Journal of Mathematical Chemistry, Vol. 43, No. 2, February 2008 (© 2006)
DOI: 10.1007/s10910-006-9219-1

On representation of DNA by line distance matrix

Milan Randić* and Jure Zupan
National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia
E-mail: mrandic@msn.com, jure.zupan@ki.si

Tomaž Pisanski
Department of Mathematics, Faculty of Mathematics and Physics,
University of Ljubljana, Ljubljana, Slovenia
E-mail: tomaz.pisanski@fmf.uni-lj.si

Received 2 September 2006; accepted 30 October 2006

Recently Line Distance (LD) matrix has been introduced as a novel route for characterization of DNA sequences. The approach was based on construction of four separate submatrices for the four nucleotides, the first row of each of which records the separation between the selected nucleotide and the remaining nucleotides of the same kind. In this article, we consider an alternative representation of DNA by LD matrix in which we construct a single matrix for each DNA sequence. The approach is illustrated on the DNA sequence of the first exon of human β-globin gene.

1. Introduction

In mid-1980 the first attempts have been reported on graphical representation of DNA sequences [1–3]. In the following years several additional or alternative graphical representation of DNA were introduced [4–7].The basic idea behind these initial graphical representations of DNA was merely to facilitate visual inspection of long DNA sequences, which could perhaps help one to qualitatively recognize similarities and dissimilarities between different DNA sequences or their segments. About 15 years later an important novel feature to accompany graphical representation of DNA was recognized: it has been proposed that graphical representations of DNA allow construction of characterization of DNA, which thus offer quantitative approach to measure the degree of similarity and dissimilarity among similar DNA sequences numerically [8]. The underlying idea of these numerical explorations of graphical DNA representations is to associate with DNA sequences, consisting of the four-letter alphabet, A, C, G, T (standing for the four nucleotides: Adenine (A), Cytosine (C),

*Correspondence author.

Guanine (G), and Thymine (T), respectively), an *ordered set* of numbers, which represent sequence invariants. Here as sequence invariants are understood various sequence properties expressed numerically, which are, of course, independent of the labels used to designate individual nucleic bases. For instance, the number of nucleotides N in a sequence is one such invariants, just as are the corresponding numbers of N_{A}, N_{C}, N_{G} or N_{T}, which tell the number of A, C, G, and T nucleotides in a DNA sequence, respectively. However, while all these numbers carry useful information about a DNA sequence, they are not very discriminating, because there are many DNA sequences that may have the same set of N_{A}, N_{C}, N_{G} or N_{T} numbers. The "art" in the search for numerical DNA characterization by invariants is in finding those mathematical properties of DNA sequences that would be sensitive to minor variations in DNA composition and at the same time capture their important structural features.

In following years additional alternative graphical representations of DNA followed [9–19], including also non-graphical representation of DNA [20–25]. In this way it was possible to express numerically the degree of similarity of dissimilarity between two DNA sequences or parts there of. There are already a number of alternative numerical characterizations of DNA sequences but, just as has been the case with QSAR (the quantitative structure-activity relationship), where a large number of molecular descriptors have been introduced [26–32] in order to characterize different structural aspects of complex molecules, so the same is to be expected for characterization of widely different DNA sequences. Hence, it is desirable to introduce additional DNA descriptors based either on their graphical representation or constructed directly from the primary sequences. While non-graphical representations of DNA have some disadvantage in respect to graphical representations of DNA by not facilitating visual inspection of sequences, they may have an important advantage over graphical representations in that numerical information extracted from such representations will not be obscured by *ad hoc* geometrical elements that different graphical representations involve. In this contribution, we consider one such novel non-graphical characterization of DNA, based on the Line Distance (LD) matrix. We will outline selected properties of novel representations on the DNA sequence based on the LD matrices and their invariants. We start by first outlining the LD matrix, the entries of which are the lengths of the line segments which define a line. We will follow with a discussion of several mathematical properties of the LD matrices that will be of interest for arriving at novel characterization of DNA in a format of DNA "profiles."

2. The line distance matrix

Let us start with by considering the line shown at the top of figure 1, which is partitioned by points: $n_0, n_1, n_2, n_3, \ldots, n_N$, which fully define a line having N segments. The particular line of figure 1 has nine segments, which are determined

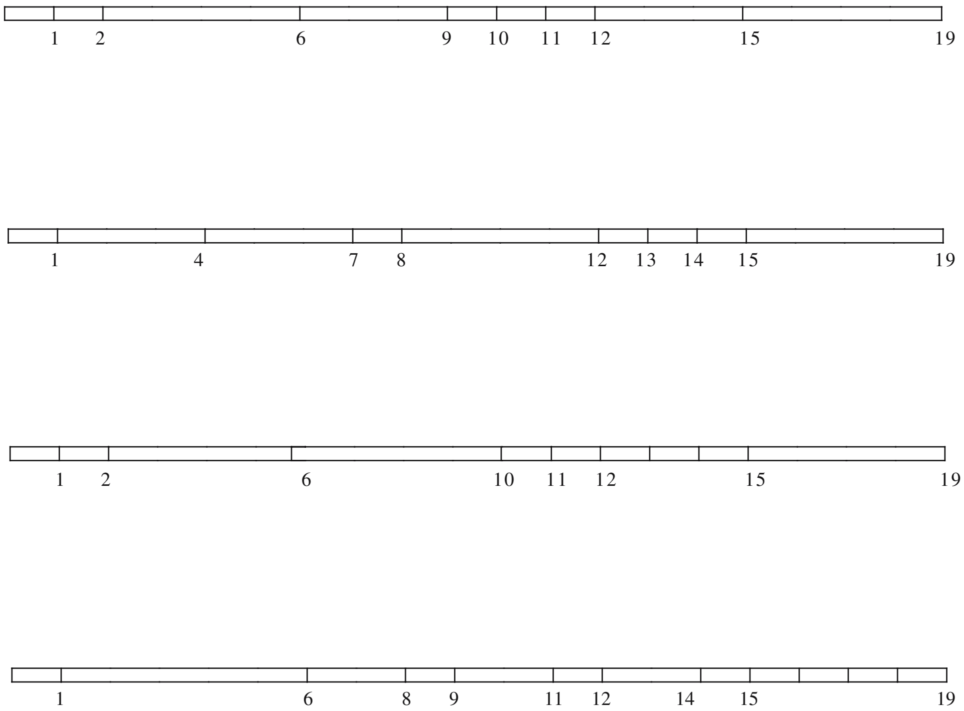


Figure 1. Four lines of the same length having different line intervals. The first two lines differ only in distribution of line segments.

by integer coordinate $x = 0, 1, 2, 6, 9, 10, 11, 12, 15,$ and 19. At the first sight it appears that the set of coordinates offers all the information on the line and that it would be hardly possible to obtain additional properties about the line, which would not follow from straightforward arithmetic and algebraic manipulations of the above numbers. For example, by subtracting the adjacent integers one obtains: 1, 1, 4, 3, 1, 1, 1, 3, 4, which are the lengths of the successive line segments. However, as we will show, it is possible to arrive at additional characterizations of lines, which offer extraction of additional properties of the line considered, which can then serve for numerical characterization of the line. Such possibility, which has been only recently considered [33–36], is to associate with partitioned lines a symmetrical $N \times N$ matrix, the elements of the first row of which are defined as the distance between the first point (n_0) and the remaining points that define individual line segments (n_i): ($n_0 - n_0$), ($n_0 - n_1$), ($n_0 - n_2$), ($n_0 - n_3$), ...

