On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization

M. Randić,[†,‡,§,#] M. Vračko,[†] A. Nandy,[||] and S. C. Basak*[,§]

National Institute of Chemistry, 1001 Ljubljana, POB 3430, Slovenia, Ames Laboratory - DOE, Iowa State University, Ames, Iowa 50011, Natural Resources Research Institute, University of Minnesota at Duluth, Miller Trunk Highway, Duluth Minnesota 55811, and Computer Division, Indian Institute of Chemical Biology, Calcutta, India

Received April 9, 2000

In this article we (1) outline the construction of a 3-D "graphical" representation of DNA primary sequences, illustrated on a portion of the human β globin gene; (2) describe a particular scheme that transforms the above 3-D spatial representation of DNA into a numerical matrix representation; (3) illustrate construction of matrix invariants for DNA sequences; and (4) suggest a data reduction based on statistical analysis of matrix invariants generated for DNA. Each of the four contributions represents a novel development that we hope will facilitate comparative studies of DNA and open new directions for representation and characterization of DNA primary sequences.

INTRODUCTION

With rapid reporting of DNA sequences derived with automated DNA sequencing techniques the problem of processing such information became acute. Usual representation of the primary sequence DNA is that of a string of letters A, G, C, T, which signify the four nucleic acid bases adenine, guanine, cytosine, and thymine, respectively. Such sequences can be very long, and even the segments of interests when comparing DNA of different species can be quite lengthy. In Table 1 we listed DNA of human β globin gene. Its length is 1424, and its first exon already involves 92 bases. Comparison of such primary sequences, and even their fragments having less than 100 bases, could be quite difficult for several reasons. Consider the list of the first exon of the β globin gene for eight different species shown in Table 2. They all look similar, but at the same time they are all sufficiently different. How similar or how different they are may depend on how such strings of letters are encoded or characterized. The standard procedures consider differences between strings due to deletion−insertion, compression−expansion, and substitution of the string elements.[1−9] These approaches have been applied to a variety of problems, from the error correcting codes in which Levenshtein has introduced metrics for string comparisons[1] to comparison of DNA sequences, comparison of protein sequences, and applications in quantitative structure−activity relationship (QSAR).[8,9] Such approaches, that have been hitherto widely used, are computer intensive.

We have recently proposed an alternative approach for comparison of sequences that is based on characterization of DNA by ordered sets of *invariants* derived for DNA sequence, rather than by a direct comparison of DNA sequences themselves. This is analogous to the use of graph invariants (topological indices) for characterization of molecules rather than use of information on their geometry and types of atoms involved. An important advantage of a characterization of structures (be it molecule or DNA) by invariants, as opposed to use of codes, is the simplicity of the comparison of numerical sequences based on invariants. The price paid is a loss of information on some aspects of the structure that accompany any characterization based on invariants. The loss of information, however, can be in part reduced by use of a larger number of descriptors (invariants), as has been well illustrated in SAR and QSAR based on mathematical descriptors for molecules.[10−12]

Graphical representations of DNA that have been developed within the past few years[13−15] offer a route to one such condensation of information coded by DNA primary sequence into a set of invariants. In Figure 1 we show few graphical representations of selected DNA as reported by Nandy.[16] The graphs are obtained by assigning to the four directions associated with the positive and the negative *x, y* axes the four nucleic acid bases A, G, C, T, such that A and T correspond to the negative *x, y* axes, respectively, and G and C correspond to the positive *x* and *y* axes, respectively. An advantage of graphical representations of DNA is that it allows visual comparisons which are easier to make. One should, however, be aware of a loss of information inherent in such graphical representations. One of the limitations is that graphical form shows the "path" of the "travel" along the primary sequence but not the "history" of the travel. *Hence, we do not know when what parts of the graphical path were retraced.* At the top of Figure 2 we show a graphical representation of the first exon of the human β globin gene, at a higher magnification. The rest of Figure 2 shows the first exon of β globin gene of several other species for comparison. As we can see upon inspection qualitative

* Corresponding author phone: (218)720-4328; fax: (218)720-4330; e-mail: sbasak@nrri.umn.edu.
† National Institute of Chemistry.
‡ Iowa State University.
§ University of Minnesota at Duluth.
|| Indian Institute of Chemical Biology.
On leave from Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311.

10.1021/ci000034q CCC: $19.00
Published on Web 08/15/2000

Table 1. DNA of Length 1424 Listing Nucleic Bases in Human Beta Globin Gene[a]

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGG
TATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCA
TGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTC
TGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCTTGG
ACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGT
TATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGC
CTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTT
CAGGGTGAGTCTATGGGACCCTTGATGTTTTCTTTCCCCTTCTTTTCTATG
GTTAAGTTCATGTCATAGGAAGGGGAGAAGTAACAGGGTACAGTTTAG
AATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCG
TTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAAT
TCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATG
CCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATACATTAAG
TAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTG
GAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTC
TATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAATG
ATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGAT
AATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTG
CATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGC
TACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTC
TTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC
ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
TATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAATG
ATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGAT
AATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTG
CATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGC
TACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTC
TTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC
ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

[a] ID HSHBB − beta globin gene sequence extract: exons: **1**−**92**, **223**−**445**, **1296**−**1424**; introns: 93−222, 446−1295. SQ Hshbb.MK1 - - segment from 62205 to 63628 of HSHBB.

Table 2. First Exon of Beta Globin Gene for Eight Species Labeled A−H

A	**human beta-globin**	**92 bases**
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG		
GCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG		
B	**goat alanine beta-globin**	**86 bases**
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGG		
TGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG		
C	**opossum beta-hemoglobin beta M-gene**	**92 bases**
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGT		
CTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG		
D	**gallus gallus beta globin**	**92 bases**
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGG		
GCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG		
E	**lemur beta-globin**	**92 bases**
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGG		
GCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG		
F	**mouse beta-a-globin**	**93 bases**
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGG		
CAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG		
G	**rabbit beta-globin**	**90 bases**
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGG		
GCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC		
H	**rat beta-globin**	**92 bases**
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGG		
GAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG		

