Characterization of DNA Primary Sequences Based on the Average Distances between Bases

Milan Randić*,† and Subhash C. Basak‡

Department of Mathematics & Computer Science, Drake University, Des Moines, Iowa 50311, and Natural Resources Research Institute, University of Minnesota at Duluth, 5031 Miller Trunk Highway, Duluth, Minnesota 55811

Received July 16, 2000

We outline numerical characterization of DNA primary sequence based on calculation of the average distance between pairs of nucleic acid bases. This leads to a representation of DNA by a condensed 4×4 symmetrical matrix, the elements of which give the average separation between pair of bases X, Y in DNA (X, Y = A, C, G, T). As an invariant of choice we consider the leading eigenvalue of the derived 4×4 matrix. Additional structurally related invariants were obtained by constructing additional "higher order" 4×4 matrices derived from the initial 4×4 matrix by raising its elements to higher powers. Suitably normalized leading eigenvalue of these matrices offer a novel characterization of DNA primary sequences, referred to as "DNA profiles". The approach is illustrated on exon 1 of human β-globin gene.

1. INTRODUCTION

An important task in analyzing available DNA data is to estimate the degree of similarity between finite sets of strings of nucleic bases. 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 Levensthein 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 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 the characterization of structures (be it small molecule or a macromolecule like DNA) by invariants, as opposed to the use of strings, is the simplicity of the comparison of numerical sequences based on invariants. The price paid, however, is a loss of information on some aspects of the structure that accompany any characterization based on invariants.[10] The loss of information, however, can be compensated in part by the use of a larger number of descriptors (invariants), as has been well illustrated in the QSAR model based on mathematical descriptors for molecules.[11−14]

The central problem to consider, if one is to use set of structural invariants instead of the structural codes, is how to find suitable invariants to characterize a given primary sequence of DNA. A way to arrive at structural invariants for a sequence is to associate a matrix with the sequence. Once a matrix has been constructed we can use a selection of matrix invariants as descriptors, which, upon ordering, offer a numerical characterization of the sequence. Recently construction of several matrices associated with DNA have been outlined,[15−20] based on graphical representations of DNA. Graphical representations of DNA have received some attention in the literature.[21−27] They result in a geometrical structure that is embedded either in a two- or three-dimensional space. A two-dimensional representation of DNA is obtained by assigning to the four nucleic bases the directions along the positive and negative x and y axes.[15] Alternately, if one assigns to the four nucleic acids the four tetrahedral direction in 3D space[16] one obtains a three-dimensional representation of DNA. From graphical representations of DNA one can construct a matrix representing DNA by calculating the Euclidean (through space) and the graph theoretical (through bonds) distances between all pairs of nucleic acid bases.

One can associate a matrix with DNA also without the use of graphical representations of DNA. One way to obtain a matrix not associated with graphical representation of DNA is to consider directly the primary sequence and to assign to each nucleic acid base two numbers: one number giving the position of a base in the DNA sequence and the other giving the position of a base in the subsequence of nucleic acid bases of the same kind. In Tables 1 and 2 we have illustrated these indicator numbers for exon 1 of human β-globin gene (92 bases). Using such labels we can represent the DNA sequence as a numerical sequence. For example, a portion of the DNA sequence corresponding to adenine (A) leads to the following numerical sequence:[18]

1/1, 2/8, 3/13, 4/20, 5/23, 6/25, 7/26, 8/37, 9/52, 10/53, 11/58, 12/59, 13/65, 14/68, 15/69, 16/80, 17/91

* Corresponding author phone: (515)292-7411; fax: (515)292-8629. Current address: 3225 Kingman Rd, Ames, IA 50014.
† Drake University.
‡ University of Minnesota at Duluth.

10.1021/ci0000981 CCC: $20.00
Published on Web 03/31/2001

Table 1. Exon-1 of Human Beta Globin Gene[a]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
C	C	T	G	A	G	G	A	G	A	A	G	T	C	T
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
G	C	C	G	T	T	A	C	T	G	C	C	C	T	G
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
T	G	G	G	G	C	A	A	G	G	T	G	A	A	C
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
G	T	G	G	A	T	G	A	A	G	T	T	G	G	T
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
G	G	T	G	A	G	G	C	C	C	T	G	G	G	C
91	92													
A	G													

[a] The nucleic bases are grouped in groups of five for better visibility.

Table 2. Same DNA Sequence Shown in Table 1 with Sequential Labels That Count Each of the Nucleic Acid Types Separately

1	1	1	2	2	3	1	2	2	3	3	4	3	4	4
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T
5	6	5	5	4	6	7	5	8	6	7	9	6	7	7
C	C	T	G	A	G	G	A	G	A	A	G	T	C	T
10	8	9	11	8	9	8	10	10	12	11	12	13	11	13
G	C	C	G	T	T	A	C	T	G	C	C	C	T	G
12	14	15	16	17	14	9	10	18	19	13	20	11	12	15
T	G	G	G	G	C	A	A	G	G	T	G	A	A	C
21	14	22	23	13	15	16	14	15	24	17	18	25	26	19
G	T	G	G	A	T	T	A	A	G	T	T	G	G	T
27	28	20	29	16	30	31	16	17	18	21	32	33	34	19
G	G	T	G	A	G	G	C	C	C	T	G	G	G	C
17	35													
A	G													

Similar sequences of length 19, 35, and 21 can be constructed for the remaining nucleic acids C, G and T, i.e., CC, GG, and TT, having 17, 19, 35, and 21 rows and columns. We will briefly return to these symmetric square matrices later (see section 7).

