On the Similarity of DNA Primary Sequences

Milan Randić* and Marjan Vračko

National Institute of Chemistry, 1001 Ljubljana, POB 3430, Slovenia

Received August 15, 1999

We consider numerical characterization of graphical representations of DNA primary sequences. In particular we consider graphical representation of DNA of β-globins of several species, including human, on the basis of the approach of A. Nandy in which nucleic bases are associated with a walk over integral points of a Cartesian x, y-coordinate system. With a so-generated graphical representation of DNA, we associate a distance/distance matrix, the elements of which are given by the quotient of the Euclidean and the graph theoretical distances, that is, through the space and through the bond distances for pairs of bases of graphical representation of DNA. We use eigenvalues of so-constructed matrices to characterize individual DNA sequences. The eigenvalues are used to construct numerical sequences, which are subsequently used for similarity/dissimilarity analysis. The results of such analysis have been compared and combined with similarity tables based on the frequency of occurrence of pairs of bases.

INTRODUCTION

DNA sequencing has become a routine and has resulted in an abundance of data on primary sequences of DNA for various species. Hence, we face the task of "digesting" such overwhelming data, which poses a number of yet unresolved problems. Sequence comparison has been considered in the literature. For example, in molecular biology such comparisons are used to resolve the questions of the homology of macromolecules.[1] Another field of application of sequence comparison is in the area of codes and error control,[2] as well as the application of comparison of computer files.[3] The string comparison technique has also been used in chemistry to measure the molecular similarity of chemical structures.[4] What is different and distinct with DNA data is that first and foremost, DNA primary sequences vary enormously in their length. They can consist of fewer than a hundred bases but can extend to over a hundred thousand bases. Even when the long sequences are broken down into segments corresponding to exons, or introns, the segments corresponding to the same position within a gene and belonging to different species may have different length. In Table 1 we show the first exon of β-globin genes belonging to eight species (including human), which differ in length from 86 to 93 bases. Clearly, a base by base comparison could not be used when sequences have different lengths. The search for optimum correspondence between sequences as outlined by Kruskal[5] involves different operations, such as trace (linking the same elements in the two sequences), alignment, or matching (spacing elements using blank spaces between). Finding the smallest number of changes (deletions, insertion, substitutions, shifts) that are necessary to match labels in two sequences is far from trivial.

An alternative approach is to consider mathematical invariants of codes (sequences) rather than codes themselves. In the case of chemical structures and chemical graphs, it is not difficult to arrive at numerous invariants that can be used for characterization and comparison of structures. There are hundreds of topological indices that have been used in structure-property-activity studies based on molecular graphs.[6−9] Construction of invariants for sequences, however, has not yet received much attention. One way to arrive at invariants for sequences is to associate a matrix with a sequence. Once we have matrix representation of a sequence, one can consider suitable matrix invariants as invariants of the sequence. This approach has been outlined recently on a primary DNA sequence of human β-globin, the sequence to be considered also in this report. One can consider a full-matrix representation of DNA, which may contain redundant

Table 1. The First Exon of the β-Globin Gene of Different Species

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		

* Corresponding author. On leave from Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311. Present address: 3225 Kingman Rd, Ames, IA 50014. Fax 51 292 8629; e-mail: milan.randic&drake.edu.

10.1021/ci9901082 CCC: $19.00
Published on Web 05/05/2000

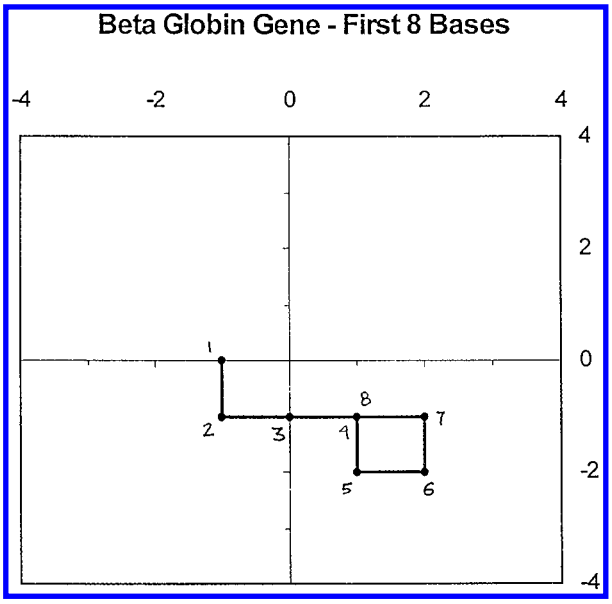

Figure 1. Nandy's graphical plot for the first eight nucleic bases of the first β-globin of Table 1.

information, or alternatively one may consider a condensed-matrix representation, which may be information-deficient.[10,11] The advantage of approaches based on matrices become apparent when one is interested in comparison of primary DNA sequences. Rather then using sequences directly, one then considers invariants associated with *n*-dimensional matrices.

Comparison and evaluation of similarity/dissimilarity of structures (represented by *n*-dimensional vectors) has been described in the literature.[12−14] These vectors have as components structural invariants extracted from matrices associated with DNA, thus formally transcribing the original primary DNA sequence in a numerical sequence, the length of which depends on the nature of invariants selected, and which can be modified if desired.