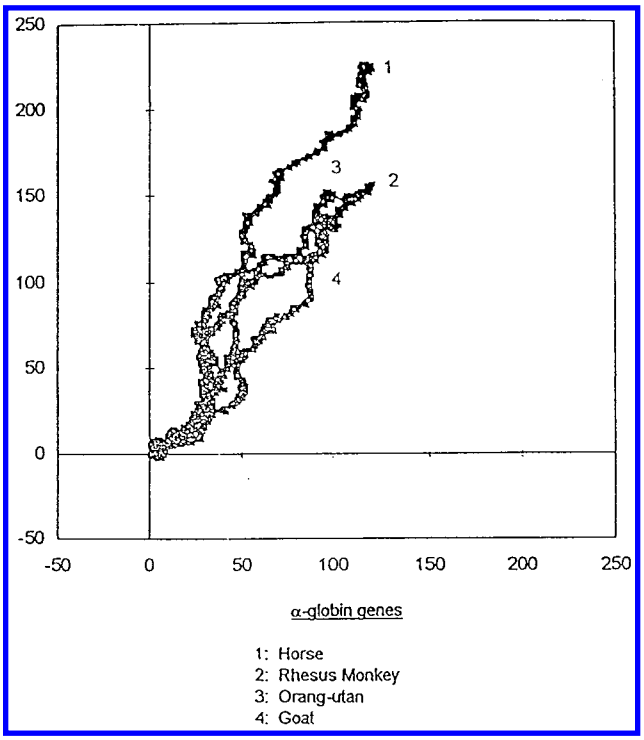

Figure 1. Few graphical representations of selected DNA that Nandy and collaborators developed.

similarities and differences between exons of different species are immediately apparent.

Mathematical curves can be represented in the form f(x, y) = 0, which corresponds to graphical projections of DNA of Figure 2, and in a parametric form $x = x(t)$ and $y = y(t)$. Clearly there is a loss of information in going from a parametric representation of a curve $x = x(t)$ and $y = y(t)$ to the analytical representation of the same curve. The f(x, y) = 0 only represents the path, while the former, if the parameter t is interpreted as time, gives the history of the movement over the path. Equally, there is loss of information when a a spatial curve is represented by its projection in the (x, y) plane (or any other plane). Hence, two routes for an improvement of graphical representations of DNA sequences appear possible: (1) to consider representation analogous to parametric representation of mathematical curves and (2)

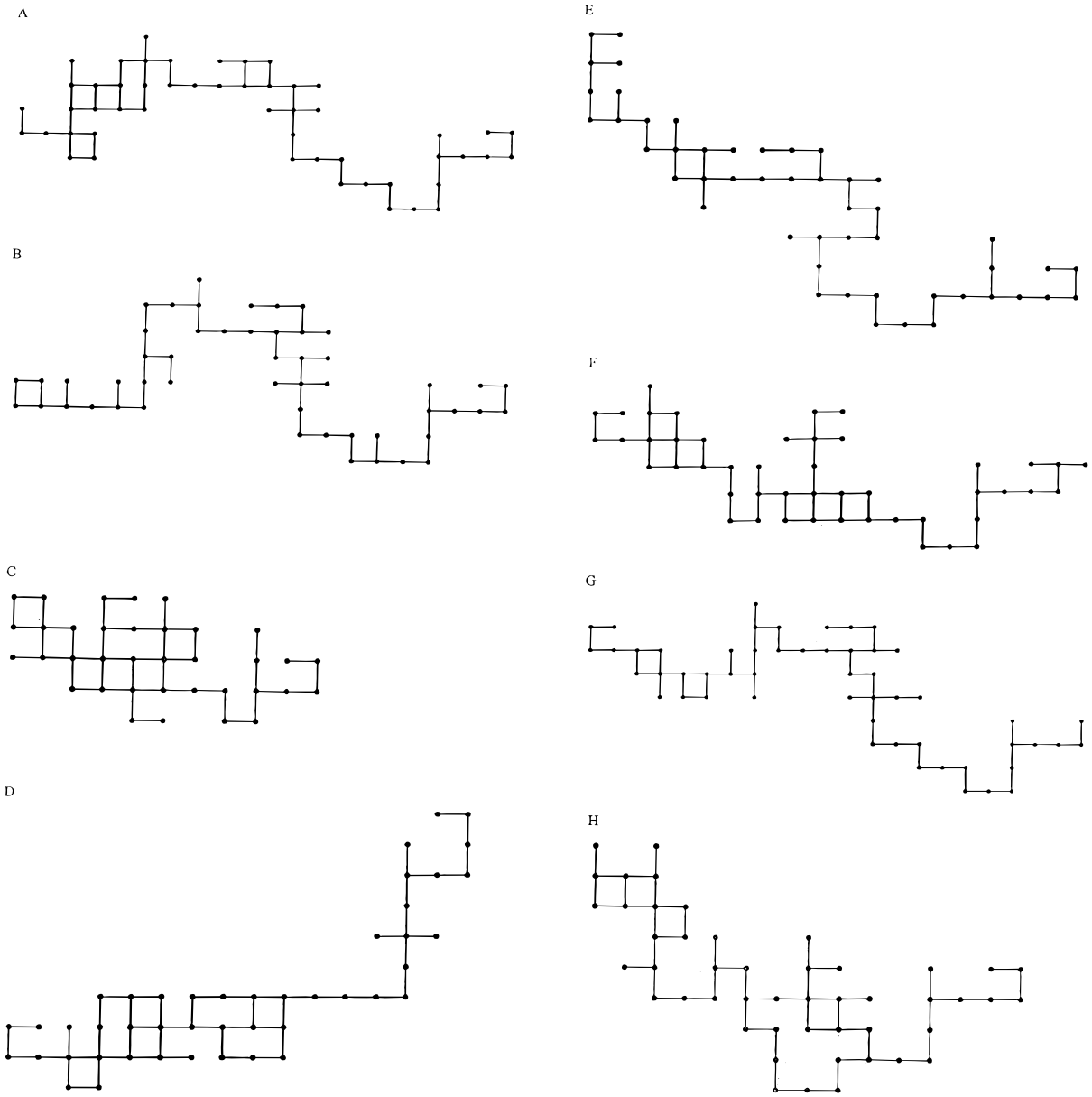

Figure 2. Graphical representations of the first exon of the human beta globin gene (top), a "detail" of Figure 1 and the remaining beta globin genes of Table 2.

to consider graphical representation of DNA sequence with "path" which is in traced in 3D space, rather than a plane. In this paper we will limit our attention to this latter problem. We will then describe a scheme which generates for a graphical spatial representation of DNA a numerical matrix. Once we arrive at a matrix representation of DNA we will search for suitable matrix invariants to be used for characterization of DNA. Finally we will consider possible condensation of derived numerical characterization of DNA in a more compact format.

3-D REPRESENTATION OF DNA PRIMARY SEQUENCE

Two-dimensional representation of DNA developed by Nandy[4] assigned to the four directions defined by the positive and the negative *x* and *y* coordinate axes to the four nucleic bases so that A and G are associated with the *x*-axis and C and T with the *y* axis. This assignment of directions differs from the assignment considered by Leong and Morgentaler,[14] who take a move to the right to correspond to A, a move to the left is C, an upward move is a T, and a downward move is G.