Still another way to obtain a matrix associated with a DNA primary sequence is to count the frequency of occurrences of pairs of bases X−Y at various separations. The frequency of X−Y bases, when summarized, leads to a reduced 4×4 matrices for DNA sequence, each of such matrices giving information on nucleic acid bases separated by different distances.[17]

In this paper we will consider the distances between pairs of nucleic acid bases rather than the frequency of the various pairs. We will show how the average distances between pairs of bases lead to a set of novel numerical invariants for the characterization of DNA.

We should mention that matrices have been used for convenient book-keeping of matching, mismatches, and deletions of a base or bases in a search for the best alignment of two sequences.[28−30] However, in such studies a matrix is always associated with two different sequences, not a single DNA sequence. In addition, such matrices were not analyzed for their properties relevant to the characterization of structure, such as construction and selection of invariants. Derivation of structural invariant for the complete characterization of DNA sequence is the hallmark of this paper.

2. REDUCED DNA MATRICES

A direct base-by-base transformation of a primary DNA sequence to a matrix will result in a matrix having many rows and columns. Such matrices need not offer immediately useful numerical insights into specifications of a particular DNA sequence. We are interested in numerical characterization of DNA and it seems desirable to construct reduced matrices which summarize information on DNA sequence as a whole, rather than using matrices based on extensive information relating to each individual pair of bases. By considering separately pairs of nucleic bases one can summarize pertinent information in a very condensed 4×4 matrix of the following form:

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

Here the individual elements XY relate to the information on the pair of bases X and Y. If the order of bases is not critical for the property considered, that is if XY is considered the same as YX, one obtains in this way a symmetric 4×4 matrix. Otherwise the 4×4 matrix would be nonsymmetrical, as was the case with the 4×4 matrices obtained when considering the frequency of a nucleic base X followed by base Y separated by distance d.[18]

The idea of a condensed 4×4 matrix can be further extended by considering triplets XYZ of nucleic acids.[20] In such a case one obtains a cubic "matrix", the elements of which are indicated by a triplet of subscripts i, j, k. The resulting "matrix" summarizes information on all 64 possible triplets of combinations of three nucleic base sequences,[29] starting with AAA and ending with TTT. In this work we will use 4×4 reduced matrices shown above in which the matrix element XY represents the average distance between X and Y in a segment of DNA considered. We will illustrate the characterization of DNA by such a matrix on exon 1 of human β-globin gene (Table 1). Hence, we will condense information contained in the segment of DNA which has 92 bases to a 4×4 matrix from which subsequently we will extract several structural invariants to be used as DNA descriptors.

3. AVERAGE X−Y BASE DISTANCE

Invariants derived from the graph theoretical distance matrix have found considerable application in the quantitative structure−activity relationship (QSAR) and the quantitative structure−property relationship (QSPR), respectively.[3−38] One of the simple such invariants is the Wiener number,[39] W, which is given as the sum of the matrix elements of the distance matrix above the main diagonal.[40] It differs from the average matrix element of the distance matrix only by a constant of proportionality. The distance matrix D(i, j) of a graph G was introduced in graph theory by Harary.[41] Its (i, j) element is defined by the length of the shortest path between vertices i and j. In linear structures, such as a string of nucleic acid bases of DNA, the distance between two sites is simply given by the difference of the corresponding sequence numbers. The Wiener number W continues to be used in chemical graph theory.[42] For acyclic molecules of a similar size (e.g. isomers) W is an indicator of the degree of molecular branching.[43−45] However, this interpretation has limitations and better alternative characterization of branching not based on the Wiener number was considered since.[46−49]

Table 3. Submatrix That Is Collecting Information on All A−A Separation Distances in the Primary DNA Sequence of Table 1

	1	8	13	20	23	25	26	37	52	53	58	59	65	68	69	80	91
1	0																
8	7	0															
13	12	5	0														
20	19	12	7	0													
23	22	15	10	3	0												
25	24	17	12	5	2	0											
26	25	18	13	6	3	1	0										
37	36	29	24	17	14	12	11	0									
52	51	44	39	32	29	27	26	15	0								
53	52	45	40	33	30	28	27	16	1	0							
58	57	50	45	38	35	33	32	21	6	5	0						
59	58	51	46	39	36	34	33	22	7	6	1	0					
65	64	57	52	45	42	40	39	28	13	12	7	6	0				
68	67	60	55	48	45	43	42	31	16	15	10	9	3	0			
69	68	61	56	49	46	44	43	32	17	16	11	10	4	1	0		
80	79	72	67	60	57	55	54	43	28	27	22	21	15	12	11	0	
91	90	83	78	71	68	66	65	54	39	38	33	32	26	23	22	11	0

As several papers have pointed out the distance matrix for characterization of molecules has some limitations,[50,51] because more distant elements are represented by larger entries in such matrix, while the opposite seems desirable. This led to consideration of a matrix of reciprocal distances and similar modifications.[52−56] On the other hand in view of the wide use and application of distance matrix and distance based invariants, it seems desirable to investigate distances properties associated with strings of nucleic acid bases in DNA as these may lead to analogous distance-based invariants to be used to characterize DNA sequences.

To illustrate the approach in Table 2 we have collected the distances between all pairs A−A measured along the DNA chain in of exon 1 of Table 1. Because base A occurs 17 times in the DNA sequence, we obtained a symmetric 17×17 the distance matrix AA, of which we have only shown the lower portion. From this 17×17 matrix we can evaluate the average matrix element by summing all the entries in the matrix and dividing it by 17^2, which is $8564/289 = 29.633218$. Observe that we included in the count of all distances between the A−A pairs also the zero A−A distances along the main diagonal. Hence, in making the average we have 17^2 in the denominator for the fraction shown. For the same 17×17 matrix the Wiener number is 4287, which when doubled gives 8564, which appears in the numerator of the fraction shown above.

