; the program for local alignment, that is identification of common molecular subsequences reported in 1981 by Smith and Waterman;[54] the program on rapid and sensitive protein similarity searches of D. J. Lipman and Pearson in 1985,[55] and the report on improved tools for biological sequence comparison in 1988.[56] Finally in 1990 came BLAST (Basic Local Alignment Search Tool) of Altschul *et al*.,[57] one of the most widely used computer program in Bioinformatics, considered as the standard algorithm for similarity analysis. With its up-dated follow up: "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs"[58] of 1997 the two algorithms were cited 80,000 times in Web of Science, together with annual citation around 5,000.

There is no doubt that most, if not all, that we know today in Bioinformatics, is due to availability of current very powerful and very useful computer programs. But that does not mean that further improvements are not possible or not likely! Very recently, at least a 45 years old problem of protein sequence alignment, for which many have believed that it cannot be solved mathematically exactly, has been solved exactly. That means without use of approximations, such as empirical parameters, statistical information, penalties for gaps, insertions and deletions, and of course, without use of trial-and-error methodology, all abundantly used in most if not all current computer programs for protein and DNA alignments. The article was entitled: "Very Efficient Search for Protein Alignment - VESPA",[59] rather than "The Exact Solution for Protein Alignment," because besides being an exact solution the algorithm is also highly efficient. What this algorithm does for a given two protein sequences is a list of labels of amino acids in two sequences that match when two sequences are not shifted, or shifted relative to one towards the other by ±1, ±2, ±3, *etc*., steps. From such information it is easy to find locations and lengths of gaps and construct alignments.

The title of the paper on exact solution of the protein alignment problem has emphasized the high efficiency of the algorithm, rather than its exactness, in view that sometimes exact solutions of problems may be lengthy, may be cumbersome, and may take more time. For example, the exact solution of the inverse problems of X-ray diffraction, the phase determination problem, solved by Hauptman and Karle,[60,61] was believed by many not to be possible to solve exactly. Hauptman and Karle reported their solution, however, it does involve heavy calculations. It took Hauptman and Karle one month to solve exactly a single crystal structure having several heavy atoms in its unit cell, that required Fourier analyses of about 6000 diffraction spots, which as Hauptman writes would take less than three minutes on computer.[62] In contrast the VESPA algorithm for exact

solution of the protein alignment of a pair of proteins having about 160 amino acids, takes about 15 minutes by using only pen and pencil (no calculator or computer). The high efficiency of the exact solution is due to the fact that the algorithm identifies also pairs of adjacent amino which appear only in one sequence and not in the other, and thus eliminate need for their further examination. To find more how was the exact solution in of protein alignment more recently extended to DNA alignment consult Ref. 63.

Let us add that there is an important distinction between computer-based alignment searches and graphical approach to the same problem. It takes at least two sequences to make computer-based comparisons. In contrast, in the case of graphical representations of DNA and proteins, one can characterize a single biological sequence! This is a very important distinction between graphical representations of bio-sequences and computer-based analyses of sequences, which one should keep in mind when discussing graphical and computer-oriented methodologies for sequence comparisons. Graphical Bioinformatics thus allows one to compile a catalogue on individual proteins and DNA, listing their various graphical (mathematical) which would allow fast search for similar sequences in the catalogue.

Four Color 2D Maps of DNA of *Bos Taurus*

The 2D maps shown in Figure 1, belonging to the first exon of β-globin gene of Cattle (*Bos taurus*), obtained by representing the DNA sequence as a spiral over regular triangle and square grids. Observe that the map regions are of different size and different shape. This is clearly seen in Figure 2, which shows the corresponding dual graphs of the maps obtaining by first replacing each map region by a vertex placed in its center and then connecting the corresponding adjacent region by edges. The dual graphs, in contrast to ordinary graphs, have a definite geometry by being embedded in the (x, y) plane where all vertices have fixed (x, y) coordinates and edges have fixed lengths. In the lower part of Figure 2 have we added the labels A, C, G, T that identify individual regions belonging to different nucleotides.

The DNA dual graphs of Figure 2 can be numerically represented by the Euclidean distance matrix. By combining the entries of the Euclidean distance matrix with the corresponding entries of the graph-theoretical distance matrix one obtains the D/D distance matrix,[64] the matrix elements of which are given by the quotient of the corresponding Euclidean and graph theoretical distances.

It is not widely known that "squared" Euclidean matrix, which is the Euclidean matrix where the elements are squared, has some interesting properties

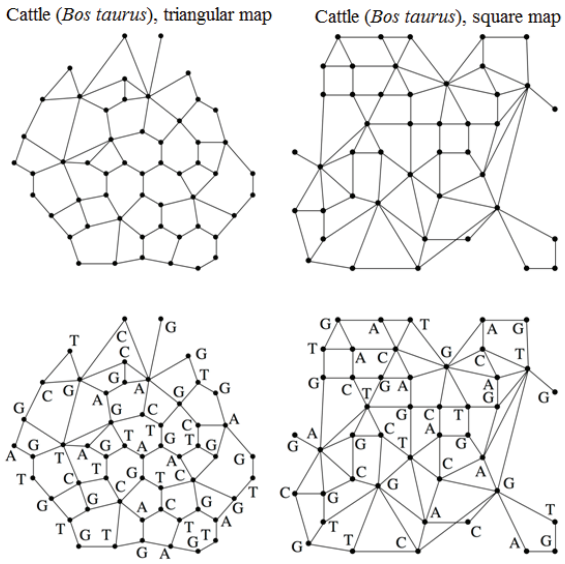

Figure 2. The dual graphs of the constructed maps obtained by fusing neighboring cells having the same label (color) into single region represented by a vertex positioned in the barycenter of the color region and connecting neighboring regions. In the lower part of the figure we identified individual vertices with corresponding nucleotides.