The nonequivalent directions are created after assignments of the first base because then there remains only one site that is *opposite* to the already selected direction; the other two sites are at *lateral* positions. If we could have three *equivalent* directions after the first assignment we would avoid considering the multiplicity of alternatives (projections). This is possible by using the directions defined by vertices of a regular tetrahedron. When looking from its center all the four directions toward the four vertices are equivalent, hence after selecting one direction the three directions remain equivalent. Hence, we will assign to four nucleic acid bases the four directions associated with the regular tetrahedron. To specify directions we will place the origin of the Cartesian (*x*, *y*, *z*) coordinate system in the center of a cube (Figure 3) so that the four corners of the cube,

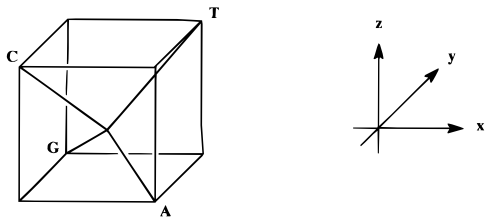

Figure 3. The tetrahedral directions assigned to A, G, C, T nucleic bases.

Table 3. Cartesian 3-D Coordinates for Initial Part of the Sequence of DNA Nucleic Bases of the First Exon

		x	*y*	*z*			*x*	*y*	*z*
1	A	+1	−1	−1	15	T	−1	+1	+1
2	T	+2	0	0	16	C	−2	0	+2
3	G	+1	+1	−1	17	C	−3	−1	+3
4	G	0	+2	−2	18	T	−2	0	+4
5	T	+1	+3	−1	19	G	−3	+1	+3
6	G	0	+4	−2	20	A	−2	0	+2
7	C	−1	+3	−1	21	G	−3	+1	+1
8	A	0	+2	−2	22	G	−4	+2	0
9	C	−1	+1	−1	23	A	−3	+1	−1
10	C	−2	0	0	24	G	−4	+2	−2
11	T	−1	+1	+1	25	A	−3	+1	−3
12	G	−2	+2	0	26	A	−1	0	−4
13	A	−1	+1	−1	27	G	−3	+1	−5
14	C	−2	0	0	28	T	−2	+2	−2

which define the tetrahedral directions, have the coordinates (+1, −1, −1), (−1, +1, −1), (−1, −1, +1), and (+1, +1, +1). To each tetrahedral direction we assign one nucleic base as follows:

$$(+1, -1, -1) \rightarrow \mathrm{A}$$

$$(-1, +1, -1) \rightarrow \mathrm{G}$$

$$(-1, -1, +1) \rightarrow \mathrm{C}$$

$$(+1, +1, +1) \rightarrow \mathrm{T}$$

The particular assignment is arbitrary, but this has no significance since all directions are equivalent. To obtain the spatial path associated with the DNA sequence, we move in *x*, *y*, *z* space in the direction that the above assignments dictates. Consider the beginning of the first exon of Table 1:

A T G G T G C A....

The first point of the spatial curve is at point (+1, −1, −1) which belongs to A, so directed from the origin. From that point we move in the direction assigned to T, (+1, +1, +1), which means that all the three coordinates of the position A, (+1, −1, −1), have to be increased by +1. We arrive then at the point (+2, 0, 0) as the location of T. From here we move in the direction defined by (−1, +1, −1) assigned to G telling that the first and the third coordinates have decreased while the second coordinate has increased. This leads to point (+1, +1, −1) as the location of G. Continuing in the direction of G we have again to decrease *x* and *z* (the first and the third coordinates) and to increase *y* (the second coordinate). Thus we come to the point (0, +2, −2). The process continues, each time we algebraically add the (*x*, *y*, *z*) coordinates of the new point to that of the last point. Continuation of this process is illustrated in Table 3 for the two dozen initial nucleic bases of the first exon. In Figure 4 we show a portion of 3-D graphical representation of DNA of Table 1.

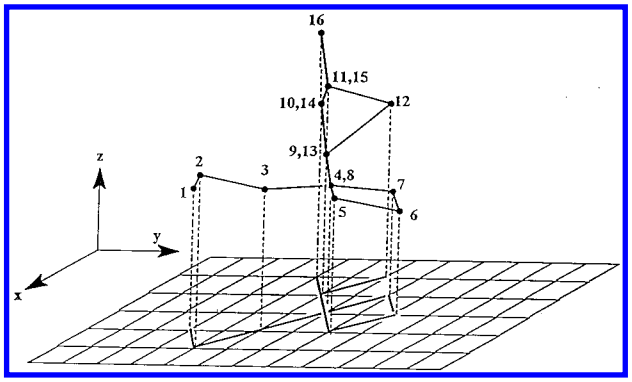

Figure 4. Portion of 3-D graphical representation of DNA of Table 1.

NUMERICAL CHARACTERIZATION OF SPATIAL REPRESENTATION OF DNA

An important advantage of graphical representations of DNA, both 2-D and 3-D, is the possibility to derive numerical characterizations for such mathematical objects. One way to arrive at numerical characterization of DNA is to associate with its graphical representation given by a curve in the space (or a plane) a matrix. Once we have a matrix we can use matrix invariants arrive at various numerical descriptors, rather than the visual description of the DNA sequence. This is analogous to the use of matrices associated with molecular graphs or molecular structure as a source for construction of topological indices rather than using molecular models (such as "sticks-and-balls or "space-filling" models) for their representation.[10]

Formally, there is no difference between a graphical "sequence chain" (in 2-D or 3-D space) or an actual polymer ("atom chain") in the space. Hence, we can transfer mathematical methods used for the characterization of molecules in structure−property and the structure−activity studies to numerical characterization of 3-D representations of the primary DNA sequence. This has been considered recently by Randić and collaborators[17] for 2-D graphical representation of DNA.

We should mention that one can also arrive at numerical descriptors that may be specific and sensitive to graphical form of a DNA without necessarily resorting to matrices. Thus, for example, Raychaudhury and Nandy[18] considered several geometrical parameters of DNA curves, such as, for example, end-to-end distance as DNA descriptors. Matrices, however, offer additional descriptors and richer characterization and can be manipulated by a computer, and one can take other advantages of linear algebra, rather than being confined to ordinary geometry.