The reason for including the zero distances (the paths of length zero) becomes more apparent when one considers the complete distance matrix for exon 1 (Table 1) which would be of size 92×92. The 17×17 matrix AA of Table 3 is the part of the 92×92 distance matrix rearranged so that its elements are partitioned into the rectangular submatrices associated with individual pairs of XY bases as shown below:

AA	AC	AG	AT	17×17	17×19	17×35	17×21
CA	CC	CG	CT	19×17	19×19	19×35	19×21
GA	GC	GG	GT	35×17	35×19	35×35	35×21
TA	TC	TG	TT	21×17	21×19	21×35	21×21

The dimensions of these rectangular matrices are given by the frequency of the nucleic bases of each type. As mentioned before it is only along the diagonal that we have quadratic submatrices with zero diagonal entry, the size of which is given similarly by the total number of the corresponding bases.

Table 4. Submatrix That Is Collecting Information on All A−C Separation Distances in the Primary DNA Sequence of Table 1

	1	8	13	20	23	25	26	37	52	53	58	59	65	68	69	80	91
7	6	1	6	13	16	18	19	30	45	46	51	52	58	61	62	73	84
9	8	1	4	11	14	16	17	28	43	44	49	50	56	59	60	71	82
10	9	2	3	10	13	15	16	27	42	43	48	49	55	58	59	70	81
14	13	6	1	6	9	11	12	23	38	39	44	45	51	54	55	66	77
16	15	8	3	4	7	9	10	21	36	37	42	43	49	52	53	64	75
17	16	9	4	3	6	8	9	20	35	36	41	42	48	51	52	63	74
29	28	21	16	9	6	4	3	8	23	24	29	30	36	39	40	51	62
32	31	24	19	12	9	7	6	5	20	21	26	27	33	36	37	48	59
33	32	25	20	13	10	8	7	4	19	20	25	26	32	35	36	47	58
38	37	30	25	18	15	13	12	1	14	15	20	21	27	30	31	42	53
41	40	33	28	21	18	16	15	4	11	12	17	18	24	27	28	39	50
42	41	34	29	22	19	17	26	5	10	11	16	17	23	26	27	38	49
43	42	35	30	23	20	18	17	6	9	10	15	16	22	25	26	37	48
51	50	43	38	31	28	26	25	14	1	2	7	8	14	17	18	29	40
60	59	52	47	40	37	35	34	23	8	7	2	1	5	8	9	20	31
83	82	75	70	63	60	58	57	46	31	30	25	24	18	15	14	3	8
84	83	76	71	64	61	59	58	47	32	31	26	25	19	16	15	4	7
85	84	77	72	65	62	60	59	48	33	32	27	26	20	17	16	5	6
90	89	82	77	70	67	65	64	53	38	37	32	31	25	22	21	10	1

Table 5. Condensed 4×4 Matrix: the Elements of Which Show the Average Separation between X−Y Nucleic Acid Bases (X, Y = A, C, G, T)[a]

AA	AC	AG	AT	29.633218	30.941176	30.998319	29.708683
	CC	CG	CT		30.116343	32.348872	30.441103
		GG	GT			15.193469	30.394558
			TT				28.961451
AA	AC	AG	AT	30.281500	30.806250	31.510714	30.071429
	CC	CG	CT		29.890000	31.871429	30.114286
		GG	GT			15.193469	30.394558
			TT				28.961451

[a] The top part corresponds to the DNA sequence of Table 1, and the bottom part corresponds to the hypothetical DNA sequence in which adenine at the position 58 is replaced by cytosine.

In Table 4 we illustrate the AC rectangular submatrix that records the distance between adenine (A) and cytosine (C). It has 17 columns and 19 rows corresponding to the number of A and C, respectively. The difference between the successive rows in the AC submatrix is constant for all the rows and the columns till the position in the column when the row label becomes bigger than the column label. Then the sense of the difference is reversed and the relative magnitudes of successive rows or columns are reversed. A similar regularity can be found also for the difference between the successive columns, except for the rows which have a label that is larger than the first column and smaller than the next column when instead of the difference we have a constant sum. These regularities may help one to find numerical errors if the distance submatrices are not constructed by computer. In the case of the AC submatrix shown in Table 4, the sum of all 17×19 matrix entries is 9994, which gives the average value of the matrix element of AC submatrix $9994/(17 \times 19) = 30.941176$. In Table 5 (top part) we have collected all XY average elements (X, Y = A, C, G, T) for the distance matrix of the exon 1 of the human β-globin gene. As a result we obtain a symmetrical 4×4 matrix, the construction of the first two elements of which has been outlined above.