CONSTRUCTION OF A *D*/*D* MATRIX FOR A DNA SEQUENCE

Several researchers have considered graphical representations for DNA sequences, in particular, Nandy,[15−17] Leong and Morgenthaler,[18] Hamori,[19,20] and others. A review by Roy, Raychaudhury, and Nandy[21] gives a comprehensive study of the graphical representation methods and their applications in viewing and analyzing DNA sequences. We will consider in more detail the graphical representation of DNA as introduced by Nandy. His approach is based on graphical forms illustrated in Figures 1 and 2, which represent a path in the (*x*, *y*)-coordinate system. The four directions along the positive and negative sides of the coordinate axes are assigned to the four nucleic bases, and the primary sequence of DNA is then "translated" into corresponding movements in the *x*, *y*-plane. By convention Nandy associated with A and G the negative and the positive directions along the *x*-axis, while C and T are similarly associated with the negative and the positive directions along the *y*-axis. Thus if we start at the origin of the coordinate system, the sequence ATGGTGCACC..., which is the beginning of the first exon in the top of Table 1, is represented by steps (−1, 0), (−1, −1), (0, −1), (1, −1), (1, −2), and so on, shown in Figure 1. The search for invariants of a sequence of labels is now transformed into a search for invariants of a graphical plot of points.

There is formal similarity between the characterization of DNA by a graphical walk in the *x*, *y*-plane and characterization of curves, mathematical or physicochemical, embedded in a plane or in three-dimensional space. By using invariants instead of the sequence of codes, the length of sequence is no longer of primary importance. This has an important advantage as now we are in a position to directly compare sequences of different length, without some prior preprocessing. The problem has thus shifted from computational (matching sequences) to conceptual: how to arrive at suitable invariants to characterize graphical sequences. However, formally whether one considers a real polymer as a curve embedded in 3-D space or a fictitious curve similarly embedded, such as the curves of Figures 1 and 2 that are embedded in 2-D space, from the mathematical point of view is irrelevant. Both type of curves may follow a similar numerical characterization. It is precisely this point of view that allowed us to represent a DNA sequence by a sequence of invariants. Considerable progress has been made in recent years in extending graph theoretical characterizations to structures having rigid 3-D geometry.[22] We will base our approach on some of these novel developments in the characterization of 3-D structures.

With graphs as chemical objects, besides the adjacency matrix which is binary and in which an entry "one" indicates neighbors, other elements being zero, and the distance matrix [23] in which entries indicate the length of the shortest distance (intervening bonds) between any pair of atoms, a dozen or more matrices have been introduced in chemical graph theory in recent years. These matrices register different structural features of a molecule. For example, the Wiener matrix [24] is constructed by considering for any pair of vertices in a graph the product of atoms on each side of the *path* separating such atoms. This represents a generalization of the procedure that Wiener described over 50 years ago for construction of the Wiener number,[25] one of the well-known and widely studied molecular invariants. The Hosoya matrix [26,27] is based on partitioning of the Hosoya topological index[28] (which counts disjoint bonds in a structure) and its generalization (which counts disjoint paths in a structure). The elements of the detour matrix, which originates with Harary [23] but which has only recently received some attention in chemical literature, indicate the longest paths between pair of vertices in a graph.

These and other matrices offer alternative characterizations of structures and serve as a source of novel invariants for chemical graphs. However, we are interested in graphs embedded in a 2-D or 3-D space. A geometry matrix, based on the *x*, *y*, *z*-coordinates of a structure, plays here an important role attributed to the adjacency matrix of a graph. Its elements are given by the Euclidean distance between a pair of points in the space. Although the geometry matrix incorporates the essential information on molecular size, shape, and form, such matrices do not directly involve information of connectivity. While this can be inferred frequently in the case of chemical structures by searching for the shortest separations, this is no longer the case with mathematical curves such as curves representing DNA sequences. An easy way to see is to consider four points at