Search for novel descriptors may be an endless project, just as this has been the case with mathematical descriptors that continue to be constructed for molecules. However, the art is in finding *useful* descriptors, and those that have plausible structural interpretation, at least within the model considered. Matrices have an additional advantage: they allow one to construct additional matrices by combining elements of different matrices as components. In this way one can arrive at additional descriptors for DNA. In this report we will confine our interest particularly to the graph

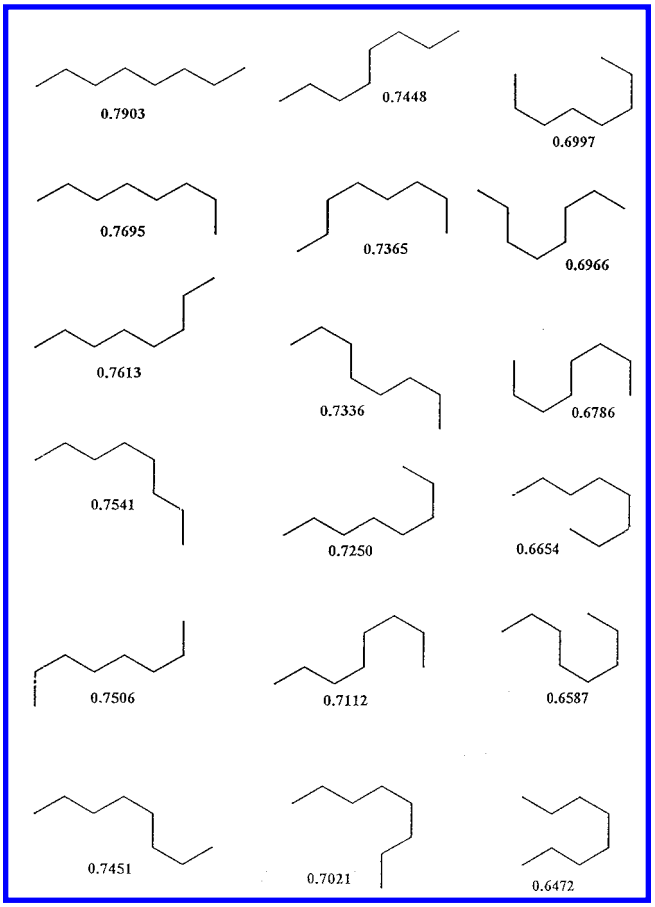

Figure 5. Conformations of eight-atom chain embedded on a graphite lattice ordered according to decreasing values of the leading eigenvalue of **D/D** matrix.

theoretical distance matrix and the Euclidean distance matrix for characterization of graphical forms of DNA.

MATRICES INVOLVING DISTANCES

The input information in a graph distance matrix[19,20] is solely confined to the information on the connectivity of the structure (system). However, when a graph is embedded in a space it assumes a fixed geometry. Then, in addition to the graph theoretical distance between a pair of vertices, we can also compute the Euclidean distances between the same pair of vertices. The Euclidean and the graph theoretical distances can be combined into a single distance/distance matrix by taking the quotient of the corresponding matrix elements.[21,22] Collection of such quotients for all pairs of vertices leads to the so-called **D/D** matrix. Matrices constructed in this way proved very promising as a tool for characterization of structures embedded in 3-D space. The normalized leading eigenvalue λ_1/n of a **D/D** matrix offers a measure of the degree of folding of a chainlike structure or a curve. In Figure 5 we illustrated configurations of an eight-atom C_8 chains embedded on a graphite lattice. Under each skeleton is given the normalized λ_1/n of **D/D** matrix. As we see the largest eigenvalue ($\lambda_1/8 = 0.7903$) is associated with the least bent *all-trans* configuration of C_8, and the smallest eigenvalue ($\lambda_1/8 = 0.6472$) belongs to the highly folded isomer TCCCT. T and C label stand for trans and cis conformations of three consecutive CC bonds (consult Table 4 for structures belonging to different labels). For chains of seven CC bonds even a smaller eigenvalue than 0.6472 is possible. It belongs to the hypothetical *all-cis* configuration CCCCC, the projection of which on hexagonal lattice gives a regular hexagon. In this structure the first and the last CC bond of C_8 would overlap, giving for $\lambda_1 = 4.6388$, which when normalized becomes $\lambda_1/8 = 0.5798$. The relative magnitudes of λ_1/n and the shape of corresponding conformations fully supports the interpretation of the normalized eigenvalue of **D/D** matrices as an index of the folding of a structure.

Table 4. Leading Eigenvalues for **D/D** Matrices of Eight-Atom Chains Embedded on a Graphite Lattice and the Leading Eigenvalues of the Corresponding Line Adjacency Matrices

conformer	λ_1/n of **D/D** matrix	λ_1/n of line adjacency matrix
TTTTT	0.7903	0.8571
TTTTC	0.7695	0.7191
TTTCT	0.7613	0.5916
TTCTT	0.7541	0.5208
CTTTC	0.7506	0.5858
TTCTC	0.7451	0.4688
TCTCT	0.7448	0.4019
TCTTC	0.7365	0.4748
CTCTC	0.7336	0.3836
TTTCC	0.7250	0.5793
TCTCC	0.7112	0.3773
TTCCT	0.7021	0.4464
CTTCC	0.6997	0.4533
TCCTC	0.6966	0.3538
CCTCC	0.6786	0.3375
TTCCC	0.6654	0.4426
CTCCC	0.6587	0.3347
TCCCT	0.6472	0.3347

A single descriptor, even though it may be instructive, offers but a limited characterization for a large system. Often additional descriptors are needed. They can be constructed by considering the so-called "higher order" **D/D** matrices.[23] These matrices are obtained by taking the powers of the quotients of two distances, rather than just using the quotients of the distances themselves. As a result we can derive for a *geometrical* (graphical-spatial) representation of DNA an *algebraic* characterization based on set of invariants, obtained by calculating the leading eigenvalue of the set of "higher order" matrices ${}^n\mathbf{D}/{}^n\mathbf{D}$. We will continue to use simplified notation **D/D** even though the **D** in the numerator stands for the Euclidean distances and the **D** in the denominator stands for graph theoretical distances.