The elements of the derived condensed matrix represents a considerable contraction of the information of the DNA sequence considered. So the following question can be immediately raised: does such a drastically simplified matrix contain enough compositional information to be useful when

Table 6. Selection of Matrix Invariants Derived from Condensed 4 × 4 Matrix[a]

matrix invariant		A/C
the maximal row sum	123.281396	122.681965
the minimal row sum	108.935218	108.970170
the average row sum	118.392476	118.465938
the leading eigenvalue	118.638256	118.707621
other eigenvalues:	−0.399856	−0.449911
	−1.150524	−0.772788
	−13.183404	−13.158502
trace (the sum of eigenvalues)	103.904481	104.32642
average matrix element	29.598119	29.616485

[a] The last column corresponds to the case of A/C substitution.

comparing different DNA sequences. That drastically condensed representation of complex systems can have useful information has been recently demonstrated by several researchers[57,58] who were able to arrive at a useful summary for properties of structural isomers by considering average properties of a large set of compounds. More relevant for our case are the characterizations of molecules using the so-called molecular "profiles".[59−63] Because a single invariant may not suffice to characterize complex systems, design of an additional set of invariants seems desirable. In particular it is desirable to have invariants which are related structurally, rather than just having a set of *ad hoc* derived invariants. Molecular profiles represent one such set of structurally related invariants. They are constructed from a set of "higher order" matrices representing a molecule again by averaging matrix elements. They can be viewed as components of a vector giving a "profile" of the sequence considered. The structurally related matrices are derived using suitable algebraically manipulations of the matrix elements of the initial distance matrix, as will be outlined in the next section.

4. INVARIANTS OF REDUCED MATRICES

From the 4 × 4 matrix of Table 4 we can select several matrix invariants, including the following: the eigenvalues, the average matrix element, the average row sum, the maximal row sum, the minimal row sum, the determinant, the trace (the sum of the diagonal elements), and if desired the coefficients of the characteristic polynomial, etc. The maximal and the minimal row sum represent, according to the theorem of Frobenius-Perron,[64] the upper and the lower bounds on the leading eigenvalue of a matrix. In Table 6 we listed several of the above-mentioned invariants for the 4 × 4 matrices of Table 5. Observe how close are the magnitudes of the average row sum and the leading eigenvalue, which is a consequence of the fact that individual matrix elements (except for the instance of CC element) are of similar magnitude. Hence, when the individual matrix elements of the 4 × 4 matrix do not differ much, the average row sum may offer a satisfactory estimate of the leading eigenvalue.

In the following we will here consider only the leading eigenvalue of the reduced matrix as the invariant used for characterization of DNA. The leading eigenvalue of matrices associated with a molecular graph have found useful interpretations. Lovasz and Pelikan pointed to the use of the leading eigenvalue of the adjacency matrix as a measure of branching.[65] Randić, Kleiner, and DeAlba[66] have interpreted the leading eigenvalue of the so-called D/D matrix (the elements of which are given as quotient of the Euclidean and graph theoretical distances for an embedded graph in 3D space) as an index of molecular folding. In another study Randić, Vračko, and Novič[67] related the leading eigenvalue of the line adjacency matrix of an embedded graph as a measure of molecular flexibility. More recently, the leading eigenvalue of the path matrix[68] was found to offer an even better, or at least more discriminatory, characterization of molecular branching.[48,49] The leading eigenvalue of the D/DD matrix (the elements of which are constructed as the quotient of the corresponding elements of the distance matrix (D) and the detour matrix (DD))[41,69−73] was suggested as a measure of molecular cyclicity.[74,75] In view of the apparent structural significance of the leading eigenvalues of various matrices associated with chemical structures it seems worthwhile to explore the use of the leading eigenvalues of condensed DNA matrices for the characterization of DNA. In passing we should add that other eigenvalues, even eigenvectors, have been considered as a source for construction of topological indices.[76]

Table 7. Eigenvalues, the Normalization Factors, and the Normalized Leading Eigenvalues of the "Higher Order" Condensed Matrices[a]

	eigenvalue	normalization	profile original	after A/C substitution
1	118.64	1	118.64	118.71
2	3577.67	$1/2^2$	894.42	895.22
3	10,875.22	$1/6^2$	3020.89	3023.57
4	3,320,808.00	$1/24^2$	5765.29	5768.33
5	101,705,087.43	$1/120^2$	7062.85	7061.40
6	3,121,851,701.12	$1/720^2$	6022.09	6014.04
7	96,005,201,290.02	$1/5040^2$	3779.49	3768.59
8	2,957,422,941,020	$1/40320^2$	1819.17	1810.29
9	91,249,864,640,008	$1/362880^2$	692.96	687.88
10	2,819,911,366,087,310	$1/3628800^2$	214.15	211.95
11	$8.728054\ 10^{16}$	$1/(11!)^2$	54.78	54.03
12	$2.705684\ 10^{18}$	$1/(12!)^2$	11.79	11.58
13	$8.400690\ 10^{19}$	$1/(13!)^2$	2.17	2.12
14	$2.612354\ 10^{21}$	$1/(14!)^2$	0.34	0.33
15	$8.136323\ 10^{22}$	$1/(15!)^2$	0.05	0.05
16	$2.538064\ 10^{24}$	$1/(16!)^2$	0.01	0.01

[a] The last column corresponds to the case of A/C substitution.

5. CONSTRUCTION OF THE LEADING EIGENVALUES "PROFILE" OF THE DNA

Besides the well-known standard product of two matrices A B in matrix algebra one can consider also the product A^B (also referred to as Kronicker's product) defined by multiplying the element a_{ij} of matrix A and the element b_{ij} of matrix B.[77] If A = B we obtain from matrix A matrix ^{2}A, the elements of ^{2}A are given by the squares of the elements of the original matrix, i.e., $^2a_{ij} = (a_{ij})^2$. The leading eigenvalue of this matrix ($^2\lambda_1$) is an additional structural invariant that can be used for characterization of the primary sequence of DNA. The process can be continued and in addition to ^{2}A one can construct a set of matrices ^{k}A by repeatedly multiplying ^{2}A matrix by A, etc. In this way we obtain a matrices A, ^{2}A, ^{3}A, ^{4}A, ^{5}A, ^{6}A, ... which yield as invariants an ordered set of the leading eigenvalues λ_1, $^2\lambda_1$, $^3\lambda_1$, $^4\lambda_1$, $^5\lambda_1$, $^6\lambda_1$, ...