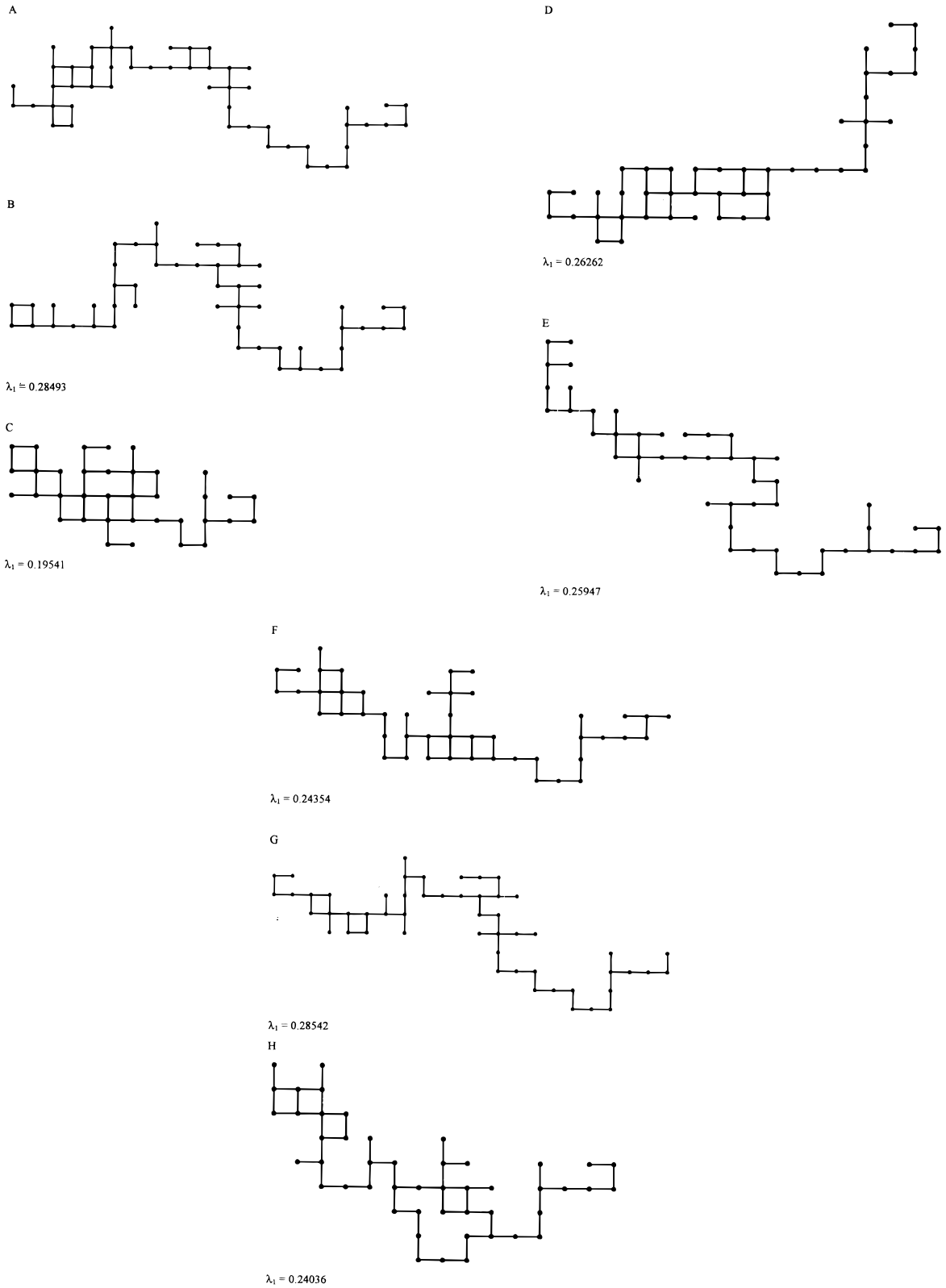

Figure 2. Graphical representations for the eight DNA sequences of Table 1.

the corners of a square. Their positions in the space fully determine the geometrical matrix, but such points can represent besides a four-member ring also an open chain of four points, or even disjoint fragments. This illustrates well that geometry matrices alone would be of little use for characterization of DNA sequences based on their graphical representation. We need to combine *geometrical* information with *connectivity* information. This has been proposed some time ago by Randić and co-workers[29] and implemented in the so-called *D*/*D* matrices. They combined the information on geometrical distances in a structure with the information on the connectivity, which can be registered by considering graph theoretical distances between points in a chainlike structure of graphical structures such as those of Figures 1 and 2. As a result one obtains the *D*/*D* matrix, the elements of which are given as the quotient of the Euclidean distance

between a pair of points (D_E) and the graph theoretical distance between the same two points (D_G). Because $D_E \leq D_G$ the elements of the corresponding *D*/*D* matrix are always smaller than 1, or at most equal to 1. As a consequence, one can construct closely related matrices D^k/D^k, obtained from the *D*/*D* matrix by raising each element separately to the *k*th power, thus increasing the number of the invariants for characterization of structures and getting more complete characterization of the object considered. In this way the descriptors for 3-D structures called *molecular profiles* have been derived.[30−34] The molecular profile approach is quite general and applies also to structures of arbitrary shape as would be molecular models that use spherical intersecting balls instead of a balls-and-sticks representation for atoms and bonds.[5] The generality of the approach has already been demonstrated by characterization of more general mathematical curves having fractal features,[36] folded models of small proteins,[37] and characterization of the degree of folding of such structures by numerical measure.[38]

EIGENVALUES AS DESCRIPTORS

Eigenvalues of a matrix are one of the best known matrix invariants. If a matrix is symmetric, as is the case with all the matrices considered here, the eigenvalues are real. A set of eigenvalues can be viewed as a characterization of a structure, but as is well-known such characterization is not unique.[39] In other words, different graphs and different structures may have the same set of eigenvalues. Such graphs, known as isospectral, have received considerable attention in mathematics,[40−42] physics,[43,44] and chemistry,[45−52] of which we only indicated some earlier contributions. While it was initially thought that the complete coincidence of all eigenvalues may be an exception rather than a rule, the subsequent research revealed that isospectral graphs are more a rule than an exception.[53] That, however, does not diminish their utility, although they would fail to discriminate structures in testing for isomorphism.