The entries in the remaining rows of the matrix above the main diagonal to the right are similarly defined by:

$$(n_i - n_i), (n_i - n_{i+1}), (n_i - n_{i+2}), (n_i - n_{i+3}), \ldots$$

Table 1
The line distance matrix for the line A of figure 1.

	1	2	3	4	5	6	7	8	9	10	Row Sum
1	0	1	2	6	9	10	11	12	15	19	85
2	1	0	1	5	8	9	10	11	14	18	77
3	2	1	0	4	7	8	9	10	13	17	71
4	6	5	4	0	3	4	5	6	9	13	55
5	9	8	7	3	0	1	2	3	6	10	49
6	10	9	8	4	1	0	1	2	5	9	49
7	11	10	9	5	2	1	0	1	4	8	51
8	12	11	10	6	3	2	1	0	3	7	55
9	15	14	13	9	6	5	4	3	0	4	73
10	19	18	17	13	10	9	8	7	4	0	105

Hence, the list of coordinates along the line: $x = 0, 1, 2, 6, 9, 10, 11, 12, 15$, and 19 represent the entries in the first row of the LD matrix. The second row is obtained by subtracting the entry above the zero on the diagonal, which is entry 1, from the corresponding elements of the first row, giving, starting with the zero on the main diagonal: 0, 1, 5, 8, 9, 10, 11, 14, and 18. The third row of the LD matrix is obtained by subtracting again 1, the entry above the diagonal zero in the third row, from the corresponding elements of the first row giving: 0, 4, 7, 8, 9, 10, 13, and 17. In the next row similarly we subtract 4, the entry above the diagonal zero in the fourth row, from the corresponding elements of the first row giving: 0, 3, 4, 5, 6, 9, and 13. In table 1, we show the complete LD matrix for the line of figure 1. Observe that matrix elements adjacent to the main diagonal represent the length of the individual line segments making the line. In fact the LD matrix is nothing but the distance matrix for the points of the line defining line segments.

The LD matrix is fully defined by the entries in its *first row*. It may appear therefore that the information of the remaining rows is redundant. In this respect the LD matrix bears relationship with the Toepliz matrices and the Hankel matrices [37–41], which are also fully determined by the elements in its first row. However, in contrast to Toepliz and Hankel matrices, the subsequent row of which are obtained by cyclic permutations of the elements of the first row here the successive rows are constructed by applying pertinent arithmetic manipulations on the matrix elements of the first row. The novelty of the present approach is that for the first time we associate matrix with a line. Introduction of a matrix for representation of a line not only allows construction of additional line invariants but also allows construction of additional matrices, the properties of which may turn out to be of interest when comparison of different sequences is considered.

3. Line matrix invariants

In this contribution, where we have in mind application of LD matrices for characterization of DNA, we are interested in line segments that are expressed by integers. Hence, as a result we will have LD matrices the elements of which are integer. We have already illustrated one such smaller matrix in table 1. In table 2, we have listed a selection of matrix invariants which can be considered as Line Matrix descriptors. The average matrix element, the average row sum, or the sum of all matrix elements is one such invariant. If the sum of all matrix elements is divided by 2 one obtains the Wiener number [42] of the LD matrix. The average matrix element and the average row sum contain the same information as the Wiener number of the matrix, the three quantities differ only in the constant of proportionality. We prefer to use the average row sum, because the smallest and the largest row sums represent the lower and the upper bound to the leading eigenvalue of the matrix, and thus often the average row sum is a good approximation to the leading eigenvalue, which not only is easy to calculate but is conceptually simpler quantity.

The eigenvalues of LD matrices represent also potential DNA descriptors. It has been recently conjectured and then proved that the LD matrices have only one positive eigenvalue [33]. This need not be so important in applications in which often one only considers the leading eigenvalues of matrices. The average line segment, which is another line invariant, can be easily obtained from the elements next to the main diagonal. In the case of the LD matrix of table 1 the average line segment happens to be 19/9 or 2.1111. Similarly one can calculate the average of the entries that are in the diagonal next to the considered diagonal, that is, the average of 2, 5, 7, 4, 2, 2, 4, 7, which or 33/8 or 4.125, and so on. In this way in the case of the LD matrix of table 1 one obtains the sequence:

2.1111, 4.1250, 6.1426, 8.0000, 9.8000, 12.0000, 14.3333, 16.5000, 19.0000 clearly one could derive this sequence also from the initial information on the line given by coordinates: 0, 1, 2, 6, 9, 10, 11, 12, 15, 19, but its derivation and

Table 2
A selection of matrix invariants of the line distance matrix of table 1.

Invariant	Description
N	The size of the matrix
$(\Sigma d_{ij})/[N(N-1)]$ (sum over i, j)	Average matrix element
$\Sigma\{(\Sigma d_{ij})/[N-1]\}/N$ (sum over j),{sum over i}	Average row sum
W	Wiener number
λ_1	The leading eigenvalue
λ_i	Eigenvalues
(Σd_{ij}) (sum over $i, i < j$)	Sequence of average diagonal sums

interpretation as the averages of selected off-diagonal elements of the LD matrix is less cumbersome.

While the average of the entries that are in the diagonal next the main diagonal and another entries of the sequence shown above could be evaluated by successive manipulations of the initial line entries there are line sequence properties associated with LD matrix, which cannot be derived from the information on the line but require construction of the matrix. Such are, for instance, the eigenvaules of the LD matrix, the characteristic polynomial, and the eigenvalues of the Lapalacian of LD matrix.

Observe that while the first row of the LD matrix conserves the information on the distribution of different segments along the line if we consider manipulating with the entries that constitute the list that gives the lengths of individual segments: 1, 1, 4, 3, 1, 1, 1, 3, and 4, we are losing information relating to the distribution of the segments. For instance, we can calculate the total length of all segments, which is 19, and can follow by considering the total length of segments squared, that is, the sum $1^2+ 1^2+ 4^2+ 3^2+ 1^2+ 1^2+ 1^2+ 3^2+ 4^2$, which is 55, and so on continue with higher powers of segments. In order to control the convergence of so constructed sequence one has to introduce $1/n!$ as normalizing factor. The resulting sequence is listed in table 3 and illustrated in figure 2.