Because matrix elements of ^{2}A, and other higher order matrices, continue to increase in magnitude upon exponentiation the corresponding leading eigenvalues $^m\lambda_1$ also increase

Table 8. Elements of the Condensed Matrix as Components of 10-Dimensional Vector for DNA Sequence of Table 1 and the Hypothetical Sequence Obtained by Substitution of a Single Adenine by Cytosine[a]

	original DNA	A/C substitution	difference
AA	29.6332	30.2815	−0.6483
AC	30.9412	30.8063	+0.1349
AG	30.9983	31.5107	−0.5124
AT	29.7087	30.0714	−0.3627
CC	30.1163	29.8900	+0.2263
CG	32.3489	31.8714	+0.4774
CT	30.4411	30.1143	+0.3268
GG	15.1935	15.1935	0
GT	30.3946	30.3946	0
TT	28.9615	28.9615	0

[a] The last column shows the difference between the two cases.

in magnitude (see the second column in Table 7). To avoid a divergent sequence of the ${}^{k}\lambda_1$ descriptors we need to normalize the derived leading eigenvalues. Using $(1/n!)^2$ as the normalization factor we obtain a converging sequence of normalized leading eigenvalues of Table 7, which represents a sequence of invariants that offers a characterization of DNA. We will refer to such a constructed sequence of descriptors/invariants as "DNA profile."

6. A TEST OF THE SENSITIVITY OF DNA PROFILES

One of the most important questions that characterizes a system by an invariant, including molecular profiles, is the sensitivity of the derived invariants to minor changes in the DNA sequence. To test the sensitivity of the "DNA profile" we have perturbed the original DNA sequence of Table 1 by replacing a single nucleic base in the position 58, which was A, by C. For the modified DNA sequence we constructed the reduced 4×4 matrix which is shown in Table 5 (the lower part). The submatrices involving A and C, i.e., the submatrices AA, AC, AG, AT, CC, CG, CT, and the corresponding symmetry equivalent submatrices CA, GA, TA, GC, and TC, which are all the submatrices in the first two columns and the first two rows, will be affected by the replacement of a single A by C. As we see by a comparison of the two 4×4 matrices of Table 5 the individual matrix elements change significantly, even if not dramatically. The parts of the 4×4 matrices corresponding to elements GG, GT, TG, and TT have not changed (as expected). As a consequence of introduced changes in matrix elements based on the "higher order" matrices the invariants listed in Table 6 have changed also. Similarly the DNA profiles have changed, as can be seen by comparing the last two columns of Table 7. Because the small difference between the leading eigenvalues are magnified by recursive multiplication the difference in magnitudes between the corresponding entries of two profiles becomes pronounced for the intermediate section of the profiles corresponding to larger "amplitudes" of the profiles. If the two profiles are viewed as vectors in a 16-dimensional space, the Euclidean distance between the two profiles is 17.69 even though the difference in the leading eigenvalues was only 0.07.

An alternative characterization of DNA is given directly by the elements of the 4×4 matrix, without calculating the leading eigenvalues. By canonical ordering (here to be taken to be alphabetical order) from the 10 distant matrix elements as (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT) we obtain a vector in a 10-dimensional vector space. In Table 8 we have listed the component of the two 10-component vectors corresponding to the two matrices of Table 5. Again we see that the single substitution of A by C induces, visible even if not large, change in the magnitudes of the components. This points to sufficient sensitivity of the 4×4 considered matrices of DNA to minor changes in the nucleic bases composition.

Table 9. Four Segments of DNA of Table 1 Each of Length Ten and the Corresponding Distance Matrices

	1	2	3	4	5	6	7	8	9	10	
	A	T	G	G	T	G	C	A	C	C	
	1	8	7	9	10	3	4	6	2	5	row sum
1	0	7	6	8	9	2	3	5	1	4	45
8	7	0	1	1	2	5	4	2	6	3	31
7	6	1	0	2	3	4	3	1	5	2	27
9	8	1	2	0	1	6	5	3	7	4	37
10	9	2	3	1	0	7	6	4	8	5	45
3	2	5	4	6	7	0	1	3	1	2	31
4	3	4	3	5	6	1	0	2	2	1	27
6	5	2	1	3	4	3	2	0	4	1	25
2	1	6	5	7	8	1	2	4	0	3	37
5	4	3	2	4	5	2	1	1	3	0	25

	11	12	13	14	15	16	17	18	19	20	
	T	G	A	C	T	C	C	T	G	A	
	3	10	4	6	7	2	9	1		8	row sum
3	0	7	1	3	4	1	6	2	2	5	31
10	7	0	6	4	3	8	1	9	5	2	45
4	1	6	0	2	3	2	5	3	1	4	27
6	3	4	2	0	1	4	3	5	1	2	25
7	4	3	3	1	0	5	2	6	2	1	27
2	1	8	2	4	5	0	7	1	3	6	37
9	6	1	5	3	2	7	0	8	4	1	37
1	2	9	3	5	6	1	8	0	4	7	45
5	2	5	1	1	2	3	4	4	0	3	25
8	5	2	4	2	1	6	1	7	3	0	31