D/D MATRICES FOR DNA

The Euclidean distance between bases in a 3-D graphical model of DNA are obtained from the 3-D coordinates of the nucleic bases listed in Table 3 using $\{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\}^{1/2}$, where x_i, y_i, z_i and x_j, y_j, z_j are the Cartesian coordinates of the points considered. To obtain the **D/D** matrix first we have to normalize the distance scale so that the Euclidean distance between adjacent vertices equals 1, not $\sqrt{3}$ (as a result of taking the side of cube to be equal 1). Then we have to divide each Euclidean distance with the number of bonds separating the two vertices to obtain the desired quotient of the two distances. In Table 5 we illustrate a part of the **D/D** matrix (corresponding to nine initial bases of DNA primary sequences of exon 1 of human β gene). The numerator combined with factor $1/\sqrt{3}$ gives the Euclidean distance between vertices i, j when the separation between adjacent bases is assigned distance 1, and

Table 5. Portion of the **D/D** Matrix for the First Exon of DNA of Table 1

0	1	$2/2\sqrt{3}$	$\sqrt{11/3}\sqrt{3}$	$4/4\sqrt{3}$	$\sqrt{27/5}\sqrt{3}$	$\sqrt{8/6}\sqrt{3}$	$\sqrt{11/7}\sqrt{3}$	$\sqrt{8/8}\sqrt{3}$
	0	1	$\sqrt{12/2}\sqrt{3}$	$\sqrt{11/3}\sqrt{3}$	$\sqrt{24/4}\sqrt{3}$	$\sqrt{19/5}\sqrt{3}$	$\sqrt{12/6}\sqrt{3}$	$\sqrt{11/7}\sqrt{3}$
		0	1	$2/2\sqrt{3}$	$\sqrt{11/3}\sqrt{3}$	$\sqrt{8/4}\sqrt{3}$	$\sqrt{3/5}\sqrt{3}$	$2/6\sqrt{3}$
			0	1	$2/2\sqrt{3}$	$\sqrt{3/3}\sqrt{3}$	0	$\sqrt{3/5}\sqrt{3}$
				0	1	$2/2\sqrt{3}$	$\sqrt{3/3}\sqrt{3}$	$\sqrt{8/4}\sqrt{3}$
					0	1	$2/2\sqrt{3}$	$\sqrt{11/3}\sqrt{3}$
						0	1	$2/2\sqrt{3}$
							0	1
								0

Table 6. Numerical Values for the Initial Portion of **D/D** Matrix and "Higher Order" **D/D** Matrices[a]

0	1	0.57735	0.63828	0.57735	0.60000	0.27217	0.27355	0.20412
		0.33333	0.40741	0.33333	0.36000	0.07407	0.07483	0.04167
		0.11111	0.16598	0.11111	0.12960	0.00549	0.00560	0.00174
		0.12345	0.02755	0.12345	0.01680	3.011−5	3.135−5	3.014−6
		1.524−4	7.590−4	1.524−4	2.821−4	9.064−10	9.831−10	9.085−12
	0	1	1	0.63828	0.70711	0.50332	0.33333	0.27355
				0.40741	0.50000	0.25333	0.11111	0.07483
				0.16598	0.25000	0.06418	0.12345	0.00560
				0.02755	0.06250	0.00412	1.524−4	3.135−5
				7.590−4	0.00391	1.696−5	2.323−8	9.831−10
		0	1	0.57735	0.63828	0.40825	0.20000	0.19245
				0.33333	0.40741	0.16667	0.04000	0.03704
				0.11111	0.16598	0.02778	0.00160	0.00137
				0.12345	0.02755	7.716−4	2.560−6	1.882−6
				1.524−4	7.590−4	5.954−7	6.554−12	3.541−12
			0	1	0.57735	0.33333	0	0.20000
					0.33333	0.11111		0.04000
					0.11111	0.12345		0.00160
					0.12345	1.524−4		2.560−6
					1.524−4	2.323−8		6.554−12
				0	1	0.57735	0.33333	0.40825
						0.33333	0.11111	0.16667
						0.11111	0.12345	0.02778
						0.12345	1.524−4	7.716−4
						1.524−4	2.323−8	5.954−7
					0	1	0.57735	0.63828
							0.33333	0.40741
							0.11111	0.16598
							0.12345	0.02755
							1.524−4	7.590−4
						0	1	0.57735
								0.33333
								0.11111
								0.12345
								1.524−4
							0	1
								0

[a] The first row is each box is the numerical value of **D/D** element, while the successive rows correspond to ^{2}**D**/^{2}**D**, ^{4}**D**/^{4}**D**, ^{8}**D**/^{8}**D**, and ^{16}**D**/^{16}**D**, respectively.

the denominator is the graph theoretical distance between the same two vertices.

The "higher order" **D/D** matrices are constructed by raising the elements of the **D/D** matrix (Table 5) to an ever increasing power. In Table 6 we show the corresponding entries of the higher order **D/D** matrices which are grouped into a single matrix where each row gives the numerical values corresponding to matrix elements of **D/D**, ^{2}**D**/^{2}**D**, ^{4}**D**/^{4}**D**, ^{8}**D**/^{8}**D**, and ^{16}**D**/^{16}**D**. As we can see all matrix elements that are smaller than one decrease as the exponents of the power increase. If one continues to raise exponents to even higher powers all the elements of ^{n}**D**/^{n}**D** matrix that are different from one would soon become very small and could be neglected. Hence, in the limit as $n \rightarrow \infty$ they are zero, and the resulting **D/D** matrix reduces to a binary matrix. In Table 7 we show the initial part of the limiting binary matrix $^{\infty}$**D**/$^{\infty}$**D** for the first exon of DNA of Table 1 again displaying only a 9×9 section. As we can see, all the elements above the main diagonal of the limiting matrix corresponding to adjacent sites in the DNA chain are necessarily equal to 1.

Table 7. Initial Portion of the Limiting (Symmetrical) Matrix of ^{n}**D**/^{n}**D** Matrix Truncated at $n = 16$[a]

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1										
2	1	0	1	1								
3		1	0	1								
4		1	1	0	1							
5				1	0	1						
6					1	0	1					
7						1	0	1				
8							1	0	1			
9								1	0	1	1	
10									1	0	1	
11									1	1	0	1
12											1	0

[a] Only zeros at the diagonal position are shown.