which allows comparison of maps of different size.[65–67] Hence, using invariants of "squared" Euclidean matrices makes possible to compare DNA sequences of different length. In Supplementary material we have listed the eigenvalues of the graph-theoretical distance matrices for the two dual graphs of Figure 2, representing the first exon of β-globin gene of cattle (*Bos taurus*) embedded over the triangular and the square maps, respectively. There one can see that the triangular based network leads to a 54 × 54 matrix, while the square network leads to a 45 × 45 matrix. Because matrices are of different size, the eigenvalues of the two matrices cannot be easily directly compared.

One possibility of using eigenvalues for comparison of graphs or networks of different size is to focus attention on the largest positive and negative eigenvalues only. As one can see from Table 2, the leading eigenvalue (λ_1) for both types of networks is of considerably larger magnitude than the remaining eigenvalues, except perhaps for several of largest negative eigenvalues. Hence, if one is to use eigenvalues for characterization of DNA one may try to use the eigenvalues of the largest absolute magnitudes and construct

Table 2. The four eigenvalue of the largest absolute magnitudes for *Bos taurus*

Cattle (*Bos taurus*)	λ_1	λ_{-3}	λ_{-2}	λ_{-1}
Triangle	165.84	−14.53	−42.57	−49.33
Square	213.68	−19.48	−58.08	−60.83

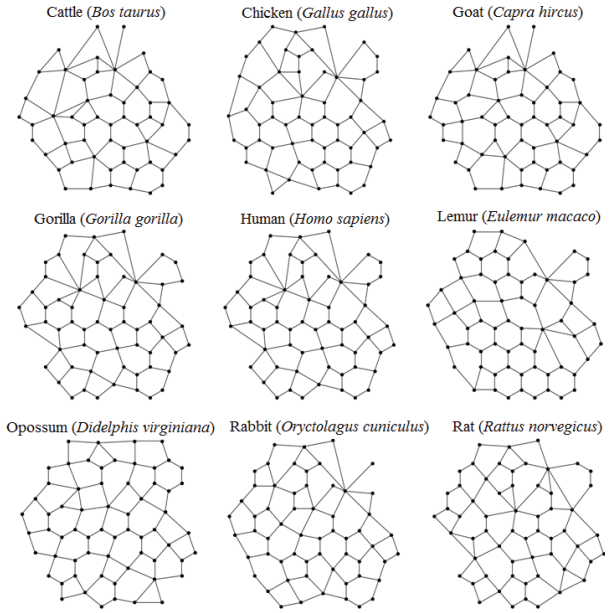

Figure 3. The dual graphs of triangular maps of the first exon of the β-globin for nine species.

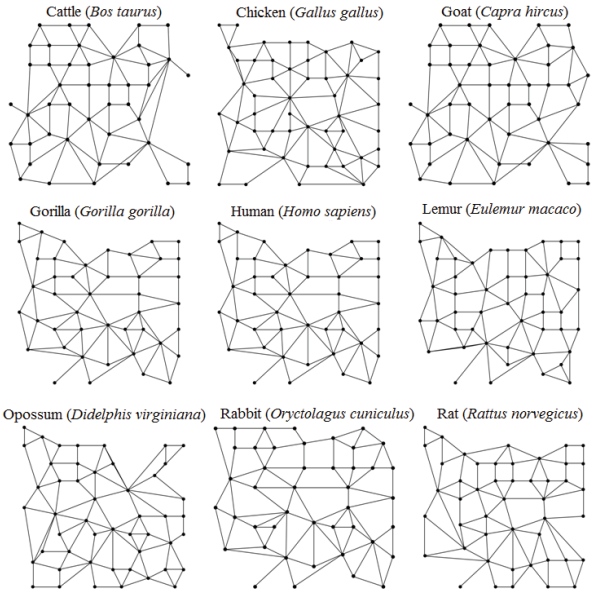

Figure 4. The dual graphs of square maps of the first exon of the β-globin for nine species.

"reduced" vectors to be used for characterization of DNA. In this way the reduced version of eigenvalue representation of *Bos taurus* networks involves only four eigenvalues shown in Table 2. Admittedly, this is an approximate route to comparison of matrices of different size that avoids impasse when comparing maps characterized by different number of eigenvalues.

Dual Maps for Nine Different Species

We will now consider dual maps for the nine species listed in Table 1 based on use of the regular triangular and square grids. We have selected the first exon of β-globin gene of these nine species to illustrate the approach. In Figure 3 and Figure 4 we have collected the resulting dual maps for the DNA sequences of Table 1. Before continuing let us examine Figure 3 and Figure 4 more closely and try visually to detect pairs of species that show greater similarity, as well as, the pairs that show lack of similarity (based on the first exon of β-globin gene). Later we will compare these qualitative estimates with the quantitative estimates based on the characterization of the maps by selected DNA similarity approaches.