	21	22	23	24	25	26	27	28	29	30	
	G	G	A	G	A	A	G	T	C	T	
	3	5	6	9	1	2	4	7	8	10	row sum
3	0	2	3	6	2	1	1	4	5	7	31
5	2	0	1	4	4	3	1	2	3	5	25
6	3	1	0	3	5	4	2	1	2	4	25
9	6	4	3	0	8	7	5	2	1	1	37
1	2	4	5	8	0	1	3	6	7	9	45
2	1	3	4	7	1	0	2	5	6	8	37
4	1	1	2	5	3	2	0	3	4	6	27
7	4	2	1	2	6	5	3	0	1	3	27
8	5	3	2	1	7	5	4	1	0	2	31
10	7	5	4	1	9	8	6	3	2	0	45

	31	32	33	34	35	36	37	38	39	40	
	G	C	C	G	T	T	A	C	T	G	
	7	2	3	8	1	4	10	5	6	9	row sum
7	0	5	4	1	6	3	3	2	1	2	27
2	5	0	1	6	1	2	8	3	4	7	37
3	4	1	0	5	2	1	7	2	3	6	31
8	1	6	5	0	7	4	2	3	2	1	31
1	6	1	2	7	0	3	9	4	5	8	45
4	3	2	1	4	3	0	6	1	2	5	27
10	3	8	7	2	9	6	0	5	4	1	45
5	2	3	2	3	4	1	5	0	1	4	25
6	1	4	3	2	5	2	4	1	0	3	25
9	2	7	6	1	8	5	1	4	3	0	37

7. NONCOMPACT MATRIX REPRESENTATION OF DNA

There is no doubt that by condensing a 92×92 matrix associated with the primary sequence of DNA of exon-1 of human beta globin gene to a 4×4 matrix relating to the

Table 10. 4 × 4 Condensed Matrices, the Row Sums, the Eigenvalues, the Trace, and the Determinant for the Four Segments of DNA of Table 8

4 × 4 matrix				row sum	eigenvalues	trace	det
			Fragment 1−10				
3.5	4.5	3.5	3.5	15	13.54495	7.66667	−19.3796
4.5	1.33333	4.33333	5.16667	15.33333	−0.30422		
3.5	4.33333	1.33333	1.83333	11	−1.03647		
3.5	5.16667	1.83333	1.5	12	−4.54495		
			Fragment 11−20				
3.5	3.5	4	4.16667	15.16667	13.91763	11.44444	−4.14815
3.5	1.33333	3.5	2.77778	11.11111	−0.29147		
4	3.5	3.5	3.83333	14.83333	−0.68171		
4.16667	2.77778	3.83333	3.11111	13.88889	−1.50000		
			Fragment 21−30				
1.33333	4.33333	2.5	4.33333	12.5	13.24035	5.45833	−18.0694
4.33333	0	5.5	1	10.83333	−0.40600		
2.5	5.5	3.125	5.5	16.625	−0.48801		
4.33333	1	5.5	1	11.83333	−6.24035		
			Fragment 31−40				
0	3.33333	4	1.66667	9	12.59011	8.44444	−10.4691
3.33333	2.66667	3.77778	3.44444	13.22222	−0.40697		
4	3.77778	4	3.88889	15.66667	−0.66468		
1.66667	3.44444	3.888889	1.77778	10.77778	−3.07402		

average distance information for various base pairs we are bound to lose a considerable amount of detailed information on DNA. Is there some less drastic option for characterization of DNA?

To consider this question we will examine several 10 × 10 submatrices of the initial 92 × 92 distance matrix. In Table 9 we show four such submatrices corresponding to the DNA subsequences 1−10, 11−20, 21−30, and 31−40 of the sequence shown in Table 1. We will pretend that these four cases simulate four DNA sequences in general, though in fact they represent fragments of a single DNA primary sequence and illustrate properties of local DNA invariants.

First to observe in Table 9 is that although the four matrices appear different they in fact represent the same matrix (shown in Table 10) in which the rows and columns have been permuted. Such matrices are related by a similarity transformation S^{-1} M S and necessarily have identical eigenvalues. Thus the leading eigenvalue is here of no particular interest. Neither are the row sums of interest since again all four matrices have the same individual row sums, listed only in a permuted order. Consequently, the Wiener index of the four matrices is the same ($W = 165$). In fact W depends only on the size of such matrices, all matrices of the same size having the same W given by the table:

size	1	2	3	4	5	6	7	8	9	10
W	0	1	4	10	20	35	56	84	120	165

The successive increments are given by twice the binomial coefficient 1, 3, 6, 10, 15, 28, ...

Hence, the complete distance matrix for a string of DNA basis is not suitable for extraction of sequence invariants. One way out of this dilemma is to consider the so-called D/D matrices mentioned earlier, the elements of which are given as quotients of two distinctive measures imposed on a sequence. Another possibility is to focus attention to diagonal submatrices AA, CC, GG, and TT, rather than considering the whole matrix. As we can see from Table 9 these diagonal submatrices vary in size and magnitudes of entries from case to case (or from a fragment to a fragment), thus they may yield useful sequence (or fragment) invariants.