In figure 3, we illustrate "Line Profiles" for additional Line matrices listed in table 4, for lines of the same length (shown in figure 1), which however have somewhat different interval entries along the main diagonal. We included these matrices in order to show how sensitive are the introduced profiles to minor changes in line intervals. As we can see the line profiles of figure 3 are of similar shape with the Line profile of figure 2, but all three line profiles though having the overall similar shape differ somewhat in their relative magnitudes. So we may conclude that characterization of lines by LD matrices appears to be sufficiently sensitive to minor changes in the length of line segments. Observe that the matrix A and B though different, have the same number of segments of the same length, and thus result in the same profile. This approach, which only considers the average segment lengths (and similar entries based on higher powers

Table 3
The line profile for line A of figure 1 shown as bar graph in figure 2.

$n = 1$	19
$n = 2$	27.5
$n = 3$	31.17
$n = 4$	28.29
$n = 5$	21.16
$n = 6$	13.41
$n = 7$	7.37
$n = 8$	3.58
$n = 9$	1.55
$n = 10$	0.61

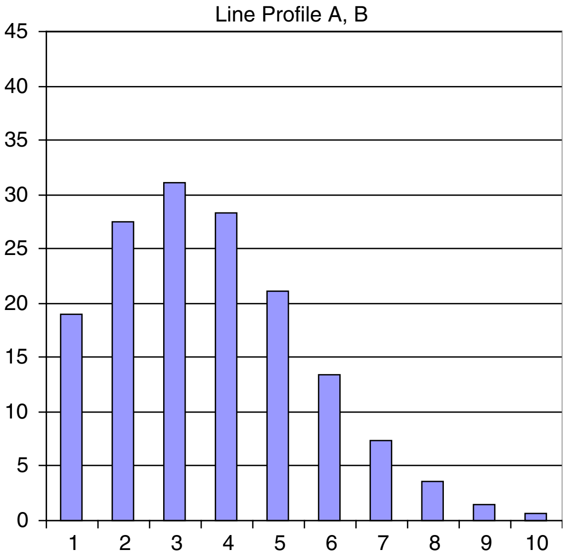


Figure 2. The line profile for lines A and B of figure 1.

of segment lengths) cannot differentiate Lines A and B, which only differ in distributions of segments along the line. Hence, in those situations, as we will see later, one has to resort to the complete LD matrices, rather then considering only their intervals (given by elements next to the main diagonal).

4. Application of line distance matrix to DNA

4.1. A, C, G, and T distance matrices

We will illustrate application of LD matrix to DNA by considering the first exon of the human β-globin gene shown below, in which for better visibility we have grouped ten nucleotides into a single "word":
ATGGTGCACC TGACTCCTGA GGAGAAGTCT GCCGTTACTG
CCCTGTGGGG CAAGGTGAAC GTGGATGAAG TTGGTGGTGA
GGCCCTGGGC AG
If we consider only adenine A then the following are the sites of 17 adenine bases in the above sequence:
1, 8, 13, 20, 23, 25, 26, 37, 52, 53, 58, 59, 65, 68, 69, 80, 91
The difference, that is, the distance, from the first point on the line gives the first row of the 17 × 17 LD matrix shown in table 5: 0, 7, 12, 19, 22, 24, 25, 36, 51, 52, 57, 58, 64, 67, 68, 79, 90
The LD matrix of table 5 is one of four such matrices representing the first exon of the human β-globin gene that can be constructed. The remaining three matri-

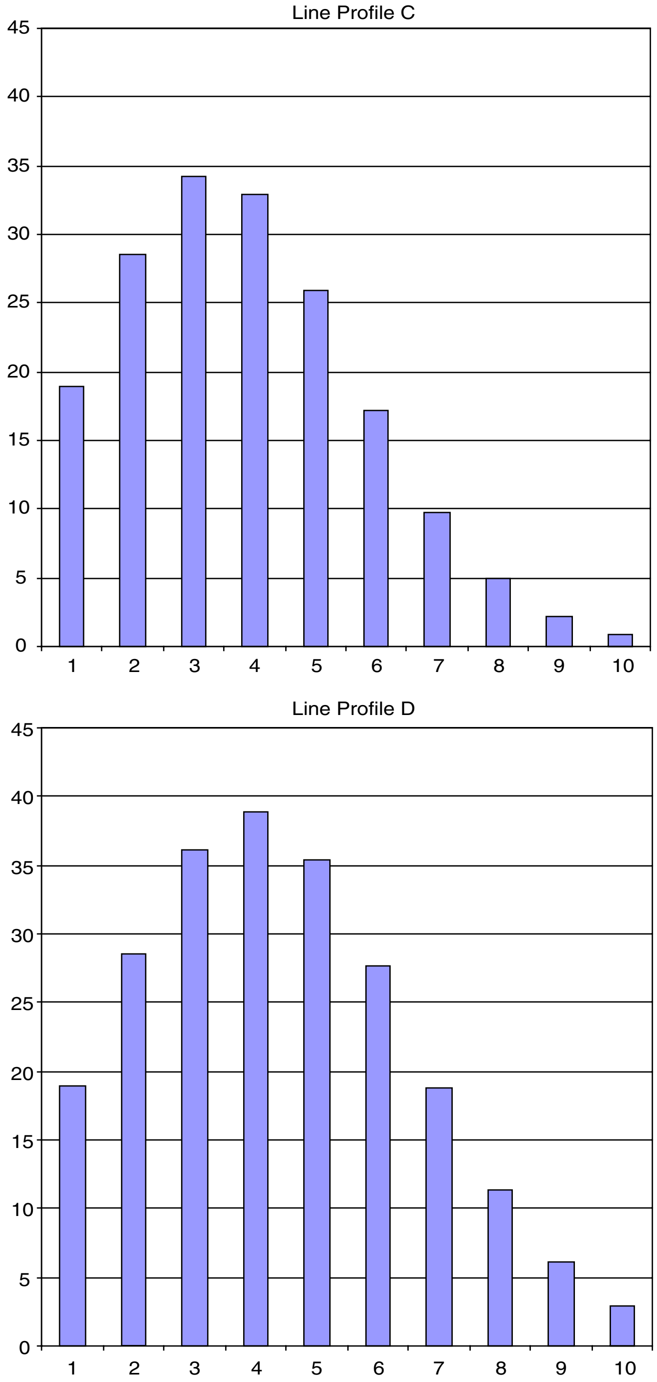


Figure 3. The line profile for lines C and D of figure 1.

Table 4
The LD matrix for the lines B, C, and D of figure 1.