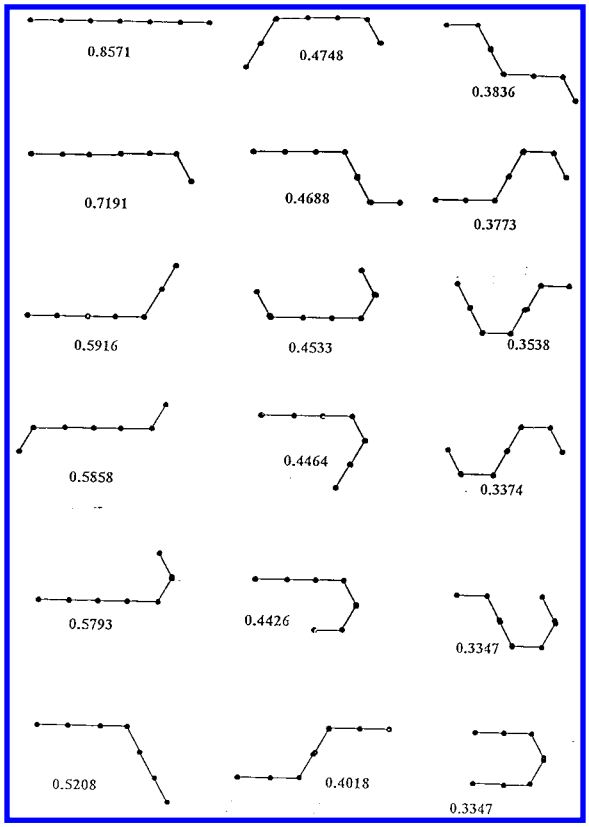

Figure 6. Conformations of line-adjacency graphs of eight-atom chains embedded on a graphite lattice ordered according to decreasing values of leading eigenvalue of line adjacency matrix. The order of isomers in Figure 5 and this figure is different.

However, entry 1 appears in addition at all sites associated with a repetition of the same nucleic base in the primary DNA sequence. For the first exon of Table 1 this happens at sites 3, 4 and 9, 10, and so on. When constructing the 3-D graphical model at these sites we continue to move in the *same* direction, and the corresponding segment of the 3-D graphical model forms a *line* segment. Hence, the elements of the limiting matrix indicate the so-called "line adjacency". The limiting matrix, referred to as the "line adjacency matrix",[24] is known in Graph Theory as the adjacency matrix of the Menger graph of a configuration.[25] For graphs of Figure 5 we show the corresponding Menger graphs. Their "line adjacency" matrix represents the limiting $^{\infty}$**D**/$^{\infty}$**D** matrices. They are also embedded in a plane because they have been derived from already embedded graphs.

A comparison of Figures 5 and 6 shows that line adjacency matrix carries *different* information than the **D**/**D** matrices from which it was algebraically constructed. The graphs in Figure 5 are ordered according to descending magnitudes of the normalized leading eigenvalue of the adjacency matrix, and the graphs in Figure 6 are ordered according to the leading eigenvalue of the limiting matrix. The resulting order is *different* from the order induced by the leading eigenvalue of **D**/**D** matrix. The leading eigenvalue of the limiting matrix can be viewed as an index of flexibility (or stiffness) of a structure, at least in some special cases.[24] Apparently structures with longer "line" segments have larger λ_1 or λ_1/n. When this is "translated" to the graphical representation relating to DNA sequences, the occurrence of "straight" segments corresponds to recurrence of the same base in a sequence repeatedly. Hence, DNA sequences with a larger number of repeating bases and longer such repeating segments will have a larger leading eigenvalue of the limiting binary matrix $^{\infty}$**D**/$^{\infty}$**D**.

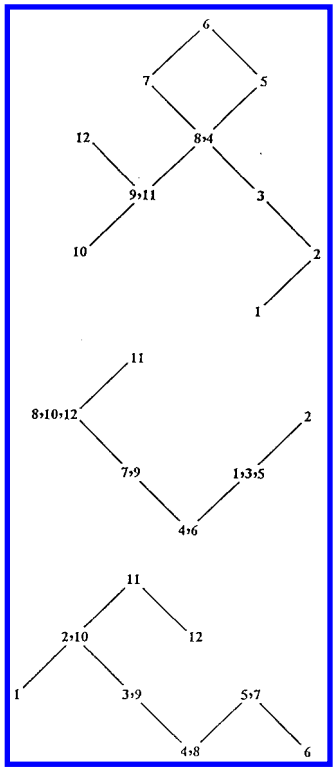

Figure 7. Projection of a portion of 3-D graphical representation of DNA of Figure 4 on the (x, y), (x, z), and (y, z) coordinate planes.

PROJECTIONS OF 3-D SPATIAL SEQUENCE REPRESENTATION

Spatial curves can be projected on coordinate planes (x, y), (x, z) or (y, z), or any plane, for that matter. The projections of 3-D spatial curves on each of the three coordinate planes is quite simple when coordinates of all the points are known. All that is needed is to ignore the coordinate perpendicular to the plane of the projection. Hence, for the first nucleic base of Table 1, A, with spatial coordinates (+1, −1, −1) we have for the projection on the x, y plane $x = 1$ and $y = -1$. For the projection of the same base on the x, z plane we have $x = 1$ and $z = -1$, while for the projection of the first nucleic base on the y, z plane we obtain $y = -1$ and $z = -1$. Hence, the projection coordinates can be read directly from Table 2 by ignoring one column, depending on the projection considered. In Figure 7 we show the three projections for the first 12 bases of exon of DNA of Table 1. It is interesting to observe that projection of the spatial 3-D representation of DNA on the (x, y) coordinate plane is identical with the 2-D graphical representation of Nandy[26,27] already depicted at the top of Figure 2. Hence, our 3-D visual representation of DNA contains automatically the 2-D graphical representation of Nandy as one of its projections. This, however, is not surprising, because if we project the four vertices of the tetrahedron having the coordinates (+1, −1, −1), (−1, +1, −1), (−1, −1, +1), (+1, +1, +1) on the (x, y) plane we obtain points (+1, −1), (−1, +1), (−1, −1), (+1, +1). The first set of points is associated with directions for A, G, C, T in 3-D as outlined in this paper, and the second set of points is associated with

directions for A, G, C, T in 2-D that coincides with that of Nandy if we rotate the coordinate system by −135°.

Similarly we find that the projection of the spatial 3-D representation of DNA on the (*x, z*) coordinate plane is identical with the 2-D graphical representation of Leong and Morgenthaler.[14] Hence, our 3-D visual representation of DNA contains alternative 2-D graphical representations as its projections. We may add that there is third yet the projection of 3-D graphical representation of DNA, the projection on the plane (*y, z*), that corresponds to the assignment of the four directions defined by the positive and the negative *x* and *y* coordinate axes to the four nucleic bases so that A and T are associated with the *x*-axis and C and G with the *y* axis. As we see from Figure 7 this projection differs from those of Nandy, Leong, and Morgenthaler and may have its own merits. Finally, we should add that one can consider projections of 3-D graphical curves of DNA on planes other than coordinate planes. While projections offer convenience of 2-D representation, all these projections are associated with some loss of information associated with the projection process.

Although the three projection paths of the 3-D representation of DNA are different, their limiting matrices are identical. This can be understood, because the form of the limiting matrix depends only on the repetition of same nucleic base in the primary sequence of DNA and that is independent of graphical representation of DNA and the projection process.