Table 11. 4 × 4 Condensed Matrices, the Row Sums, the Eigenvalues, the Trace, and the Determinant for the Three Segments of DNA Obtained by Replacement of A at Positions 3, 5, and 6 by C

4 × 4	matrix			row sum	eigenvalues	trace	det
			Replacement A_3 by C				
0.5	3	2.75	3.5	9.75	12.97593	7	−31.8438
3	3	3.75	3.5	13.25	−0.39490		
2.75	3.75	2.5	5.5	14.5	−1.53547		
3.5	3.5	5.5	1	13.5	−4.04456		
			Replacement A_5 by C				
1.5	3	2.5	4.5	11.5	12.87524	7	−15.6250
3	2	4	2.5	11.5	−0.35129		
2.5	4	2.5	5.5	14.5	−0.71897		
4.5	2.5	5.5	1	13.5	−4.87524		
			Replacement A_6 by C				
1	3.5	2.25	5	11.75	12.87702	6	−20.8438
3.5	1.5	4.25	2	11.25	−0.38998		
2.25	4.25	2.5	5.5	14.5	−0.71969		
5	2	5.5	1	13.5	−5.76735		

In Table 11 we have constructed the reduced 4 × 4 matrices for the four fragment sequences of Table 9. As is to be expected the reduced matrices, the elements of which represent normalized average matrix elements, show variations between different sequence fragments. The leading eigenvalues of such matrices (or ordered sequence AA, AC, AG, AT, CC, CG, CT, GG, GT, TT of the average elements) constitute local sequence invariants. As we see from Table 11 the eigenvalues for the four fragments show considerable variation in magnitudes.

Finally, let us illustrate on one of 10 × 10 matrices how sensitive are matrix invariants on a replacement of adenine (A) by cytosine (C). We selected the third matrix of Table 9 where the DNA fragment has only one cytosine base (at position 9) and three adenine bases (at positions 3, 5, and 6). In Table 12 we show the corresponding 4 × 4 matrices, the row sums, the eigenvalues, the trace, and the determinants. As we see the selected invariants of the 4 × 4 condensed matrices are quite sensitive on the location of the

replacement of adenine by cytosine. The eigenvalues, and in particular the leading eigenvalue, apparently show minor variations, while the trace and the determinant (which also appear in the characteristic polynomial of the corresponding eigenvalue problem) show apparently unpredictable changes, suggesting that the coefficient of the characteristic polynomial may be less suitable as descriptors for DNA sequences by not showing simpler regularity in their variations.

8. CONCLUDING REMARKS

By constructing a DNA "profile" we succeeded in replacing the primary sequence of DNA by a sequence of numerical invariants. Comparison of two DNA sequences is now transformed into a comparison of the corresponding sequences of mathematical descriptors of DNA which is a straightforward mathematical exercise. Direct comparison of sequences based on invariants can lead to partial ordering in addition to the traditional table of similarity/dissimilarity among sequences. Future applications of this approach and possible modifications will demonstrate which of the various methods outlined here for characterization of DNA may be useful for specific problems. As has been the case with the introduction of numerous topological indices in QSAR, different structural invariants may play a dominant role in different applications.

ACKNOWLEDGMENT

This is contribution number 294 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by Grants F49620-98-1-0015 and F49620-01-1-0098 from the U.S. Air Force.

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-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.
(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) It is generally believed that finite list of simple invariants cannot uniquely represent graph or a molecular structure. In the words of Frank Harary: "No decent complete set of invariants for a graph is known" (see ref 41 p 11).
(11) 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.
(12) Balaban, A. T.; Ivanciuc, O. Historical developments of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Sci. Publ.: Amsterdam, 1999; pp 21−57.
(13) Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Sci. Publ.: Amsterdam, 1999; pp 563−593.
(14) Randić, M.; Novič, M.; Vračko, M. *Molecular Descriptors, New and Old, Lecture Notes in Chemistry*; Submitted for publication.
(15) Randić, M.; Nandy, A.; Basak, S. C. On the numerical characterization of DNA primary sequences. *J. Math. Chem.* Submitted for publication.
(16) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−1244.
(17) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.
(18) Randić, M. On characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34
(19) Randić, M.; Vračko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.
(20) Nandy, A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Current Sci.* **1994**, *66*, 309−313.
(21) Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Current Sci.* **1996**, *70*, 661−668.
(22) Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci. (CABIOS)* **1996**, *12*, 55−62.
(23) Leong, P. M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci. (CABIOS)* **1995**, *12*, 503−511.
(24) Hamori, E. Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques* **1989**, *7*, 710−720.
(25) Hamori, E. Visualization of biological information encoded in DNA. In *Frontiers of Computing Science;* Pickover, C., Tewskbury, S. K., Eds.; J. Wiley and Sons: New York 1994; Vol. 3: Scientific Visualization, pp 90−121.
(26) Roy, A.; Raychaudhary, C.; Nandy, A. Novel techniques of graphical representation and analysis of DNA - - A review. *J. Biosci.* **1998**, *23*, 55−71.
(27) Randić, M.; Guo, X.; Basak, S. C. Characterization of DNA based on occurrence of triplets of nucleic bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619−626.
(28) Tinoco, I., Jr.; Uhlenbeck, O, C.; Levine, M. D. Estimation of secondary structure in ribonucleic acids. *Nature* **1971**, *230*, 362−367.
(29) Max, E. E.; Maizel, J. V., Jr.; Leder, P. The nucleotide sequence of a 5.5- kilobase DNA segment containing the mouse κ immunoglobulin J and C region genes. *J. Biol. Chem.* **1981**, *256*, 5116−5120.
(30) Goad, W. B.; Kanehisa, M. I. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nuclei Acids Res.* **1982**, *10*, 247−263.
(31) Trinajstić, N. *Chemical Graph Theory;* CRC Press: Boca Raton, FL, 1992
(32) Buckley, F.; Harary, F. *Distance in Graphs;* Addison-Wesley: Reading, MA, 1990.
(33) Randić, M.; Pompe, M. The variable molecular descriptors based on distance related matrices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575−581.
(34) Randić, M.; Balaban, A. T.; Basak, S. C. On structural interpretation of distance related topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 593−601.
(35) Mihalić, Z.; Veljan, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The distance matrix in chemistry. *J. Math. Chem.* **1992**, *11*, 223−258.
(36) Ivanciuc, O.; Ivanciuc, T. matrices and structural descriptors computed from molecular graph distances. In *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers J., Balaban, A. T., Eds.; Gordon and Breach Sci. Publ.: Amsterdam, 1999; pp 221−227.
(37) Mihalić, Z.; Nikolić, S.; Trinajstić, N. Comparative study of molecular descriptors derived from the distance matrix. *J. Chem. Inf. Comput. Sci.* **1999**, *32*, 28−37.
(38) Lučić, B.; Lukovits, I.; Nikolić, S.; Trinajstić, N. On distance indices in QSPR modeling. *J. Chem. Inf. Comput. Sci.* Submitted for publication.
(39) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(40) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(41) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(42) *MATCH (communications in mathematical and computer chemistry);* Gutman, I., Klavzar, S., Mohar, B., guest Eds.; Published by A. Kerber, Department of Mathematics, University of Bayreuth: Bayreuth, Germany.