	1	2	3	4	5	6	7	8	9	10	Row Sum
B											
1	0	1	4	7	8	12	13	14	15	19	93
2		0	3	6	7	11	12	13	14	18	85
3			0	3	4	8	9	10	11	15	67
4				0	1	5	6	7	8	12	55
5					0	4	5	6	7	11	53
6						0	1	2	3	7	53
7							0	1	2	6	55
8								0	1	5	59
9									0	4	65
10										0	97
										Total	682
C											
1	0	1	2	6	10	11	12	13	15	19	89
2		0	1	5	9	10	11	12	14	18	81
3			0	4	8	9	10	11	13	17	75
4				0	4	5	6	7	9	13	59
5					0	1	2	3	5	9	51
6						0	1	2	4	8	51
7							0	1	3	7	53
8								0	2	6	57
9									0	4	69
10										0	101
										Total	686
D											
1	0	1	6	8	9	11	12	14	15	19	95
2		0	5	7	8	10	11	13	14	18	87
3			0	2	3	5	6	8	9	13	57
4				0	1	3	4	6	7	11	49
5					0	2	3	5	6	10	47
6						0	1	3	4	8	47
7							0	2	3	7	49
8								0	1	5	57
9									0	4	63
10										0	95
										Total	646

ces associated with cytosine, guanine, and thymine are obtained by considering the successive sites for C, G, and T, respectively:

C: 7, 9, 10, 14, 16, 17, 29, 32, 33, 38, 41, 42, 43, 51, 60, 83, 84, 85, 90
G: 3, 4, 6, 12, 19, 21, 22, 24, 27, 31, 34, 40, 45, 47, 48, 49, 50, 54, 55,
57, 61, 63, 64, 67, 70, 73, 74, 76, 77, 79, 81, 82, 87, 88, 89, 92
T: 2, 5, 11, 15, 18, 28, 30, 35, 36, 39, 44, 46, 56, 62, 66, 71, 72, 75, 78, 86

Which lead to 18 × 18, 35 × 35, and 19 × 19 matrices.

Table 5
The line distance matrix for the line depicting separations of successive adenine bases of the first exon of human β-globin gene.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	7	12	19	22	24	25	36	51	52	57	58	64	67	68	79	90
2		0	5	12	15	17	18	29	44	45	50	51	57	60	61	72	83
3			0	7	10	12	13	24	39	40	45	46	52	55	56	67	79
4				0	3	5	6	17	32	33	38	39	45	48	49	60	72
5					0	2	3	14	29	30	35	36	42	45	46	57	69
6						0	1	12	27	28	33	34	40	43	44	55	67
7							0	11	26	27	32	33	39	42	43	54	66
8								0	15	16	21	22	28	31	32	43	55
9									0	1	6	7	13	16	17	28	40
10										0	5	6	12	15	16	27	39
11											0	1	7	10	11	22	34
12												0	6	9	10	21	33
13													0	3	4	16	27
14														0	1	13	24
15															0	12	23
16																0	11
17																	0

Selected properties of such matrices can be used in a comparative study of β-globin gene DNA sequences of different species. In figures 4–7, we have depicted the corresponding A, C, G, and T profiles, respectively, which are obtained by raising the LD matrix elements (such as those shown for Distance matrix belonging to adenine in table 5) to higher powers, while using $1/n!$ as the normalization factor. As one can see all the four profiles are of similar shape, but they differ dramatically when the y-coordinate scale s are compared. In the case of A the maximal height of profile "bars" approaches 350,000, in the case of C the maximal height of profile "bars" exceeds 800 millions, in the case of G the maximal height of profile "bars" does not even reaches 1400 and In the case of T the maximal height of profile "bars" approaches 6000. The critical quantity which determined the "height" of the profiles is the maximal segment separation of the identical bases, which are in the case of A, C, G, and T matrices 15, 23, 9, and 10, respectively.

4.2. Global DNA distance matrices

The LD matrix has been introduced [33–36] for characterization of distribution of the four bases in DNA by considering separately each of the four bases. In this way a single sequence of DNA has been represented by four LD matrices, the size of which depended on the numbers N_A, N_C, N_G or N_T. Conse-

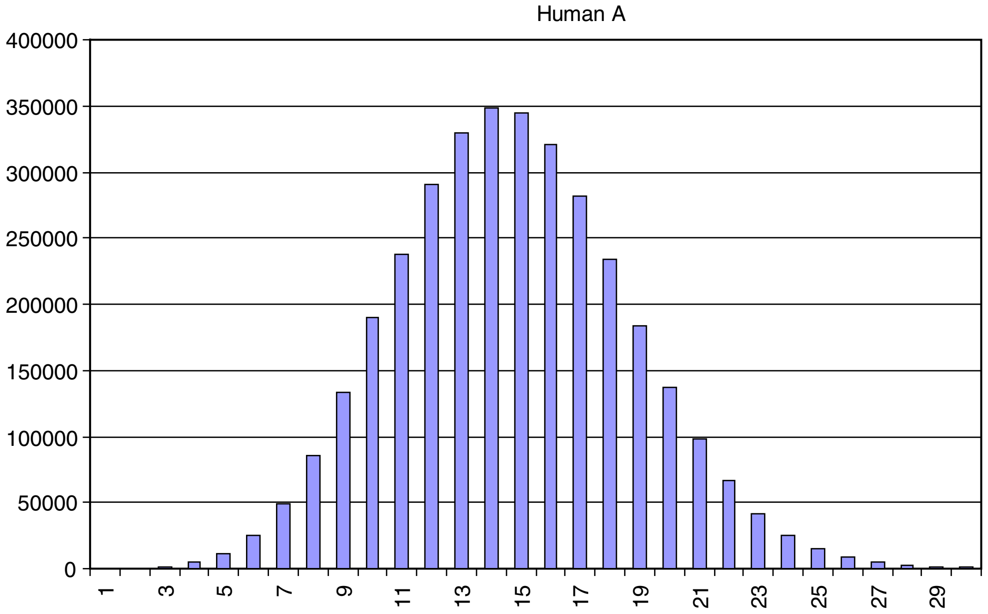


Figure 4. The line profile based on 17 × 17 line distance matrix for adenine.

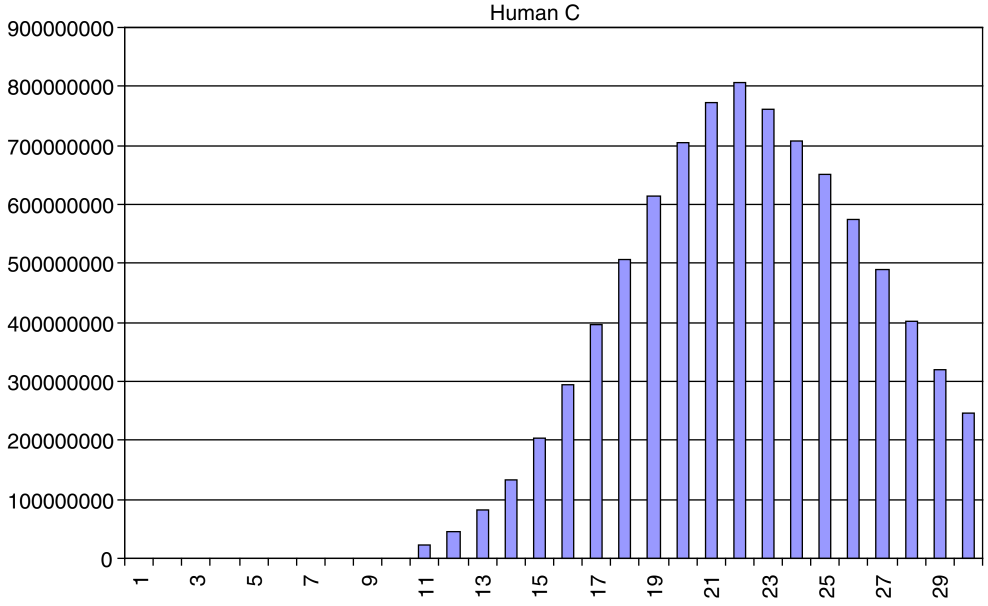


Figure 5. The line profile based on 187 × 18 line distance matrix for cytosine.