Because the DNA sequences considered are of different length, and in addition the four-color map of DNA having the same lengths can lead to matrices of different size, we will first consider the approximate characterization of DNA maps based on 4-component vectors involving the largest positive and three negative eigenvalues of the distance matrix shown in Table 2. It is clear from Figure 3 and Figure 4 that gorilla and human are by far more similar to one another than both are to cattle. In Table 3 are shown the largest (by absolute values) eigenvalues for the three species. The upper part of Table 3 shows the eigenvalues based on graph-theoretical distances. In the lower part of Table 3 are the eigenvalues based on the Euclidean distance separating vertices.

Observe that in the first case human and gorilla have identical eigenvalues (and this is true for all eigenvalues, not only the four shown), which happens because the corresponding (dual) graphs are identical, even though the embedded graphs differ slightly, due to the presence of an additional base on the end of the DNA sequence of gorilla. As one can see the geometry-based distance matrices carry more information than the graph-based distance matrices, but both approaches

Table 3. The leading eigenvalue and largest negative eigenvalues for the three species considered

	Square grid - graph theoretical distance				Square grid - geometric distance			
	λ_1	λ_{-3}	λ_{-2}	λ_{-1}	λ_1	λ_{-3}	λ_{-2}	λ_{-1}
Cattle (*Bos taurus*)	155.27	−15.227	−38.092	−46.089	213.676	−19.4754	−58.0783	−60.8261
Gorilla (*Gorilla gorilla*)	157.18	−12.727	−39.386	−43.507	223.794	−18.7731	−53.9635	−73.6406
Human (*Homo sapiens*)	157.18	−12.727	−39.386	−43.507	224.06	−18.9597	−53.8893	−73.8855

Table 4. The absolute magnitudes of four non-zero eigenvalues of the Euclidean distance matrix of nine species, the elements of which have been squared

	λ_1	λ_2	λ_3	λ_4
Cattle (*Bos taurus*)				
Triangle	655.83	328.97	268.61	58.25
Square	1327.09	620.33	582.74	124.02
Chicken (*Gallus gallus*)				
Triangle	763.34	392.77	309.42	61.15
Square	1682.70	872.92	645.76	164.02
Goat (*Capra hircus*)				
Triangle	680.99	348.52	273.83	58.651
Square	1452.91	728.58	597.91	126.42
Gorilla (*Gorilla gorilla*)				
Triangle	793.10	408.68	322.02	62.41
Square	1475.42	800.81	544.22	130.39
Human (*Homo sapiens*)				
Triangle	790.84	408.26	320.48	62.11
Square	1479.83	805.21	543.38	131.23
Lemur (*Eulemur macaco*)				
Triangle	809.12	430.96	314.73	63.43
Square	1561.82	819.60	595.66	146.55
Opossum (*Didelphis virginiana*)				
Triangle	846.47	407.23	376.56	62.68
Square	1714.75	915.68	656.37	142.70
Rabbit (*Oryctolagus cuniculus*)				
Triangle	751.01	411.76	297.09	60.15
Square	1339.35	719.67	514.05	105.63
Rat (*Rattus norvegicus*)				
Triangle	752.078	381.784	312.582	57.7113
Square	1563.6	802.362	612.397	148.843

have their advantages. If one is to screen large number of sequences, the use of graph-based distance matrices is faster and can serve for pre-screening data. The geometry-based distance matrices, which carry more information, can be used later on subset of matrices for increased differentiation among DNA sequences of higher similarity.

More on Comparison of Maps of Different Size

Instead of using truncated eigenvalue data as outlined above one can alternatively use powers of the distance matrix, which allow an exact comparison of non-zero eigenvalues of matrices of different size. According to already mentioned interesting and intriguing theorem of Linear Algebra when the individual elements of distance matrix (of distances between points lying in general position) are squared, there are only four eigenvalues different from zero, of these one is positive and three are negative.[65] This theorem is a special case of a more general property of Euclidean distance matrices, when their individual matrix elements are raised to higher powers. When entries of the Euclidean distance matrix are raised to the second, third, fourth and fifths powers *etc*., the number of non-zero eigenvalues of such matrices are: 4, 9, 16, 25, *etc*., respectively. Moreover, when one considers Euclidean matrices in 3D, 4D, 5D, *etc*., then for the case of the squared matrix elements one finds not four but five, six, and seven non-zero eigenvalues, respectively, one of which is always positive while all the other are negative. These properties of the Euclidean distance matrix elements can be even combined, leading to generalizations which may be of considerable interest for comparative studies of maps. Here we will consider only the squared Euclidean distance matrix of the dual maps illustrated in Figure 3 and Figure 4.