MATRIX INVARIANTS OF DNA

The search for a matrix representation of DNA primary sequence was motivated by desire to have numerical descriptors for DNA that are sequence invariants. Numerical characterization of DNA primary sequences will make comparisons of different DNA sequences much simpler than comparison based on alphabet symbols or the corresponding codes. Moreover, it will lead to quantitative measure of similarity and may open a novel method of characterizations for the same set of sequences. Matrices not only offer various inherent invariants as a tool for such comparisons but also allow one to consider modifications of matrix elements and in this way may further enrich the tool for comparative study of DNA. In this report we will continue to confine our attention to **D/D** matrix of DNA, but it will be clear that the outlined schemes are equally valid not only for the "higher order" **D/D** matrices but also for other matrices that one can associate with DNA.

Among numerous matrix (and graph) invariants we will consider first the average matrix element, which in the case of the graph theoretical distance matrix, except for normalization, is related to the Wiener number, a well-known graph theoretical invariant.[28,29] Alternatively one can consider the average row sum, which differs from the average matrix element and the Wiener number again only by normalization factor. The average row sum has an advantage, particularly when the individual row sums do not differ widely, because it may suggest an approximate value for the leading eigenvalue of the matrix. According to the Frobenius−Perron theorem of linear algebra the largest and the smallest row sums represent the upper and the lower bounds, respectively, for the leading eigenvalue (λ_1) of a symmetric matrix.[30] In Table 8 we have listed the upper bounds, the lower bounds, and the leading eigenvalue for truncated sequence of DNA for $n = 1$ to $n = 9$. Observe how closely the average row sum (given in the last column) approximates the leading eigenvalue, particularly for shorter segments of the matrix.

Table 8. The Upper Bounds, the Lower Bounds, the Leading Eigenvalue, and the Average Row Sums for Truncated Matrices of DNA

	row sum max	λ_1	row sum min	row sum average
1	0	0	0	0
2	1	1	1	1
3	2	1.732051	1.57735	1.718233
4	3	2.629245	2.21563	2.607815
5	3.63828	3.238402	2.79298	3.203444
6	4.34539	3.869193	3.39298	3.843783
7	4.84871	4.242930	3.09442	4.178791
8	5.18204	4.455833	2.71756	4.335833
9	5.45559	4.737987	3.49400	4.508241

Table 9. Average Matrix Element as a Function of Gradually Truncated **D/D** Matrix

	x, y, z	*x, y*	*x, z*	*z, y*
1	0	0	0	0
2	0.86603	0.70711	0.70711	0.70711
3	1.21424	1.07298	0.62854	1.07298
4	1.74711	1.52917	1.06066	1.52917
5	2.00204	1.82479	0.90510	1.82479
6	2.34274	2.16431	1.02138	2.16431
7	2.35303	2.25833	1.23982	2.07952
8	2.23832	2.11133	1.21440	1.97442
9	2.25630	2.13265	1.24376	1.89102
10	2.46497	2.24965	1.54132	1.94565
11	2.51032	2.19350	1.71264	2.03576
12	2.55077	2.23357	1.80319	2.00313
13	2.47111	2.15924	1.75222	1.92259
14	2.51231	2.20976	1.79277	1.89779
15	2.50930	2.12249	1.84061	1.92616
16	2.63879	2.14294	2.01366	2.04107

The leading eigenvalue of a matrix is an important matrix invariant. We have already mentioned that λ_1/n of the **D/D** matrix is an index of the folding of a structure, and λ_1/n of the limiting matrix can be viewed as an index of the flexibility of a system. Similarly, the λ_1 of the adjacency matrix and λ_1 of the path matrix represent alternative indices of (molecular) branching,[31,32] while λ_1 of the **D/DD** matrix, where **DD** represents the detour matrix,[33,34] is an index of the cyclicity of a system.[35,36] The average row sum may give a similar insight into a system as the leading eigenvalue. The average row sum, however, can be easily computed, while computation of eigenvalues of large matrices is more involved, and, of course, the DNA sequences could be very long. For example, the 1424 bases of Table 1, of which we considered the first exon only (92 bases), are a part of 73 326 base pairs.[37]

The average row sum, and also the average matrix element of a **D/D** matrix, will depend on the size of the matrix as is seen from Table 9 where under the heading *x, y, z* we have listed the average matrix element as a function of *n*, the size of the matrix at truncation of DNA sequence. The same was true for the leading eigenvalue of the truncated DNA sequences (Table 8).

The dramatic condensation of data illustrated above may be excessive for some more ambitious comparisons of DNA sequences. In such cases, one can, in addition to **D/D** matrix, also consider the leading eigenvalue or the average element

Table 10. Leading Eigenvalue of the **D/D** Matrix and Higher Order **D/D** Matrices for $n = 2$ to $n = 20$ Showing the Convergence for λ_1 and the Limit for $n \rightarrow \infty$

power	λ_1	power	λ_1
1	4.73797	12	2.35418
2	3.54855	13	2.35143
3	2.99558	14	2.34966
4	2.71223	15	2.34851
5	2.55903	16	2.34777
6	2.47313	17	2.34729
7	2.42349	18	2.34696
8	2.39409	19	2.34675
9	2.37629	20	2.34661
10	2.36537	limit	2.34631654447882
11	2.35850		

of ^{2}**D**/^{2}**D** matrix, of ^{3}**D**/^{3}**D** matrix, and so on. A dozen ^{n}**D**/^{n}**D** matrices can in this way offer a sufficient number of invariants for more extensive comparisons of DNA sequences. In Table 10 we report the leading eigenvalue for a 9×9 ^{n}**D**/^{n}**D** matrices for $n = 1$ to $n = 20$, which illustrate the "profile," the sequence of descriptors, for the particular fragment of DNA. As n increases the value of the leading eigenvalue λ_1 converges to a limiting value. The limit can be easily computed as it represents the leading eigenvalue of the binary matrix of the same size (here 9×9). Using so constructed "profiles" the calculation of the similarities of DNA sequences is transformed into a calculation of similarities of the corresponding numerical sequences of DNA descriptors, the task which is not computer intensive if compared to the similar studies using alignment methodologies. Of course, it yet remains to be investigated which set of invariants may offer optimal characterization for DNA comparisons and how sensitive are such "profiles" to minor changes in DNA composition. In a recent study in which the DNA sequence was characterized by average distances between various nucleic acid bases it was shown that the "distance profiles", constructed analogously to the here reported "leading eigenvalue profile", is very sensitive already when a single nucleic base has been changed (i.e., the case of mutation).[41]