The magnitudes of individual eigenvalues are bounded by the largest and the smallest row (or column) sums of the matrix considered, as stated by the theorem of Frobenius and Perron.[54] In particular, the largest positive eigenvalue, the eigenvector of which has no nodal surfaces, can at best be equal to the largest row sum (or column sum) of a matrix, or be smaller. Hence, for matrices, the elements of which do not vary dramatically, the average row sum, or the largest row sum, can approximate the leading eigenvalue.[55] This is of some interest when considering large matrices as it is much simpler to calculate row sums than matrix eigenvalues, and it may be more economical particularly when one is interested only in the leading eigenvalue of a matrix.

Among all eigenvalues the leading eigenvalue of a matrix, λ_1, often plays a special role. In the case of the adjacency matrix of trees, Lovasz and Pelikan [56] suggested the leading eigenvalue λ_1 as an index of molecular branching. More recently it was shown that the leading eigenvalue of a substituted path matrix, $\lambda\lambda_1$, gives even better characterization of molecular branching.[57−60] The leading eigenvalue of the *D*/*D* matrix has been interpreted as an index of folding of a structure, and the leading eigenvalue of the so-called line-adjacency matrix, known in the mathematical literature as the Menger graph of a configuration,[61] apparently represents flexibility of a structure.[62] Finally, the leading eigenvalue of the *D*/*DD* matrix, where *D* represents the graph theoretical distance matrix and DD the corresponding detour matrix, has been suggested for characterization of cyclic structures.[63,64] In view of these interpretations of the leading eigenvalues, it is of interest to see if the leading eigenvalue λ_1, and eigenvalues in general, of matrices associated with a graphical representation of DNA can be useful for characterization of DNA, and particularly if they can be used for computing similarity between different DNA sequences.

Table 2. All the Eigenvalues of the First DNA Sequence of Table 1

1	0.27136	24	0.00196	47	−0.00852	70	−0.01282
2	0.07010	25	0.00071	48	−0.00886	71	−0.01303
3	0.06208	26	0.00063	49	−0.00912	72	−0.01319
4	0.04361	27	−0.0009	50	−0.00944	73	−0.01335
5	0.03325	28	−0.00069	51	−0.01001	74	−0.01369
6	0.03236	29	−0.00109	52	−0.01008	75	−0.01416
7	0.02842	30	−0.00215	53	−0.01013	76	−0.01447
8	0.02563	31	−0.00242	54	−0.01045	77	−0.01476
9	0.02429	32	−0.00292	55	−0.01084	78	−0.01560
10	0.02071	33	−0.00319	56	−0.01088	79	−0.01568
11	0.01820	34	−0.00336	57	−0.01115	80	−0.01589
12	0.01565	35	−0.00395	58	−0.01131	81	−0.01633
13	0.01527	36	−0.00476	59	−0.01149	82	−0.01695
14	0.01438	37	−0.00509	60	−0.01164	83	−0.01745
15	0.01314	38	−0.00534	61	−0.01180	84	−0.01827
16	0.01070	39	−0.00595	62	−0.01195	85	−0.01842
17	0.00995	40	−0.00619	63	−0.01202	86	−0.01941
18	0.00931	41	−0.00631	64	−0.01208	87	−0.02077
19	0.00678	42	−0.00657	65	−0.01211	88	−0.02149
20	0.00650	43	−0.00728	66	−0.01224	89	−0.02249
21	0.00596	44	−0.00776	67	−0.01227	90	−0.02367
22	0.00451	45	−0.00802	68	−0.01259	91	−0.02493
23	0.00349	46	−0.00824	69	−0.01278	92	−0.02688

D/*D* MATRICES FOR DNA

The *D*/*D* matrix is constructed for graphs embedded in space, hence, a structure having rigid geometry, and does not exist for ordinary molecular graphs that are devoid of rigid geometrical form. The elements of this matrix in a way measure the compactness of the structure because if the structure is extended the average element of the matrix will be relatively large in comparison with a structure that is highly compact and takes smaller space. As already mentioned, the entries in a matrix represent the quotient of the spatial distance and the distance measured along the bond in a structure. Numerical illustration of the *D*/*D* matrix for a short segments of DNA has been discussed at some length elsewhere.[65] In this contribution we will consider DNA sequences of Table 1, the length of which varies between 86 and 93 bases. We will refer to these sequences by labels A−H. For each of the sequence we have first constructed a *D*/*D* matrix and then calculated its eigenvalues. The *D*/*D* matrices are symmetric, but dense;[66] that is, except for the diagonal entries it has but few zero elements. The nondiagonal zero entry occurs only when the graphical path representing DNA crosses itself.