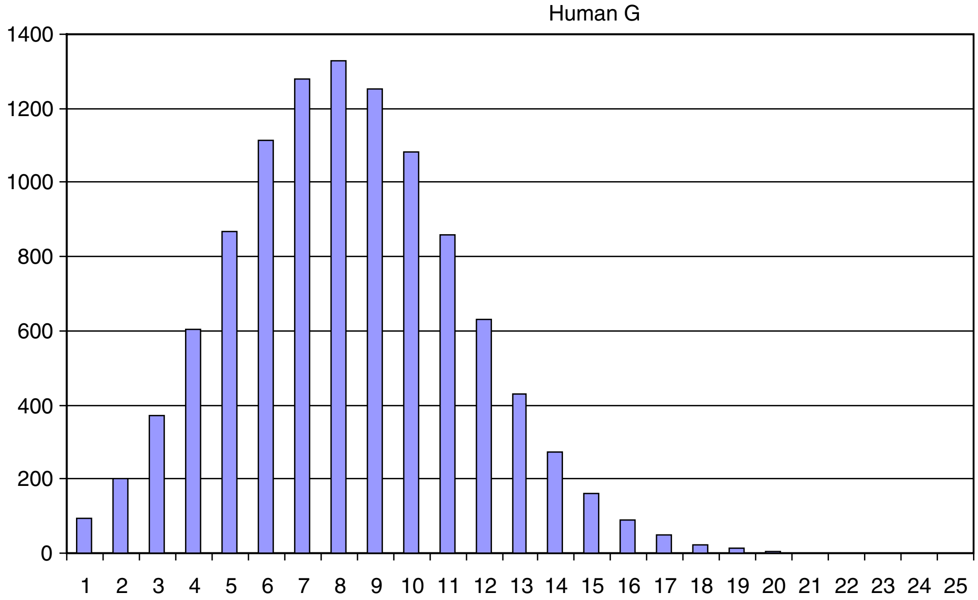


Figure 6. The line profile based on 35 × 35 line distance matrix for guanine.

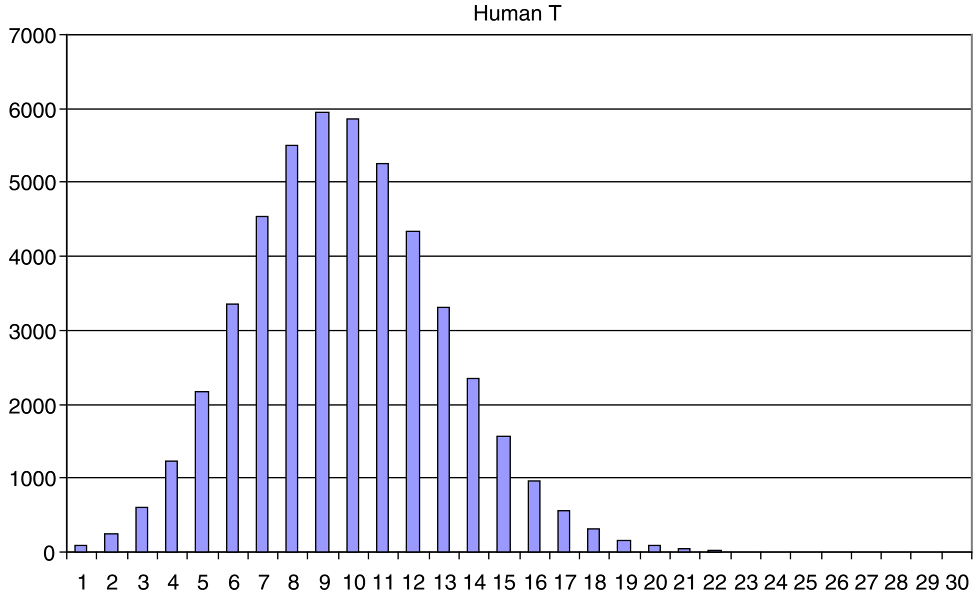


Figure 7. The line profile based on 19 × 19 line distance matrix for thymine.

Table 6
The first exon of human β-globin gene and accompanying numerical sequence in which entries indicate $\pi/2$ segments between successive bases placed on the periphery of a circle.

A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A
0	1	1	4	3	1	1	1	3	4	2	1	2	3	2	2	4	2	1	2
G	G	A	G	A	A	G	T	C	T	G	C	C	G	T	T	A	C	T	G
2	4	2	2	2	4	2	3	2	2	1	1	4	3	3	4	3	3	2	1
C	C	C	T	G	T	G	G	G	G	C	A	A	G	G	T	G	A	A	C
1	4	4	2	1	3	1	4	4	4	1	1	4	2	4	3	1	2	4	3
G	T	G	G	A	T	G	A	A	G	T	T	G	G	T	G	G	T	G	A
3	3	1	4	2	1	1	2	4	2	3	4	1	4	3	1	4	3	1	2
G	G	C	C	C	T	G	G	G	C	A	G								
2	4	1	4	4	2	1	4	4	1	1	2								

quently sequences of DNA for which $N_{\mathrm{A}} + N_{\mathrm{C}} + N_{\mathrm{G}} + N_{\mathrm{T}} =$ constant will be treated differently if they have the same N_{A}, N_{C}, N_{G} or N_{T} or not, which translates into different treatment of permutation of two nucleotides from substitution (which would change N_{A}, N_{C}, N_{G} or N_{T}. It would be desirable and of interest to develop computational approach in which DNA representation would be based on a single LD matrix, rather than a set of four such matrices. That this indeed is possible will be demonstrated in this section.

Let us consider the same DNA sequence of the first exon of human β-globin gene shown before. Our intention is to replace the four-letter sequence by numbers, and we want these numbers to be associated with distance between nucleotides. One way of doing this is shown in table 6. We will associate with the four nucleotides A, T, G, and C the four locations on the unit circle. Such choice is related to the selection of the four directions of the positive and the negative coordinate system as Nandy used in his graphical representation of DNA [6], and the four corners of a square of Jeffrey in his constructions of "chaos game" representation of DNA [43]. The chaos game is an invention of Barnsley [43], who considered geometrical construction based on selecting an arbitrary n polygon and selecting at random a point inside it. From this point one moves assumed fraction of distance toward another vertex of the polygon selected at random and continues the "game" indefinitely. Jeffrey's construction of highly compact representations of DNA is based on using a square, starting from the center and moving toward the corner assigned to the base. From this point one moves half a way to the corner assigned to the second base, and so on. In this way one is always on the interior of the A, C, G, T-square, regardless how many bases in DNA one considers.