Squared Euclidean Distance Matrices

In Table 4 we have collected for the nine species of Table 1 the non-zero eigenvalues belonging to a distance matrix, the elements of which are the squared Euclidean distance for dual graphs. The graphs were derived for the triangular network and the square network as the template for construction of the corresponding DNA spirals. Some overall characteristics are apparent. The magnitudes of the eigenvalues are smaller for the triangular case and larger for the square case. One can also notice that the variations of the leading eigenvalues (λ_1) among the species are considerable. For instance, in the case of square maps the λ_1 values are in the range 1327–1715. Observe also that there is some parallelism between the leading eigenvalues

Table 5. Comparison of the exact and the approximate extreme eigenvalues of squared distance matrices for three species

	Exact				Approximate			
	λ_1	λ_{-3}	λ_{-2}	λ_{-1}	λ_1	λ_{-3}	λ_{-2}	λ_{-1}
Bos taurus	1327.09	−124.02	−582.74	−620.33	213.68	−19.48	−58.08	−60.83
Homo sapiens	1479.83	−131.23	−543.38	−805.21	224.06	−18.96	−53.89	−73.89
Gorilla gorilla	1475.42	−130.39	−544.22	−800.81	223.794	−18.78	−53.96	−73.64

Table 6. The similarity/dissimilarity table for the nine species with spiral representations of DNA based on triangular grid

Triangle grid	Cattle	Chicken	Goat	Gorilla	Human	Lemur	Opossum	Rabbit	Rat
Cattle	0	47.2	11.6	60.1	59.2	68.2	83.5	45.4	42.5
Chicken		0	35.9	12.9	12.0	21.5	38.7	13.6	5.9
Goat			0	48.9	47.9	56.7	73.0	33.9	31.4
Gorilla				0	1.0	10.2	27.4	21.6	18.0
Human					0	10.7	28.4	20.7	17.2
Lemur						0	27.3	25.5	27.1
Opossum							0	49.0	42.0
Rabbit								0	16.1
Rat									0

associated with different grids. Thus the smallest leading eigenvalues occur for cattle (*Bos taurus*) and the largest for opossum (*Didelphis virginiana*), regardless of the selection of the grid used for construction of the 2D map.

In Table 5 we show the leading eigenvalues for cattle, gorilla and human of the squared Euclidean distance matrix for square grid representation of the first exon of DNA using graph theoretical distances and Euclidean distances for elements of the distance matrix. A comparison of the exact eigenvalues with the truncated values, which represent an approximate characterization of DNA, shows considerably different values. Nevertheless the relative values are proportional. However, when one compares the characterization of different species using the exact eigenvalues and the corresponding truncated spectral entries, a close look reveals minor deviations. For example, while for the exact eigenvalues for the relative magnitudes for all eigenvalues holds: $|\lambda_i(\text{human})| > |\lambda_i(\text{gorilla})|$, the same is not the case for all approximate eigenvalues.

Similarities and Dissimilarities Among the DNA Sequences

Different graphical representation will generally involve different artifacts in the analysis, and will be accompanied by different signal-to-noise ratios, for different segments of DNA sequences. In this respect graphical representations that do not depend on assignment of the four bases to the four non-equivalent geometrical elements of the underlying geometrical template for construction of graphical representations have some advantage. It is difficult to satisfy such condition with 2D graphical representations, while such can be constructed 3D, as illustrated by Randić, Zupan and Balaban,[66] at an obvious cost of less clear visual properties of such representations. The way to reduce the chance of including false similarity results when comparing different DNA sequences is to use several 2D graphical representations simultaneously. One may expect that "accidental" coincidences in the similarity-dissimilarity testing will be then more likely to be discerned. The combined information may detect which results of the relative similarities of DNA sequences are reliable.

In Table 6 and Table 7 we have collected similarity and dissimilarity for the nine species based on comparison of the first exon of their β-globin gene. Table 6 is based on a graphical representation of DNA using the triangle network as the background for construction of the spiral which leads to the 2D DNA four color map. Table 7 corresponds to the same construction based on the underlying square Cartesian grid. All the similarity/dissimilarity values have been normalized so that the smallest entry in each table is equal to 1. This occurs for

Table 7. The similarity/dissimilarity table for the nine species with spiral representations of DNA based on square grid

Square grid	Cattle	Chicken	Goat	Gorilla	Human	Lemur	Opossum	Rabbit	Rat
Cattle	0.00	69.79	26.29	37.35	38.34	48.73	77.78	19.36	47.46
Chicken		0.00	43.86	38.50	37.75	22.45	9.23	63.50	22.58
Goat			0.00	14.64	15.45	22.61	51.64	22.55	21.40
Gorilla				0.00	1.00	16.33	45.49	25.73	17.82
Human					0.00	15.69	44.68	26.65	17.35
Lemur						0.00	30.05	41.07	3.82
Opossum							0.00	70.70	30.60
Rabbit								0.00	41.32
Rat									0.00

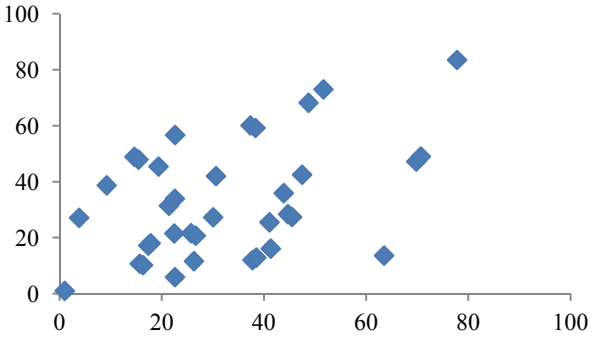

Figure 5. Matrix entries of the "triangle" (*y*-coordinate) and "square" (*x*-coordinate) similarities (Table 5) plotted one against the other.

the pairs human-gorilla for maps based on both grids.