The *D*/*D* matrices for DNA fragments of Figure 1 are of the order 100 × 100. Their eigenvalues have been calculated with the computer program written by one of the coauthors (M.V.), which is based on a reduction of the matrix to the tridiagonal form, known as the Givens and Householder reduction.[67] In Table 2 we have listed all the eigenvalues of

Table 3. The Leading Five Eigenvalues of DNA Sequences of Table 1

β-globin	λ_1	λ_2	λ_3	λ_4	λ_5
A	0.27136	0.07010	0.06208	0.04361	0.03315
B	0.28493	0.08166	0.04139	0.04090	0.03702
C	0.19541	0.09635	0.06126	0.05568	0.03553
D	0.26262	0.08191	0.06037	0.04722	0.03388
E	0.25947	0.03898	0.06710	0.04082	0.03590
F	0.24354	0.09458	0.05760	0.05484	0.03481
G	0.28542	0.07432	0.05257	0.04272	0.03858
H	0.24036	0.09900	0.07022	0.04890	0.03780

the *D*/*D* matrix for the first DNA sequence of Table 1 in order to illustrate the fact that except for the first eigenvalue all other eigenvalues of *D*/*D* matrices of graphical representation of DNA of Nandy are visibly smaller. Hence, λ_1 will play the dominant role in characterization of DNA while other eigenvalues will have a lesser influence, particularly when considering similarity analysis based on the eigenvalues. Because of this we decided to truncate the list of eigenvalues and limit the attention to the first eigenvalue and the several following largest values.

In Table 3 we have listed the leading eigenvalues of the *D*/*D* matrix for the eight DNA sequences of Table 1. We could have included a few more eigenvalues, but the overall picture would hardly change because the differences between successive eigenvalues steadily decrease, and after a few initial steps they become too small to make a difference.

CHARACTERIZATION BY THE LEADING EIGENVALUE

The leading eigenvalue of the *D*/*D* matrix, as has been mentioned earlier, gives a measure of the degree of folding of long chains. The smaller the value of λ_1, the more folded the corresponding graphical representation of DNA. It follows therefore that among the eight β-globins the opossum β-hemoglobin (β-M-gene DVHBBB), sequence C of Table 3, is the most folded. On the other hand, the graphical representations of rabbit β-globin (sequence G) and the goat β-globins (sequence B) are the least folded of the eight sequences considered. In Figure 2 we show the projections, that is, the paths, of the graphical representations, for the eight β-globins of Table 1. From Figure 2, already by visual inspection, one observes that opossum β-globin representation is considerably different from the others. It is confined within a 10×4 rectangle, which in comparison to other graphical representations occupies a rather compact space. Equally, it can be seen from Figure 2 that the β-globins of goat and rabbit appear most extended. This fully agrees with these sequences having the largest *D*/*D* leading eigenvalues, which means that they are the least folded structures. Thus, we see that indeed the leading eigenvalue of the *D*/*D* matrix is a good index of the degree of folding even for a structure, such as the sequences of DNA considered, that intersects itself.

The diagrams of Figure 2 show paths over which one walks, as one follows the sequence of individual bases in DNA. The distinction between the path representation and the walk representation is quite analogous to the representation of a curve in analytical geometry as $F(x, y) = 0$, and the parametric representation of the same curve as $x = x(t)$ and $y = y(t)$, the latter of which corresponds to the walk representation. Because elements of the *D*/*D* matrix correspond to walks, the projections of DNA that are similar may still have considerably different leading eigenvalues. For example, the path E (lemur) and G (rabbit) appear similar, but the corresponding leading eigenvalues are not so similar, $\lambda_1 = 0.259\,47$ and $\lambda_1 = 0.285\,42$, respectively. On the other hand, mouse (F) and rat (H) β-globins have close magnitudes of the leading eigenvalues; the values are $\lambda_1 = 0.243\,54$ and $\lambda_1 = 0.240\,36$ respectively, but the path projections show different form. Clearly *D*/*D* matrix elements are sensitive not only to the form of the path projection but also to the pattern of the walk over such paths. For example, two walks shown in Figure 3 both have the same projection, a square path. One walk represents a cyclic walk over the square path, while the other involves several repetitious steps. The corresponding *D*/*D* matrices (shown in Table 4) have different forms. For each row the row sum of the first matrix is greater than the corresponding row sum for the second matrix. Consequently, the leading eigenvalues of the second matrix will be smaller, suggesting that the walk with repeating steps is more folded, which agrees with expectations.

Figure 2 well illustrates the advantages of visual representation of the DNA sequence of Nandy. On the other hand,

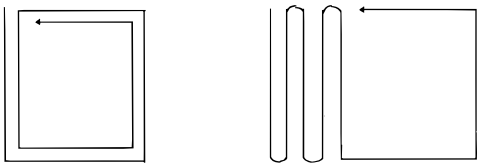

Figure 3. Alternative walks over a square path.

Table 4. *D*/*D* matrices and the Row Sums for the Two Folded Curves of Figure 3 Having the Same Path Projection

	1	2	3	4	5	6	7	8	Row sum
1	0	1	0	1/3	0	1/5	√2/6	1/7	1.91189
2		0	1	0	1/3	0	1/5	√2/6	2.76904
3			0	1	0	1/3	2/4	1/5	3.03333
4				0	1	0	1/3	2/4	3.16667
5					0	1	√2/2	1/3	3.37377
6						0	1	√2/2	3.24044
7							0	1	3.97614
8								0	3.11900

	1	2	3	4	5	6	7	8	Row sum
1	0	1	√2/2	1/3	0	1/5	√2/6	1/7	2.61900
2		0	1	√2/2	1/3	0	1/5	√2/6	3.47614
3			0	1	√2/2	1/3	0	1/5	3.94755
4				0	1	√2/2	1/3	0	4.08088
5					0	1	√2/2	1/3	4.08088
6						0	1	√2/2	3.94755
7							0	1	3.47614
8								0	2.61900