In contrast to the approach of Jeffrey here one starts with the first nucleotide A and move along the circle periphery in positive sense (i.e., anticlockwise) to record the number of "quarter"-segments needed to reach the next base (figure 8). We measure the distance between the successive bases by counting the

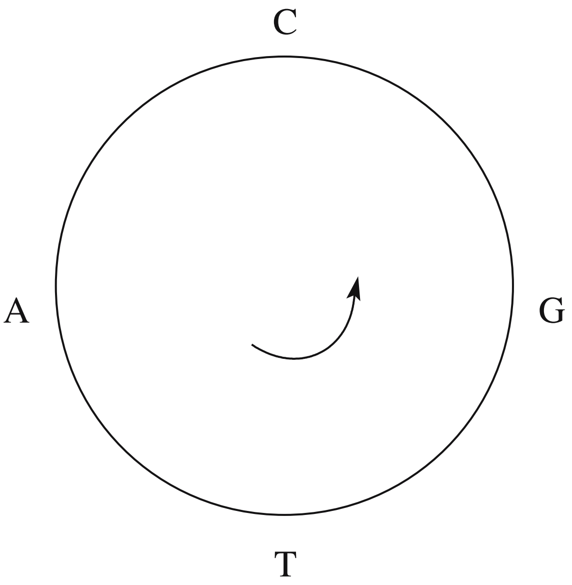


Figure 8. The unit circle with the four bases placed on the circumference of the circle.

quadrants (or $\pi/2$ arcs) that ore required to pass to come from one base to the next. Thus the distance from the first base A to the adjacent T equals 1, from T to the adjacent G again the distance is 1 but the distance between the two adjacent G bases equals 4, because we have to pass a full 360° to come again to G. In the second row of table 6 we have listed the distances between the adjacent bases, which, except for the initial zero, can take values between 1 and 4. The entry 1 are associated with the successive pairs of nucleotides AT, TG, GC, and CA, the entry 2 with the pairs AG, TC, GA, and CT, the entry 3 with the pairs AC, TA, GT, and CG, and finally the entry 4 belonging to distances between pairs of the identical bases. Observe that the entries 1–4 in the second row of table 6 are not associated with individual nucleic bases, as has been the case with the graphical representation of DNA based on the "four" horizontal lines [9]. Here the same entry can stand sometimes for A, sometimes for T, G or C. It may appear thus that there is loss of information in transforming the four letters DNA sequence to the four integers sequence of table 6 but this is not the case. It is not difficult to reconstruct the DNA sequence from the entries of the second row when one knows the cyclic ordering A–T–G–C.

The 10 × 10 matrix of table 1 is in fact the initial part of the 92 × 92 LD matrix of the DNA sequence of the first exon of human β-globin gene, the segments of which are listed just below DNA sequence of table 6. In figure 9–11 we have shown the resulting profiles for the 92 bases of the first exon of human β-globine gene, mouse β-globin gene and lemur β-globin gene. Visual inspection

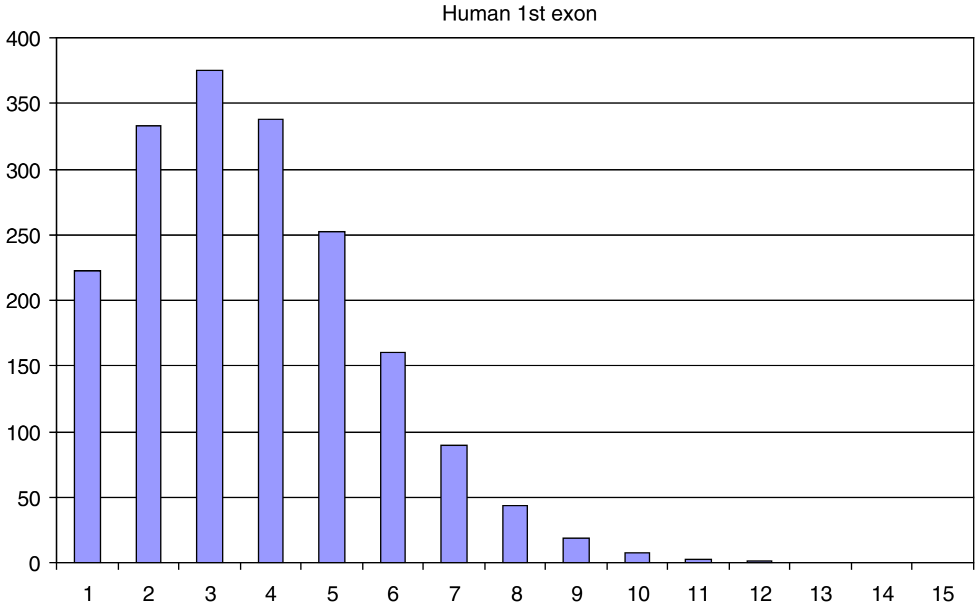


Figure 9. The line profile based on the 92 × 92 line distance matrix for the first exon of human β-globin gene.

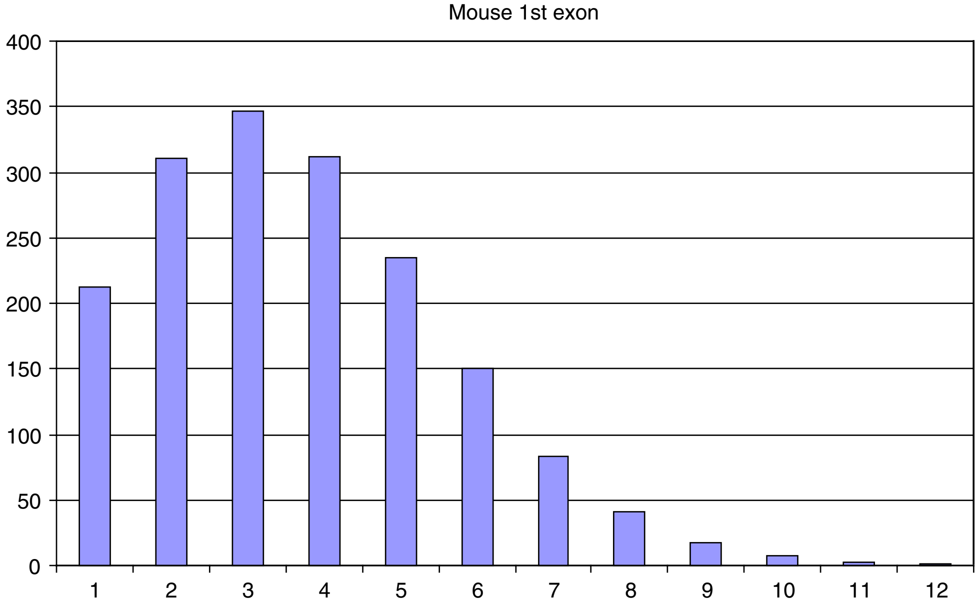


Figure 10. The line profile based on the 92 × 92 line distance matrix for the first exon of mouse β-globin gene.

