On A Four-Dimensional Representation of DNA Primary Sequences

Milan Randić[†,§] and Alexandru T. Balaban*[,‡]

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 53311, and
Department of Marine Sciences, Texas A&M University at Galveston, Galveston, Texas 77553

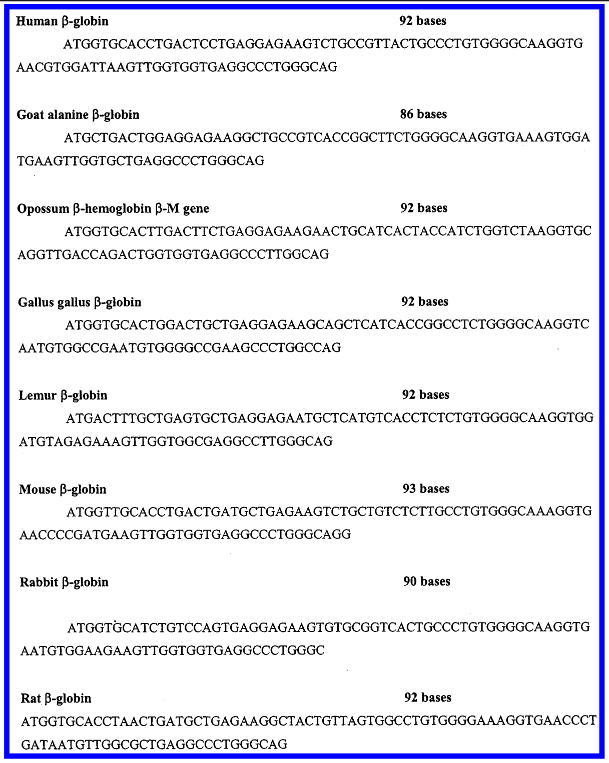
Received September 4, 2002

We consider a four-dimensional representation of DNA primary sequences by assigning to each of the four basic amino acids A, T, G, C directions along the four orthogonal coordinate axes. Advantages and limitations of the novel representation of DNA primary sequences are discussed, and the use of the 4-D representation is illustrated by constructing novel sequence invariants. Comparisons with the similarity/dissimilarity results based on 2-D and 3-D representations for a set of eight short DNA sequences corresponding to the first exon of beta globin in eight species, including human, are considered to illustrate the use of our novel sequence invariants based on the entries in derived sequence matrices restricted to a selected width of a band along the main diagonal.

INTRODUCTION

Graphical representations of DNA were introduced to facilitate comparison of DNA sequences and observing differences in their structure.[1,2] As pointed out by Leong and Morgenthaler,[1] 2-D plots are useful on three different levels. First, they give visual communication of the results of an analysis. Next, such plots help in checking for the presence of an effect by human eye rather than computer algorithm. Finally, plots are the primary tools for identifying unsuspected structures in data. From the various visual displays associated with DNA, the *random walk plot*, which takes a form of a trajectory over the Cartesian coordinate system, offers a useful display of the excess of one type of base over another type. Figure 1 of ref 3 illustrates this walk for the first exon of human β-globin (listed in Table 1); such a 2-D representation of DNA is obtained by assigning moves either left−right to A and G or up−down to C and T, respectively. Alternatively, one can consider that the occurrence of an A moves the curve to the right, C to the left, G downward, and T upward, as considered by Becker, Chambers, and Wilks.[4] There are two apparent disadvantages of such 2-D representations. First the choice of axes for various bases is arbitrary as there is no natural preference for any of several possible assignments of elementary moves along x and y axes. The second disadvantage is the loss of information associated with any repeating left−right or up−down movements that overlap. Moreover, the random walk trajectory may intersect itself, which then does not allow a reconstruction of the primary sequence of DNA from its graphical representation. Both these disadvantages have been addressed and successfully resolved. Randić et al.[5] considered a three-dimensional representation of DNA in which the four directions associated with tetrahedral angles have been used as the elementary directions for the four bases A, T, G, C.