In Figure 5 we have plotted the corresponding entries of the "triangle" and "square" similarity tables one against the other. If the two tables would have the same information all the points should lie on the diagonal line with the slope $m = 1$. As one can see there is a considerable scatter of points in Figure 5, which indicates that the four-color maps do not fully parallel each other. Thus, if either Table 6 or Table 7 is viewed in isolation they will pass some incorrect information. It is good that the four-color maps carry different information and are sensitive to minor variations in DNA sequences. In order to extract meaningful information from such 2D representations of DNA we have to view both similarity tables simultaneously. If we consider the five smallest entries in each table we find the following pairs as the most similar:

- Triangle grid: gorilla-human; chicken-rat; gorilla-lemur; and human-lemur.
- Square grid: gorilla-human; lemur-rat; and chicken-opossum.

Observe that the only pair that we find in both cases is: gorilla-human. Hence, we can be certain that indeed there is a considerable similarity in the first exon of the β-globin gene of gorilla and human.

One should be suspicious about the similarity of the pairs of DNA sequences that appear only once, *i. e.*, in one of the two tables and of apparent similarity of more "distant" species in the evolution trees. However a recent very ambitious comparative study on various species that was extended to whole chromosomes has shown that indeed there could be considerable similarity in some DNA data between two relatively distant species.

The Table 6 and Table 7 can also be used to identify the least similar species by searching for the largest entries. Among the five largest entries one finds:

- Triangle: cattle-opossum; goat-opossum; and cattle-lemur.
- Square: cattle-opossum; rabbit-opossum; cattle-chicken; chicken-rabbit and goat-opossum.

From the above one can suspect that cattle-opossum and goat-opossum, which appear in both similarity / dissimilarity tables, are indeed among the least similar. The above indicates that graphical approaches to similarity of DNA, as any other comparative studies using mathematical invariants as descriptors (which also extends to use of topological indices as descriptors for molecules) while possibly giving false positives (indicating pairs as similar that need not be) are not giving false negatives (indicating as similar cases that are not).

Let us return to Figure 5 and identify the points that are close to the diagonal, namely with close being within ±20. In Table 6 and Table 7 one can identified those points and observe that these points vary in their magnitudes from 1 (the smallest entry for human-gorilla) to 83.5 (the maximal value for cattle-opossum). Regardless of their magnitudes all the points in Table 6 and Table 7 close to the diagonal appear to be beyond doubt. Small entries in both tables, such as chicken-gorilla, chicken-human, chicken-rabbit, and chicken-rat in Table 6 are clearly all false-positive of Table 6. Similarly the cases of chicken-opossum, and lemur-rat in Table 7 are clearly all false-positives of Table 7.

Comparison with Alternative Similarity Analyses

It is of interest to see how the similarity/dissimilarity Table 6 and Table 7 compare with other similarity/dissimilarity tables. We have selected one such table, Table 8, which could be viewed as a kind of standard tables for such comparison. Table 8, which was kindly supplied to us by one of the reviewers of the manuscript, is derived using Clustal Omega.[68,69] It basically determines similarities between DNA sequences based on the number and size of gaps after DNA sequences have been aligned. The first thing to notice is that although qualitatively our two tables (Table 6 and Table 7) parallel Table 8, they also show differences. This is not alarming, because all these tables measure similarities with respect to distinct properties of DNA sequences. When one compares the similarities of Table 6 and Table 7 with the "standards" of Table 8 one can see some agreements but also some disagreements, particularly when one considers less similar case. However, significant and more important is that both approaches agree in identifying the most similar DNA sequences. Actually, in the case of our graphical approach one can identify the most similar pairs of DNA directly from Figure 3 and Figure 4, without using Clustal Omega program,[68] or our numerical similarity values. In short, we may conclude that the four-color model appears to hold well for pairs of DNA which are very similar, but the four-color map approach without further modifications is not suitable for characterization of the degree of similarities of less similar systems, without further

Table 8. Similarity based on Clustal Omega[67]

	Cattle	Chicken	Goat	Gorilla	Human	Lemur	Opossum	Rabbit	Rat
Cattle	0	25.6	3.5	8.1	8.1	18.6	30.6	13.1	24.4
Chicken		0	22.1	26.1	26.1	35.9	28.6	30.0	35.9
Goat			0	10.5	10.5	20.9	31.8	15.5	22.1
Gorilla				0	0.0	26.1	28.6	10.0	19.6
Human					0	26.1	28.6	10.0	19.6
Lemur						0	42.9	27.8	34.8
Opossum							0	33.7	40.7
Rabbit								0	
Rat									0

modification. One such modification would be to base comparison of the four-color maps on construction of maps for pair-wise aligned DNA sequences. It does not seem that such modifications will be difficult to construct in view that DNA alignment problem has been very recently exactly solved.[63]

CONCLUSION

Graphical representation of DNA as a 2D map is a novelty that has only recently been considered. It has an apparent advantage of compressing information on DNA to a relatively small space, and significantly, there is no loss of information in such representations, because DNA sequence can be reconstructed from their graphical four color map image. As we have seen some such representations of DNA have considerable global sensitivity to minor perturbations in a sequence, though locally may be less sensitive. It is therefore useful to consider simultaneously at least a pair of such representations of DNA and combining their information in order to identify false positive and false negative entries in the similarity tables. In view of existence of a number of alternative graphical representations of proteins a question can be posed whether the four color representation of DNA has some special features to be of interest and competitive, and be of potential interest in problems of biology.