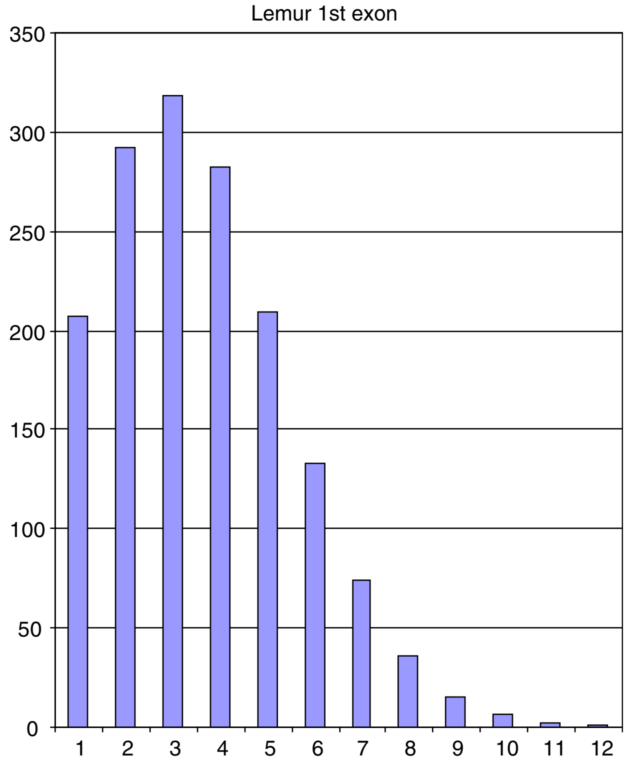
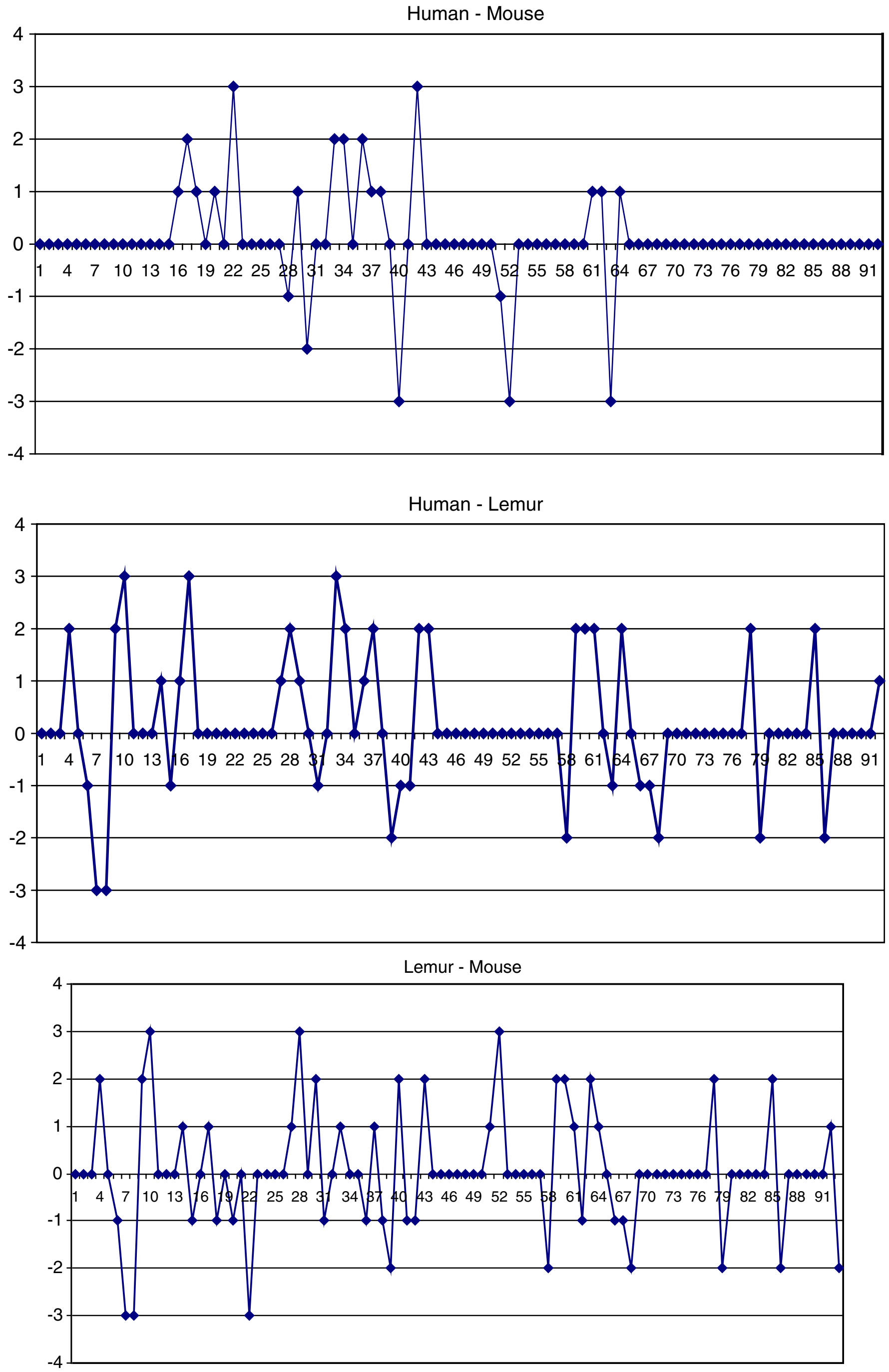


Figure 11. The line profile based on the 92 × 92 line distance matrix for the first exon of lemur β-globin gene.

shows that all the three profiles are similar, though smaller differences can also be noticed, particularly when one compares the scale on the y-coordinates.

5. Comparison of different DNA sequences

It has been only recently recognized that "transformation" of alphabetic biosequences (such as A, C, G, and T into numerical sequences offers novel important possibilities in comparative study of DNA [44–49]. Simple arithmetic manipulations allow one to identify similar segments in lengthy DNA sequences. The same applies also to numerical sequences such as the DNA sequence of table 6, in which entries between 1 and 4 appear. In figure 12, we have plotted for human, mouse and lemur the difference between the corresponding sequences of their first exon of β-globin gene. As we see in the case of the pairs human–mouse and human–lemur we can immediately by inspection see a greater similarity for the pair human–mouse than human–lemur. Simple quantitative index of similarity is given by the quotient $N_{\mathrm{O}}/N_{\varnothing}$, where N_{O} and $N_{\varnothing}$ stands for the


Human - Mouse

Human - Lemur

Lemur - Mouse

Figure 12. The numerical difference between DNA sequences of the first exon of human, mouse, and lemur globin gene.

count of zeros and the count of non-zero elements in the difference plot. In the case of figure 12, we have $N_{\mathrm{O}}/N_{\varnothing} = 71/21$, and $N_{\mathrm{O}}/N_{\varnothing} = 56/36$, respectively, for the human–mouse and human–lemur cases. These results may have not been expected, but is not surprising after recent study that has shown a considerable similarity between human and rodents, than was suspected [50]. Finally, we may observe from Figure 12 that lemur–mouse is the least similar pair of the three considered, with $N_{\mathrm{O}}/N_{\varnothing} = 49/43$.