Table 1. First Exon of β-Globin for Eight Vertebrates



Human β-globin **92 bases**
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTG
AACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG

Goat alanine β-globin **86 bases**
ATGCTGACTGGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGA
TGAAGTTGGTGCTGAGGCCCTGGGCAG

Opossum β-hemoglobin β-M gene **92 bases**
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGC
AGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG

Gallus gallus β-globin **92 bases**
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTC
AATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG

Lemur β-globin **92 bases**
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGG
ATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG

Mouse β-globin **93 bases**
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTG
AACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG

Rabbit β-globin **90 bases**
ATGGTGCATCTGTCCAGTGAGGAGAAGTGTGCGGTCACTGCCCTGTGGGGCAAGGTG
AATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC

Rat β-globin **92 bases**
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCT
GATAATGTTGGCGCTGAGGCCCTGGGCAG

One can observe that while in a 2-D representation once A is assigned to the x axis (say, for the move to the left), then the x and y axes no longer offer three equivalent directions for the remaining three bases. In contrast, the assignment of A to any of the four tetrahedral directions leaves the remaining three directions fully equivalent. As it has been demonstrated,[5] the 3-D representation contains the alternative 2-D representation as projections on various coordinate planes. The 3-D random walk over a tetrahedral grid, however, can also intersect itself and also remains associated with the loss of information that accompanies repeated left−right or up−down moves. However, as Guo et al.[6] demonstrated, by changing the angles between the elementary

* Corresponding author e-mail: balabana@tamug.tamu.edu.
[†] Drake University.
[‡] Texas A&M University at Galveston.
[§] Affiliated with National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia. Home Fax: 1-515-292-8629.

10.1021/ci020051a CCC: $25.00 © 2003 American Chemical Society
Published on Web 01/23/2003

Table 2. 4-D Coordinates for the First 15 Bases of the First Exon of Human β-Globin and Goat β-Globin

human β-globin						goat β-globin					
no.	base	A	T	G	C	no.	base	A	T	G	C
1	A	1	0	0	0	1	A	1	0	0	0
2	T	1	1	0	0	2	T	1	1	0	0
3	G	1	1	1	0	3	G	1	1	1	0
4	G	1	1	2	0	4	C	1	1	1	1
5	T	1	2	2	0	5	T	1	2	1	1
6	G	1	2	3	0	6	G	1	2	2	1
7	C	1	2	3	1	7	A	2	2	2	1
8	A	2	2	3	1	8	C	2	2	2	2
9	C	2	2	3	2	9	T	2	3	2	2
10	C	2	2	3	3	10	G	2	3	3	2
11	T	2	3	3	3	11	C	2	3	3	3
12	G	2	3	4	3	12	T	2	4	3	3
13	A	3	3	4	3	13	G	2	4	4	3
14	C	3	3	4	4	14	A	3	4	4	3
15	T	3	4	4	4	15	G	3	4	5	3

directions one can obtain a 2-D graphical representation of DNA primary sequence that does not have the deficiencies previously mentioned. The approach of Guo et al. can, in principle, be extended also to the 3-D graphical representation of DNA.

There are clear advantages of simple 2-D and 3-D representations based on Cartesian and tetrahedral grids, respectively, in that they offer a simple visual representation of DNA in the case of 2-D representation and with computer graphics even 3-D can be visually examined. The change of the angles between the elementary directions to avoid overlap introduces an arbitrary element—the selection of novel angles. This is not so critical for visual inspection of DNA sequences, but it will affect the numerical characterization of such graphical representations. It seems desirable to avoid intersections and repeated overlapping moves of 2-D and 3-D graphical representations of DNA and also to avoid arbitrary modifications of such moves that may affect adversely the numerical analysis of graphical DNA plots. As we will demonstrate in this contribution, this is possible by adopting a four-dimensional representation of DNA primary sequences—of course at the price of partly losing a visual grasp of such sequences. Thus we gain in the mathematical clarity and definiteness of DNA representations and we get rid of the geometrical arbitrariness and ambiguities of 2-D and 3-D graphical representations of DNA.