The four color representation of DNA has been initially introduced in 2005. Very recently we came across a publication report by G, Agüero-Chapin *et al*.,[70] illustrating one biological application of the four color representation of DNA. These authors explored the adenylation domain repertoire of non ribosomal peptide synthetases using an ensemble of sequence-search methods, which definitely illustrates use of the four-color DNA maps for considering problems of Biology. We may also add recent work of Z. Zhang and collaborator on visualization of DNA sequences,[71] in which they have modified our four-color DNA maps to five-color maps of square shape, which allow them to replace maps by the square matrices and use matrix properties in comparative study of map similarities.

We should also mention that besides use of gap mismatches and Euclidian distance as measures of similarity there are numerous additional similarity indices that have been used as alternative that will produce additional similarity scales. In the case of DNA, which can be represented as binary sequences, according to Consonni and Todeschini[72] there have been to that date more than 50 different similarity coefficients, to which these authors added five new ones. Hence, no single similarity table, be it the Clustal Omega approach, or the Four-color map approach, has claims to be "the solution" when it comes to similarity studies of biological objects. Clustal Omega approach[67] may have here an advantage because it compares bio-sequences that have been aligned. However, in view that the protein alignment problem and the related DNA alignment problems finally have been exactly solved using a very efficient algorithm, this open a novel direction in protein and DNA similarity studies. Besides (*i*) the computer based alignment search and similarity analyses, and (*ii*) sequence non-aligned based graphical approaches, one can expect rise of (*iii*) sequence aligned based graphical approaches, which, just as sequence non-aligned based graphical approaches can capture diverse sequence properties, besides gap-based similarity measures.

Acknowledgements. MR wishes to thank the Laboratory of Chemometrics at the National Institute of Chemistry for cordial hospitality. This work has been supported in part by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia under research grant P1-0017, P1-0294, N1-001 and N1-0012. We thank Professor A. T. Balaban (Texas A&M University at Galveston, TX) for his comments on the manuscript and both referees for useful suggestions that improved the presentation of our results. Referee # 2 was kind to send us the Clustal Omega[67] similarity table (Table 8) for comparison. We also thank Dr. Bono Lučić, the Guest Editor, for his patience and help with the manuscript references.