Table 5. (Top) Similarity/Dissimilarity Tables for the Eight DNA Sequences of Table 1 Based on the Leading Eigenvalue (the Values above the Main Diagonal) and Two Leading Eigenvalues (the Values below the Main Diagonal); (Bottom) Similarity/Dissimilarity Tables for the Eight DNA Sequences of Table 1 Based on the Three Leading Eigenvalues (the Values above the Main Diagonal) and Four Leading Eigenvalues (the Values Below the Main Diagonal)

	A	B	C	D	E	F	G	H
A	0	0.00136	0.0760	0.0087	0.0119	0.0278	0.0141	0.0310
B	0.0178	0	0.0895	0.0223	0.0255	0.0414	**0.0005**	0.0446
C	0.0804	0.0907	0	0.672	0.0641	0.0481	0.0900	0.0450
D	0.0147	0.0223	0.0687	0	**0.0032**	0.0191	0.0228	0.0223
E	0.0160	0.0254	0.0659	**0.0033**	0	0.0159	0.0260	0.0191
F	0.0370	0.0434	0.0482	0.0229	0.0211	0	0.0419	**0.0032**
G	0.0147	**0.0074**	0.0927	0.0240	0.0256	0.0465	0	0.0451
H	0.0425	0.0479	0.0450	0.0281	0.0573	**0.0056**	0.0514	0

	A	B	C	D	E	F	G	H
A	0	0.0277	0.0804	0.0148	0.0168	0.0373	0.0175	0.0432
B	0.0278	0	0.0930	0.0296	0.0365	0.0465	**0.0138**	0.0561
C	0.0813	0.0941	0	0.687	0.0662	0.0483	0.0931	0.0459
D	0.0152	0.0302	0.0693	0	**0.0075**	0.0277	0.0253	0.0298
E	0.0170	0.0365	0.0678	**0.0099**	0	0.0231	0.0294	0.0574
F	0.0389	0.0484	0.0483	0.0288	0.0270	0	0.0468	**0.0138**
G	0.0175	**0.0139**	0.0940	0.0257	0.0295	0.0483	0	0.0544
H	0.0436	0.0566	0.0464	0.0299	0.0580	**0.0150**	0.0547	0

use of the leading eigenvalue of *D*/*D* matrices well illustrates advantages of quantitative numerical characterization of graphical representation of DNA. Hence, the two approaches, the geometrical representation and the algebraic characterization of DNA, are complementary.

SIMILARITY AMONG DNA BASED ON THE LEADING EIGENVALUES

Although the leading eigenvalue of *D*/*D* matrices gives useful insights on folding of curves representing DNA, we can go beyond the description of DNA based on a single eigenvalue by constructing sequences that enlist increasing number of eigenvalues. Thus, for example, the first DNA of Table 1 can be characterized by the following sequence:

$$A_1 = 0.27136$$

$$A_2 = 0.27136, \quad 0.07010$$

$$A_3 = 0.27136, \quad 0.07010, \quad 0.06280$$

$$A_4 = 0.27136, \quad 0.07010, \quad 0.06280, \quad 0.04361$$

etc.

In Tables 5 and 6 we collected the similarity/dissimilarity values based on use of Euclidean distance between vectors for the eight DNA sequences of Table 1 represented by sequences having from one to four leading eigenvalues as members constructed as outlined above. A small entry in the similarity/dissimilarity matrix *may* point to DNA sequences that are similar, while the large entries indicate with certainty that the corresponding sequences are not similar. We have emphasized "may" because there could be spurious "similarity" caused by loss of information associated when one characterized a structure by its eigenvalues. Isospectral graphs are such an illustration.[39−53] On the other hand, if two structures are similar they are likely to have similar row sums and consequently similar eigenvalues.

First, observe that the relative magnitudes of the similarity /dissimilarity measure have changed little when instead of a single eigenvalue we use two, three, or four. We have compressed in Table 5 similarities based on single eigenvalue (the values above the main diagonal) and two leading eigenvalues (the values below the main diagonal), and in the lower part of Table 5 we have compressed similarities based on three leading eigenvalues (the values above the main diagonal) and four leading eigenvalues (the values below the main diagonal). For all data we have emphasized (by bold print) the three smallest values, which potentially indicate the most similar sequences in order to show that the results are essentially the same whether we use one or several leading eigenvalues. As we see from the top part of Table 5, the three smallest entries are (B, G), (D, E), (F, H). The last corresponds to the pair of β-globins of mouse and rat, which are expected to be very similar because of their evolutionary relationship. At the same time we see that opossum β-globin shows greater differences with all other species, including human. The situation has not much changed when we characterize the β-globins with two, three, or four leading eigenvalues. Again the same three pairs of entries show the smallest magnitudes in the similarity/ dissimilarity matrix. There are some variations in the relative magnitudes for similarity/dissimilarity among the selected DNAs in different sections of Table 5. However, the relative

Table 6. Similarity/Dissimilarity Tables for the Eight DNA Sequences of Table 1 Based on the Frequency of Occurrence of Pairs of Nucleic Bases[a]

	A	B	C	D	E	F	G	H
A	0	49	82	94	56	20	34	41
B		0	95	53	47	52	**78**	50
C			0	114	54	93	132	111
D				0	**118**	105	110	107
E					0	61	118	77
F						0	55	**40**
G							0	99
H								0

[a] Only one of the several smallest entries of Tables 5 and 6 remains small, pointing to genuine similarity.

values for a number of entries in the similarity/dissimilarity tables compressed in Table 5 have remained remarkably constant after additional eigenvalues were involved for the characterization of DNA path/walk sequences. For example, this is the case with most combinations involving H, except (F, H), pointing to the dominant role of the leading eigenvalue. In some cases inclusion of additional eigenvalues increases the dissimilarity among the sequences, as we can see for the pair (D, E).