(43) Bonchev, D.; Trinajstić, N. On topological characterization of molecular branching. *Int. J. Quantum Chem: Quantum Chem. Symp.* **1978**, *12*, 293−303.
(44) Bonchev, D.; Trinajstić, N.; Information theory, Distance matrix, and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(45) Bonchev, D. Topological order in molecules 1. Molecular branching revisited. *J. Mol. Struct. (THEOCHEM)* **1995**, *336*, 137−156.
(46) Bertz, S. H. Branching in graphs and molecules. *Discrete Appl. Math.* **1988**, *19*, 65−83.
(47) Ivanciuc, O.; Ivanciuc, Y.; Carbol-Bass, D.; Balaban, A. T. Investigation of Alkane branching with topological indices. *Pegasus* Submitted for publication.
(48) Randić, M. On structural ordering and branching of acyclic saturated hydrocarbons. *J. Math. Chem.* **1998**, *24*, 345−358.
(49) Randić, M.; Guo, X.; Bobst, S. Use of path matrices for characterization of molecular structures. *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* **2000**, *51*, 305−322.
(50) Randić, M. Linear combinations of path numbers as molecular descriptors. *New J. Chem.* **1997**, *21*, 945−951.
(51) Balaban, A. T.; Mills, D.; Ivanciuc, O.; Basak, S. C. Reverse Wiener Indices. *Croat. Chem. Acta* **2000**, *73*, 923−941.
(52) Balaban, A. T.; at al. The complementary distance matrix, a new molecular graph metric. Work in progress.
(53) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary index for the characterization of chemical graphs. *J. Math. Chem.* **1993**, *12*, 235−250.
(54) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *J. Math. Chem.* **1993**, *12*, 309−318.
(55) Diudea, M. V. Indices of reciprocal properties of Harary indices. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 292−299.
(56) Randić, M, Pompe M. Variable Molecular Descriptors; Poster presented at the 2nd Indo-U.S. Workshop on Mathematical Chemistry, Duluth, MN, May 29−June 3, 2000.
(57) Dobrynin, A. A.; Gutman, I. The average Wiener index of trees and chemical trees. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 679−683.
(58) Bytautas, L.; Klein, D. J. Mean Wiener numbers and other mean extensions for alkane trees. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 471−481.
(59) Randić, M. Molecular profiles - - Novel geometry-dependent molecular descriptors. *New J. Chem.* **1995**, *19*, 781−791.
(60) Randić, M.; Razinger, M. On characterization of molecular shapes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 594−606.
(61) Randić, M.; Krilov, G. On characterization of molecular surfaces. *Int. J. Quantum Chem.* **1997**, *65*, 1065−1076.
(62) Randić, M.; Krilov, G. Characterization of 3-D sequences of proteins. *Chem. Phys. Lett.* **1997**, *721*, 115−119.
(63) Randić, M.; Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.
(64) Gantmacher, F. *Theory of matrices;* Chelsea Publ: New York, 1959; Vol. II, Chapter 13.
(65) Lovasz, L.; Pelikan, J. I. On the eigenvalues of trees. *Period. Math. Hung.* **1973**, *3*, 175−182.
(66) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286.
(67) Randić, M.; Vračko, M.; Novič, M. Eigenvalues as molecular descriptors. In *QSAR/QSPR by Molecular Descriptors*; Diudea, M. V., Ed.; Nova Publ.: In press.
(68) Randić, M.; Plavšić, D.; Razinger, M. Double invariants. *MATCH* **1997**, *35*, 243−259.
(69) Amić, D.; Trinajstić, N. On the detour matrix. *Croat. Chem. Acta* **1995**, *68*, 53−62.
(70) Lukovits, I. The detour index. *Croat. Chem. Acta* **1996**, *69*, 873−882.
(71) Lukovits, I.; Razinger, M. On calculation of the detour index. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 283−286.
(72) Trinajstić, N.; Nikolić, S.; Lučić, B.; Amić, D.; Mihalić, Z. The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 631−638.
(73) Randić, M. DeAlba, L. M.; Harris, F. E. Graphs with identical detour matrix. *Croat. Chem. Acta* **1998**, *71*, 53−68.
(74) Randić, M. On characterization of Cyclic structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.
(75) Pisanski, T. Plavšić, D.; Randić, M. On numerical characterization of cyclicity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 520−523.
(76) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517−523.
(77) This operation is available on MATLAB as a.∧b (power (a, b): Element- by-element power operation); Hanselman, D., Littlefield, B., Eds.; Mastering MATLAB 5, Prentice Hall: Upper Saddle River, NJ, 1998.

CI0000981