REFERENCES

1. E. Hamori and J. J. Ruskin, *J. Biol. Chem.* **258** (1983) 1318–1327.
2. E. Hamori, *Nature* **314** (1985) 585–586.
3. E. Hamori, *BioTechniques* **7** (1989) 710–715.
4. H. J. Jeffrey, *Nucleic Acid Res.* **18** (1990) 2163–2170.
5. M. F. Barnsley and H. Rising, *Fractals Everywhere*, 2nd Ed. Academic Press, Boston 1993.
6. M. Randić, M. Vračko, A. Nandy, and S. C. Basak, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.
7. M. Randić and M. Vračko, *J. Chem. Inf. Comput. Sci.* **40** (2000) 599–606.
8. M. Randić, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1330–1338.
9. M. Randić, F. Witzmann, M. Vračko, and S. C. Basak, *Med. Chem. Res*. **10** (2001) 456–479.
10. M. Randić, J. Zupan, and M. Novič, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1339–1344.
11. M. Randić, *Int. J. Quantum Chem*. **90** (2002) 848–858.
12. M. Randić, in: *Handbook of Proteomics Methods* Conn PM. Ed. Humana Press Inc: Totowa, NJ., 2003, pp. 429–450.
13. M. Randić, M. Novič, and D. Plavšić, *Int. J. Quantum Chem*. **113** (2013) 2413–2446.
14. M. Randić, M. Vračko, N. Lerš, and D. Plavšić, *Chem. Phys. Lett.* **371** (2003) 1–6.
15. M. Randić, M. Vračko, N. Lerš, and D. Plavšić, *Chem. Phys. Lett.* **371** (2003) 202–207.
16. M. Randić, J. Zupan, D. Vikić-Topić, and D. Plavšić, *Chem. Phys. Lett.* **431** (2006) 375–379.
17. A. Nandy, *Curr. Sci.* **66** (1994) 309–314.
18. P. M. Leong and S. Morgenthaler, *Comput. Appl. Biosci*. **11** (1995) 503–507.
19. M. A. Gates, *J. Theor. Biol*. **119** (1986) 319–328.
20. J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in Qsar and Qspr*, Gordon and Breach, Reading, U. K., 1999.
21. J. Devillers, *Comparative Qsar*, Francis & Taylor, Washingon, D. C., 1998.
22. O. Ivanciuc, in: *Handbook of Chemoinformatic,*. Gasteiger, Editor, Wiley-VCH, Weinheim, 2003, Vol. 3. pp 981–1003.
23. R. Todeschini and V. Consonni, in: *Handbook of Chemoinformatics* (J. Gasteiger. Editor. Wiley-VCH. Weinheim. 2003.Vo. l. 3. pp 1003–1033.
24. M. Randić, *Chem. Phys. Lett.* **386** (2004) 468–471.
25. M. Randić, N. Lerš, D. Plavšić, S. C. Basak, and A. T. Balaban, *Chem. Phys. Lett.* **407** (2005) 205–208.
26. M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić-Topić, and D. Plavšić, *J. Mol. Graph. Model.* **27** (2009) 637–641.
27. M. Randić, J. Zupan, and T. Pisanski, *J. Math. Chem*. **43** (2008) 624–692.
28. M. Randić and J. Zupan, *SAR & QSAR in Environ. Res*. **15** (2004) 191–205.
29. M. Randić, *Chem. Phys. Lett*. **317** (2003) 29–34.
30. M. Randić, M. Vračko, J. Zupan, and M. Novič, *Chem. Phys. Lett*. **373** (2003) 558–562.
31. X. F. Guo, M. Randić, and S. C. Basak, *Chem. Phys. Lett*. **350** (2001) 106–112.
32. M. Randić, *J. Chem. Inf. Comput. Sci*. **40** (2000) 50–56.
33. M. Randić, M. Vračko, M. Novič, and D. Plavšić, *Int. J. Quantum Chem*. **109** (2009) 2982–2995.
34. M. Randić and D. Plavšić, *Chem. Phys. Lett*. **476** (2009) 277–278.
35. M. Randić, M. Novič, A. R. Choudhury, and D. Plavšić, *SAR QSAR Environ Res*. **23** (2012) 327–343.
36. M. Randić, J. Zupan, A. T. Balaban, D. Vikić-Topić, and D. Plavšić, *Chem. Rev.* **11** (2011) 790–862; in particular pp. 809–811.
37. M. Randić, M. Novič, and M. Vračko, *SAR & QSAR in Environ. Res*. **19** (2008) 339–349.
38. M. Randić, J. Zupan, and D. Vikić-Topić, *J. Mol. Graphics & Modelling* **26** (2007) 290–305.
39. M. Randić, *Chem. Phys. Lett.* **444** (2007) 176–180.
40. M. Randić, D. Butina, and J. Zupan, *Chem. Phys. Lett*. **419** (2006) 528–532.
41. M. Randić, *SAR & QSAR in Environ. Res*. **15** (2004) 147–157.
42. M. Randić and R. Orel. *J. Math. Chem.,* in press.
43. R. Orel and M. Randić. *J. Math. Chem*. **50** (2012) 2689–2702.
44. M. Randić and R. Orel. *J. Math. Chem*. **49** (2011) 1759–1768.
45. M. Randić, M. Novič, M. Vračko, and D. Plavšić, *J. Theor. Biol.* **266** (2010) 21–28.
46. M. Randić, *J. Proteome Res.* **5** (2006) 1575–1579.
47. M. Randić, F. A. Witzmann, V. Kodali, and S. C. Basak, *J. Chem. Inf. Model*. **46** (2006) 116–122.
48. M. Randić and E. Estrada, *J. Proteome Res*. **4** (2005) 2133–2136.
49. M. Randić, N. Novič, and M. Vračko, *J. Chem. Inf. and Modeling* **45** (2005) 1205–1213.
50. M. Randić, in: *Handbook of Proteomics Methods,* P. M. Conn (ed.). Humana Press. Inc. Totowa, NJ, 2003, pp. 429–450.
51. M. Randić, M. Novič, and M. Vračko, *J. Proteome Res.* **1** (2002) 217–226.
52. M. Randić, *J. Math. Chem*. **43** (2007) 756–772.
53. S. Needleman and C. D. Wunsch, *J. Molecular Biol*. **48** (1970) 443–453.
54. T. F. Smith and M. S. Waterman, *J. Molecular Biol*. **147** (1981) 195–197.
55. D. J. Lipman and W. R. Pearson, *Science* **227** (1985) 1435–1441.
56. W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. USA* **85** (1988) 2444–2448.
57. S. F. Altschul, W. Gish, W. Miller, and D. J. Lipman, *Journal of Molecular Biology* **215** (1990) 403–410.
58. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Nucleic Acid Res.* **25** (1997) 3389–3402.
59. M. Randić, *J. Comput. Chem*. **33** (2012) 702–707.
60. H. Hauptman and J. Karle, *Acta Cryst.* **1** (1948) 70–75.
61. H. Hauptman and J. Karle, *Acta Cryst.* **6** (1953) 131–135.
62. H. Hauptman, *On the Beauty of Science*, Prometheus Books, Amherst, NY, 2008.
63. M. Randić, *J. Comput. Chem*. **34** (2013) 77–82.
64. M. Randić, A. F. Kleiner, and L. M. De Alba, *J. Chem. Inf. Comput. Sci.* **34** (1994) 77–82.
65. M. Kunz and Z. Rádl, *J. Chem. Inf. Comput. Sci.* **38** (1998) 374–378.
66. M. Randić, J. Zupan, and A. T. Balaban, *Chem. Phys. Lett*. **397** (2004) 247–252.
67. B. Horvat, G. Jaklič, I. Kavkar, and M. Randić, *J. Math. Chem*., (in press)
68. Clustal Omega server, https://www.ebi.ac.uk/Tools/msa/clustalo/ (accessed on December 14, 2013).
69. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D Thompson, and D. G. Higgins, *Mol. Syst. Biol*. **7** (2011) 1–6 (art. no. 539).
70. G. Agüero-Chapin, R. Molina-Ruiz, E. Maldonado, G. de la Riva, A. Sanchez- Rodrigues, V. Vasconcelos, and A. Antunes. *PLOS ONE* (2013), (in press).
71. Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L.Pan, and Y. Ye, *MATCH Commun. Math. Comput. Chem*. **68** (2012) 621–637.
72. V. Consonni and R. Todeschini, *MATCH Commun. Math. Comput. Chem*. **68** (2012) 581–592.