6. Summary

We have outlined construction of DNA "profiles" as novel descriptors of DNA or fragments thereof. Profiles are calculated from the information on the lengths of segments of lines that serve for representation of DNA. As has been outlined the constructed DNA "profiles" appear to be sensitive enough of changes in compositions of DNA and at the same time it appears that they carry important structural information in their numerical representations. It is interesting to observe that although the corresponding segments of DNA belonging to the first exon of β-globin gene of human and mouse have greater number of identical bases in the corresponding locations in the DNA sequence the overall similarity between human and lemur when expressed via the corresponding DNA profiles (figure 12) show greater similarity than do human and mouse case.

References

[1] E. Hamori and J. Ruskin, J. Biol. Chem. 258 (1983) 1318.
[2] E. Hamori, Nature 314 (1985) 585.
[3] M.A. Gates, J. Theor. Biol. 119 (1986) 319.
[4] E. Hamori, BioTechniques 7 (1989) 710.
[5] H.J. Jeffrey, Nucleic Acids Res. 18 (1990) 2163.
[6] A. Nandy, Curr. Sci. 66 (1994) 309.
[7] P.M. Leong and S. Morgenthaler, Comput. Appl. Biosci. 11 (1995) 503.
[8] M. Randić, M. Vračko, A. Nandy and S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[9] M. Randić, N. Lerš and D. Plavšić, Chem. Phys. Lett. 368 (2003) 1.
[10] M. Randić, Chem. Phys. Lett. 386 (2004) 468.
[11] M. Randić, M. Vračko, N. Lerš, and D. Plavšić, Chem. Phys. Lett. 371 (2003) 202.
[12] M. Randić, M. Vračko, J. Zupan and M. Novič, Chem. Phys. Lett. 373 (2003) 558.
[13] M. Randić and J. Zupan, SAR and QSAR Environ. Res. 15 (2004) 147.
[14] M. Randić, Period. Biol. 107 (2005) 415.
[15] M. Randić, D. Vikić-Topić, A. Graovac, N. Lerš and D. Plavšić, Period. Biol. 107 (2005) 437.
[16] X. Liu, Q. Dai and T. Wang, J. Mol. Graph. Model (in press).
[17] J. Zupan and M. Randić, J. Chem. Inf. Model. 45 (2005) 309.
[18] X. Guo, M. Randić, S.C. Basak, Chem. Phys. Let. 350 (2001) 106.
[19] M. Randić, N. Lerš, D. Plavšić, S.C. Basak and A.T. Balaban, Chem. Phys. Lett. 407 (2005) 205.
[20] M. Randić, J. Chem. Inf. Comput. Sci. 40 (2000) 50.

[21] M. Randić, Chem. Phys. Lett. 317 (2000) 29–342.

[22] M. Randić and S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 561.

[23] M. Randić and M. Vračko, J. Chem. Inf. Comput. Sci. 40 (2000) 599.

[24] M. Randić, X. Guo and S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 619.

[25] M. Randić and A.T. Balaban, J. Chem. Inf. Comput. Sci. 43 (2003) 532.

[26] M. Randić, in: *The Encyclopedia of Computational Chemistry,* eds. P.V.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III and P.R. Schreiner, (John Wiley, Chichester, 1998) p. 3018.

[27] A.T. Balaban, in: *Encyclopedia of Analytical Chemistry*, ed. R.A. Meyers, (Wiley Chichester, UK, 2000).

[28] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Methods and Principles in Medicinal Chemistry, Vol. 11, eds. R. Mannhold, H. Kubinyi and H. Timmerman (Wiley-VCH, New York)

[29] A.R. Katritzky, V. Lobanov and M. Karelson, *CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis)* (University of Florida, Gainesville, FL, 1994).

[30] A.T. Balaban, in: *Topological Indices and Related Descriptors in QSAR and QSPR* eds. J. Devillers, A.T. Balaban, (Gordon and Breach, Amsterdam, The Netherlands, 1999) p. 403.

[31] E. Estrada, in: *Topological Indices and Related Descriptors in QSAR, and QSPR*, eds. J. Devillers, A.T. Balaban, (Gordon and Breach, Amsterdam, 1999) p 403.

[32] S.C. Basak, G.D. Grunwald and G.J. Niemi, in: *From Chemical Topology to Three-dimensional Geometry,* ed. A.T. Balaban, (Plenum Press, New York 1977) p. 73.

[33] G. Jaklič, T. Pisanski and M. Randić, presented at Complex Objects Visualization 2005, University of Primorska, Koper, Slovenia, 16–19 November 2005.

[34] G. Jaklič, T. Pisanski and M. Randić, MATCH, Comm. Math. Chem. (submitted).

[35] G. Jaklič, T. Pisanski and M. Randić, J. Comput. Biol. (in press).

[36] G. Jaklič, T. Pisanski and M. Randić, Poster: Ghent University, Belgium. Book of abstracts, p. 28, February 6–9 2006.

[37] Brookes, M., The Matrix Reference Manual [on line] http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro. html.2005

[38] M. Hladnik, D, Marušič and T. Pisanski, Discrete Math. 244 (2002) 137.

[39] M. Hladnik, Linear Algebra Appl. 286 (1999) 261.

[40] P.J. Davis, *Circulant Matrices* (Wiley New York, 1979).

[41] C.R. Putnam, Pac. J. Math. 14 (1964) 651.

[42] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.

[43] M.F. Barnsley and H. Rising, *Fractals Everywhere, 2nd ed.* (Academic Press, Boston, MA, 1993).

[44] M. Randić, J. Math. Chem. (submitted)

[45] M. Randić, J. Zupan, D.Vikić-Topić and D. Plavšić, Chem. Phys. Lett. 431 (2006) 357.

[46] M. Randić, Acta Chim. Slovenica (in press).

[47] M. Randić, M. Novič, D. Vikić-Topić, N. Lerš and D. Plavšić, J. Mol. Graph. Model. (submitted).

[48] M. Randić, Chem. Phys. Lett. (submitted).

[49] M. Randić and D. Juretić, J. Chem. Inf. Model. (submitted).

[50] Y. Hashimotoa, T. Niikura, H. Tajima, T. Yasukawa, H. Sudo, Y. Ito, Y. Kita, M. Kawasumi, K. Koumaya, M. Doyu, G. Sobue, T. Koide, S. Tsuji, J. Lang, K. Kurokawa and I. Nishimotoa, Proc. Natl. Acad. Sci. USA 98 (2001) 6336.