CONCLUDING REMARKS

In this article we (1) outlined a construction of a 3-D "graphical" representation of DNA primary sequences, illustrated on a portion of the human β globin gene; (2) described a particular scheme that allows 3-D spatial representation of DNA to be transformed into a numerical matrix representation; (3) illustrated derivation of a set of matrix invariants from the matrix representation of DNA; and (4) suggested a relative simple data reduction based on statistical analysis of generated DNA matrix invariants. Each of the four contributions, in our view, not only will facilitate comparative studies of DNA but also open possibilities for further developments of condensation of primary DNA sequence information. The outlined 3-D representation, for example, can be modified by use of the sequential labels as the fourth coordinate in order to avoid 3-D spatial curves overlap itself. The numerical matrix characterization offers many alternatives, from the use of different distance measures to the use of different matrix forms. In addition to the possibility of selecting matrix invariants, which is almost unlimited, we have the possibility of selecting different matrices to start the process of condensation of data. Hence, we anticipate here an expansion, if not explosion, of alternatives that may parallel the expansion of the topological indices proposed for the characterization of molecular structure−property-activity relationships and introduction of novel matrices for chemical graphs. The most significant aspect considered in this contribution may turn out to be the data reduction step when a large number of input data are condensed into a substantially smaller set of derived parameters. This important aspect of DNA data analysis has only recently received some attention,[38−40] but, in view of the exponential growth of the automated DNA sequencing techniques, the problem of digesting novel information, no doubt, will require novel ideas that go beyond just listings of nucleic bases of a primary sequence. The construction of sequence "profiles", illustrated in this report, may be one way of data reduction, in addition to the recently proposed grouping of data for different nucleic acids separately, which allow large ($n \times n$) matrices (where n can run into the hundreds or the thousands) to be condensed to small (4×4) matrices where the rows and the columns are associated with the four nucleic bases A, G, C, and T. Needless to say that the outlined approach is suitable for characterization of local fragments of DNA, which is precisely how one may look on the truncated DNA fragment considered in this work.

ACKNOWLEDGMENT

We thank Professor Ch. Rücker (Freiburg, Germany) for critical reading of the manuscript and numerous suggestions that lead to improved presentation of the material.

REFERENCES AND NOTES

(1) Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernet. Control Theor.* **1966**, *10*, 707−710.
(2) Sankoff, D. Matching sequences under deletion-insertion constraints. *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *68*, 4−6.
(3) Kruskal, J. B. An overview of sequence comparison. In *Time wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*; Sankoff, D., Kruskal, J. B., Eds.; Addison-Wesley: London, 1983; pp 1−40.
(4) Waterman, M. S. General methods of sequence comparison. *Bull. Math. Biol.* **1984**, *46*, 473−500.
(5) Smith, T. F.; Waterman, M. S. Comparison of biosequences. *Adv. Appl. Math.* **1981**, *2*, 482−489.
(6) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.
(7) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(8) Jerman-Blazic, B.; Fabič, I.; Randić, M. Comparison of sequences as a method for evaluation of the molecular similarity. *J. Comput. Chem.* **1986**, *7*, 176−188.
(9) Jerman-Blažič, B.; Fabič, I.; Randić, M. Application of string comparison techniques in QSAR Studies. In *QSAR in Drug Design and Toxicology;* Hadzi, D., Jerman-Blažič, B., Eds.; Elsevier Sci. Publ.: Amsterdam, The Netherlands, 1987; pp 52−54.
(10) Randić, M. *Topological indices, The Encyclopedia of Computational Chemistry;* Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley and Sons: Chichester, 1998; pp 3018−3032.
(11) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.
(12) Randić, M.; Basak, S. C. Variable molecular descriptors. In *Some Aspects of Mathematical Chemistry;* Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; to be published by Visva Bharati University, Santiniketan, West Bengal, India
(13) Roy, A.; Raychaudhury, C.; Nandy, A. A novel techniques of graphical representation and analysis of DNA sequences − A Review. *J. Biosci.* **1998**, *23*, 55.
(14) Leong, P. M.; Mogenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *12*, 503−511.

(15) Hamori, E. Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques.* **1989**, *7*, 710−720.
(16) Nandy, A. New graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309.
(17) Randić, M.; Nandy, A.; Basak, S. C. On numerical characterization of DNA primary sequences. *J. Math. Chem.*, submitted for publication.
(18) Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.
(19) Harary, F. *Graph Theory;* Addison-Wesley: Reading, MA, 1969.
(20) Buckley, F.; Harary, F. *Distance in Graphs;* Addison-Wesley: Reading, MA, 1990.
(21) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277.
(22) Randić, M.; Razinger, M. *On characterization of 3D molecular structure*, in: *From Chemical Topology to Three-Dimensional Geometry;* Balaban, A. T., Ed.; Plenum Press: New York, 1997.
(23) Randić, M.; Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.
(24) Randić, M.; Vračko, M.; Novič, M. *Eigenvalues as molecular descriptors*, in: *QSAR/QSPR by Molecular Descriptors;* Diudea, M. V., Ed.; Nova Publ.: in press.
(25) Coxeter, H. S. M. *Bull. Am. Math. Soc.* **1950**, *56*, 413.
(26) Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci.* **1996**, *70*, 661−668.
(27) Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for maromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.
(28) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(29) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, Fl, 1992.
(30) Gantmacher, F. *Theory of Matrices;* Chelsea Publ: New York, 1959; Vol. II, Chapter 13.
(31) Randić, M. On molecular branching. *Acta Chim. Slovenica* **1997**, *44*, 57−77.
(32) Randić, M. J. On structural ordering and branching of acyclic saturated hydrocarbons. *Math. Chem.* **1998**, *24*, 345−358.
(33) Amić, D.; Trinajstić, N. On the detour matrix *Croat. Chem. Acta* **1995**, *68*, 53−62.
(34) Trinajstić, N.; Nikolić, S.; Lučuć, B.; Amić, D.; Mihalić, Z. The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 631.
(35) Randić, M. J. On characterization of cyclic structures. *Chem. Inf. Comput. Sci.* **1997**, *37*, 1063.
(36) Pisanski, T.; Plavšić, D.; Randić, M. On numerical characterization of cyclicity *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 520−523.
(37) EMBL Nucleic Bases Sequence Database (rel. 31) ID HSHBB Accession number V01317.
(38) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.
(39) Randić, M.; Vračko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.
(40) Randić, M. On characterization of DNA primary sequences by a condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34.
(41) Randić, M.; Basak, S. C. Characterization of DNA based on the average distances between the nucleic acid bases. *J. Chem. Inf. Comput. Sci.*, submitted for publication.

CI000034Q