Because of an inherent loss of information associated with eigenvalues, we have to use the results given in Table 5 with some caution. The best is to combine such information with similar information obtained by some other characterization of the set of structures. It is possible when distinct structural features are used for characterization of the same set of structures that the loss of information accompanying data reduction will not overlap and will be different in the two cases. Recently these same DNA sequences were characterized using the frequency of occurrence of pairs of nucleic bases in a sequence.[10] In Table 6 we reproduced the results of similarity based on such characterization. As we can see, again the pair (F, H) belonging to mouse and rat exons is associated with a small entry (great similarity), while the other two pairs that had small entries in Table 5, (B, G) and (D, E), now have relatively large entries. This shows that the apparent similarity of (B, G) and (D, E) of Table 5 does not signify genuine similarity of these pairs of β-globins. The lack of similarity between the corresponding DNA primary sequences agrees with the more distant evolutionary relationship of the corresponding species.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology of the Republic of Slovenia through Grant No. J1-8901.

REFERENCES AND NOTES

(1) Needleman, S. B.; Wunsch, C. D. *J. Mol. Biol.* **1970**, *48*, 443.
(2) Levinshtein, V. I. *Probl. Peredachi Inf.* **1971**, *7*, 30.
(3) Kernigham, B. W.; Lesk, M. E.; Ossana, J. F. *Bell. Syst. Technol. J.* **1978**, *57*, 215.
(4) Jerman-Blažič, B.; Fabič, I.; Randić, M. Comparison of sequences as a method for evaluation of the molecular similarity. *J. Comput. Chem.* **1986**, *7*, 176−188.
(5) Kruskal, J. *SIAM Rev.* **1983**, *25*, 201.
(6) Randić, M. Topological Indices. In *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 & Sons: Chichester, 1998; pp 3018−3032.
(7) Balaban, A. T. Historical developments of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devilers, J., Balaban, A. T., Eds., pp 21−57.
(8) Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devilers, J., Balaban, A. T., Eds., pp 563−593.
(9) Randić, M.; Novič, M.; Vračko, M. *Molecular Descriptors, New and Old*; Lecture Notes in Chemistry, submitted.
(10) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.
(11) Randić, N. On characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34
(12) Johnson, M. A., Maggiora, G. M., Eds. *Concept and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(13) Randić, M. Similarity methods of interest in chemistry. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, S. I., Ed.; Gordon & Breech: Amsterdam, 1996; pp 1−100.
(14) Jerman-Blažič, B.; Fabič, I.; Randić, M. Evaluation of the molecular similarity and property prediction for QSAR purposes. *Chemom. Intell. Lab. Syst.* **1989**, *6*, 49−63.
(15) Nandy, A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309−313.
(16) 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.
(17) Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.* **1996**, *12*, 55−62.
(18) Leong, P. M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *12*, 503−511.
(19) Hamori, E. Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques* **1989**, *7*, 710−720.
(20) Hamori, E. Visualization of biological information encoded in DNA. In *Frontiers of Computing Science, Vol. 3: Scientific Visualization*; Pickover, C., Tewskbury, S. K., Eds.; Plenum Press: New York, 1994; pp 91−121.
(21) Roy, A.; Raychaudhary, C.; Nandy, A. Novel techniques of graphical representation and analysis of DNA−A review. *J. Biosci.* **1998**, *23*, 55−71.
(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; pp 159−236.
(23) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(24) Randić, M. Novel molecular descriptor for structure−property studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.
(25) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(26) Hermann, A.; Zinn, P. List operation on chemical graphs. 6. Comparative study of combinatorial topological indexes of the Hosoya type. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 551−560.
(27) Randić, M. Hosoya matrix−a source of new molecular descriptors. *Croat. Chem. Acta* **1994**, *67*, 415−429.
(28) Hosoya, H. Topological index. A newly proposed quantity charaterizing the topological nature of structural isomers of saturated hydrocarbons, *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(29) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286.
(30) Randić, M.; Razinger, M. On characterization of molecular shapes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 594−606.
(31) Randić, M. Molecular profiles−Novel geometry-dependent molecular descriptors. *New J. Chem.* **1995**, *19*, 781−791.
(32) Randić, M. On characterization of conformations of nine-membered rings. *Int. J. Quantum Chem: Quantum Biol. Symp*. **1995**, *22*, 61−73.
(33) Randić, M. Molecular bonding profiles. *J. Math. Chem.* **1996**, *19*, 375−392.
(34) Randić, M.; Krilov, G. Bond profiles for cuboctahedron and twist cuboctahedron, *Int. J. Quantum Chem: Quantum Biol. Symp*. **1996**, *23*, 127−139.
(35) Randić, M.; Krilov, G. On characterization of molecular surfaces. *Int. J. Quantum Chem.* **1997**, *65*, 1065−1076.
(36) Bytautas, L.; Klein, D. J.; Randić, M.; Pisanski, T. Foldedness in linear polymers: a difference between graphical and Euclidean distances. *Discrete Mathematical Chemistry DIMACS Workshop on Discrete Mathematical Chemistry*; (Hansen, P.; Folwer, P.; Zheng, M., Eds.), Amer. Math. Soc.: Providence, RI, 2000, 21−57.
(37) Randić, M.; Krilov, G. Characterization of 3-D sequences of proteins. *Chem. Phys. Lett.* **1997**, *721*, 115−119.
(38) Randić, M.; Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.
(39) Collatz, L.; Sinogowitz, U. *Abh. Math. Sem. Univ. Hamburg* **1957**, *21*, 63.
(40) Harary, F.; King C.; Read, R. C. Cospectral graphs and digraphs, *Bull. London Math. Soc.* **1975**, *3*, 321.
(41) Godsil, C.; McKay, B. Some computational results on the spectra of graphs. *Lecture Notes Math.* **1976**, *560*, 72.
(42) Cvetković, D. M.; Doob, M.; Sachs, H. *Spectra of Graphs−Theory and Applications*; Academic Press: New York, 1980.
(43) Baker, C. A. Drum shapes and isospectral graphs. *J. Math. Phys.* **1966**, *7*, 2238.
(44) Fisher, M. E. On hearing the shape of a drum, *J. Combin. Theor.* **1966**, *1*, 105.
(45) Balaban, A. T.; Harary, F. The characteristic polynomial does not uniquely determine the topology of a molecule. *J. Chem. Docum.* **1971**, *11*, 258.
(46) Herndon, W. C.; Ellzey, M. L. Isospectral graphs and molecules. *Tetrahedron* **1975**, *31*, 99.
(47) Randić, M.; Trinajstić, N.; Živković, T. Molecular graphs having identical spectra. *J. Chem. Soc., Faraday Trans. 2* **1976**, *72*, 244.