SUPPLEMENTARY MATERIAL

Table S1. Cattle (*Bos taurus*): The eigenvalues for triangle and square maps

Cattle (*Bos taurus*)	Triangle grid	Square grid	Cattle (*Bos taurus*)	Triangle grid	Square grid
1	165.838	213.676	28	−0.6278	−0.823809
2	−49.329	−60.8261	29	−0.5988	−0.782625
3	−42.5682	−58.0783	30	−0.5812	−0.761899
4	−14.5282	−19.4754	31	−0.5646	−0.740955
5	−8.3779	−14.0429	32	−0.5454	−0.73865
6	−7.4158	−7.43305	33	−0.5277	−0.716817
7	−3.8541	−5.45275	34	−0.5183	−0.711153
8	−3.4813	−5.08666	35	−0.5018	−0.669258
9	−3.2333	−3.64131	36	−0.4806	−0.646672
10	−2.6904	−3.13543	37	−0.4795	−0.640988
11	−2.0786	−2.77815	38	−0.4701	−0.615453
12	−1.9557	−2.55075	39	−0.4361	−0.600133
13	−1.7015	−2.19047	40	−0.4186	−0.587079
14	−1.4153	−1.871	41	−0.4026	−0.573943
15	−1.2938	−1.63039	42	−0.4004	−0.548024
16	−1.1988	−1.57115	43	−0.3851	−0.538311
17	−1.0696	−1.53554	44	−0.3773	−0.524602
18	−1.0637	−1.32045	45	−0.3733	−0.499955
19	−0.9489	−1.27537	46	−0.3656	
20	−0.9108	−1.16248	47	−0.3545	
21	−0.8151	−1.08871	48	−0.3502	
22	−0.8074	−1.07443	49	−0.3398	
23	−0.7515	−1.03544	50	−0.3264	
24	−0.7211	−0.983105	51	−0.3172	
25	−0.7008	−0.950591	52	−0.3046	
26	−0.6511	−0.895029	53	−0.2995	
27	−0.6431	−0.871074	54	−0.2862	

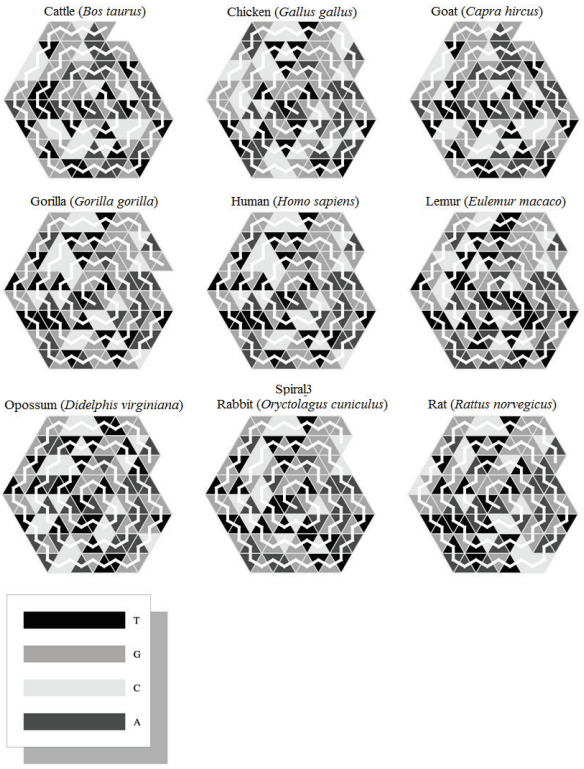

Figure S1. Four-color triangular maps of the first exon of the β-globin for nine species.

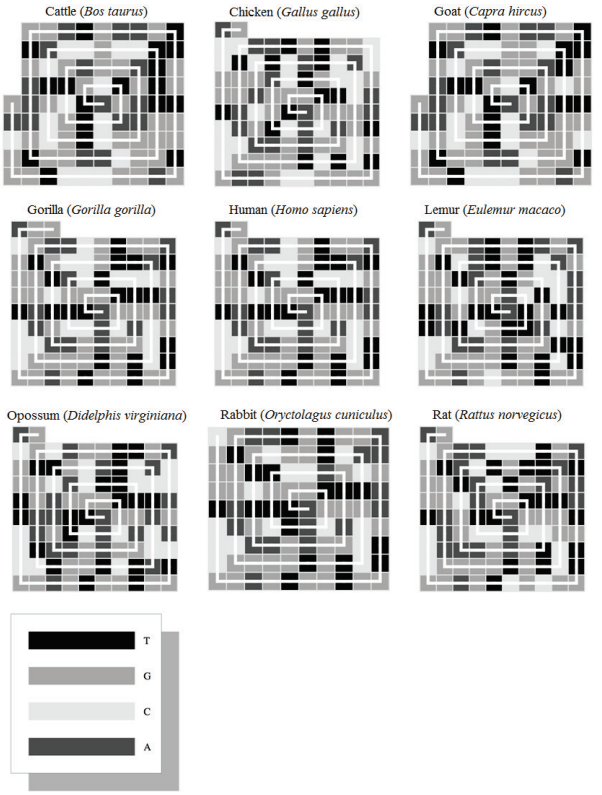

Figure S2. Four-color square maps of the first exon of the β-globin for nine species.