(48) Heilbronner, E. Some comments on cospectral graphs. *MATCH* **1979**, *5*, 105.
(49) Živković, T.; Trinajstić, N.; Randić, M. On conjugated molecules having identical topological spectra. *Mol. Phys.* **1975**, *30*, 517.
(50) Jiang, Y. Problems on isospectral molecules. *Sci. Sin.* **1984**, *27*, 236.
(51) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. Kleiner, A. F.; Randić, M. On irreducible endospectral graphs. *J. Math. Phys.* **1986**, *27*, 2601−2612.
(52) Randić, M.; Baker, B. Isospectral multitrees. *J. Math. Chem.* **1988**, *2*, 249−265.
(53) Schwenk, A. J. Almost all trees are cospectral. In *New Directions in the Theory of Graphs*; Harary, F., Ed.; Academic Press: New York, 1973; pp 275−307.
(54) Gantmacher, F. *Theory of Matrices*; Chelsea Publishers: New York, 1959; Vol. II, Chapter 13.
(55) Randić, M.; Guo, X.; Oxley, T.; Krishnapryan, H.; Naylor, L. Wiener matrix invariants. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361−367.
(56) Lovasz, L.; Pelikan, J. I. On the eigenvalues of trees. *Period. Math. Hung.* **1973**, *3*, 175−182.
(57) Randić, M.; Plavšić, D.; Razinger, M. Double invariants. *MATCH* **1997**, *35*, 243−259.
(58) Randić, M. On structural ordering and branching of acyclic saturated hydrocarbons. *J. Math. Chem.* **1998**, *24*, 345−358.
(59) Randić, M.; Guo, X.; Bobst, S. Use of path matrices for characterization of molecular structures. *Discrete Mathematical Chemistry DIMACS Workshop on Discrete Mathematical Chemistry*; (Hansen, P.; Folwer, P.; Zheng, M., Eds.), Amer. Math. Soc.: Providence, RI, 2000, pp 305−322.
(60) Randić, M. On molecular branching. *Acta Chim. Sloven.* **1997**, *44*, 57−77.
(61) Coxeter, H. S. M. *Bull. Am. Math. Soc.* **1950**, *56*, 413.
(62) Randić, M.; Vračko, M.; Eigenvalues as molecular descriptors. In *QSAR/QSPR Studies by Molecular Descriptors*; Diudea, M. V., Ed.; Nova Publishers: Commack, NY.
(63) Randić, M. On characterization of cyclic structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.
(64) Pisanski, T. Plavšić, D.; Randić, M. On numerical characterization of cyclicity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 000−000.
(65) Randić, M.; Nandy, A.; Basak, S. C. On the numerical characterization of DNA primary sequences. *J. Math. Chem.*, submitted.
(66) Randić, M.; DeAlba, M. L. Dense Graphs and Sparse Matrices. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1078−1081.
(67) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recepies, The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1988; pp 335−377.

CI9901082