FOUR-DIMENSIONAL REPRESENTATION OF DNA

We will illustrate the four-dimensional characterization of DNA on the first exon of human β-globin listed at the top of Table 1. In 4-D space points, vectors, and directions have four components, and we will assign the following basic elementary directions to the four bases:

A (1, 0, 0, 0)
T (0, 1, 0, 0)
G (0, 0, 1, 0)
C (0, 0, 0, 1)

Because the four directions of 4-D space are fully equivalent, the above selection is equivalent to any other permutation of labels and directions, hence this selection should not be viewed as introducing any arbitrary decision to influence numerical analysis that follows. In Table 2 we have indicated 4-D coordinates for the first 15 bases of the first exon of human β-globin and goat β-globin in order to illustrate differences between the initial stages of the two sequences.

Since we have no graphical representation to associate with a random walk in 4-D space, we have constructed the distance matrix for each such random walk in which any (i, j) entry is the Euclidean distance between corresponding points in 4-D space given by

$$\mathrm{D}ij = \sqrt{\{(\mathrm{A}_i - \mathrm{A}_j)^2 + (\mathrm{T}_i - \mathrm{T}_j)^2 + (\mathrm{G}_i - \mathrm{G}_j)^2 + (\mathrm{C}_i - \mathrm{C}_j)^2\}}$$

where A_i, T_i, G_i, C_i and A_j, T_j, G_j, C_j are the four coordinates of the ith and jth nucleic acid listed in Table 2 for cases i, $j = 1, ..., 15$ for both the human and goat first exon of β-globin.

In Table 3 we show a fragment of the 92 × 92 distance matrix corresponding to the four-dimensional representation of the first exon of the human β-globin. This matrix, which summarizes all the distances between the nucleic bases in the four-dimensional representation, forms our basic information on individual DNA sequences. Instead of visual inspection of DNA graphic representations we have now the opportunity for mathematical analysis and comparison of distance matrices belonging to different DNA sequences. Constructing suitable structural, sequence, and matrix invariants facilitates such comparisons. In this contribution we will point out novel sequence invariants that can be extracted from the distance matrix, such as that shown in Table 3.

SEQUENCE INVARIANTS

An invariant, by definition, is a quantity (numerical value) that is independent of any graphical representation of a structure or assignment of labels used to construct the matrix representing a structure. Similarly, sequence invariants are quantities that can be attributed to a sequence independently of labels used to identify individual elements of a sequence or any graphical representation of a sequence. There is, however, an important difference between sequence invariants and matrix invariants in that in a sequence the adjacency of neighboring elements is firmly fixed and is not lost in alternative graphical or labeling uses. Thus, for instance the first two bases A T of the first exon of the human β-globin will remain adjacent regardless of labels used for A and T, which need not be 1 and 2.

The sequence invariant that we will introduce will be explained on the 15 × 15 fragment of the distance matrix of Table 3. One can observe first that in each row of the table (and because the matrix is symmetrical the same holds for the corresponding columns) the entries increase from left to right. This is important to recognize because if the same fragment of the distance matrix is based on some other labeling of nucleic acid bases, this will no longer be the case. However, knowing this important property of naturally labeled sequences by consecutive numbers, any arbitrary labeled matrix can easily be rearranged by first placing the smallest entry 1 next to the main diagonal, then the next smallest entry (either $\sqrt{2}$, or $\sqrt{4} = 2$) next to 1, and so on till all entries are arranged in increasing order as we move from the diagonal zero to the right.

Table 3. A 15 × 15 Fragment of the 92 × 92 Distance Matrix Belonging to the 4-D Representation of the First Exon of Human β-Globin

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	1	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{8}$	$\sqrt{13}$	$\sqrt{14}$	$\sqrt{15}$	$\sqrt{18}$	$\sqrt{23}$	$\sqrt{28}$	$\sqrt{35}$	$\sqrt{38}$	$\sqrt{45}$	$\sqrt{50}$
2		0	1	$\sqrt{4}$	$\sqrt{5}$	$\sqrt{10}$	$\sqrt{11}$	$\sqrt{12}$	$\sqrt{15}$	$\sqrt{20}$	$\sqrt{23}$	$\sqrt{30}$	$\sqrt{33}$	$\sqrt{40}$	$\sqrt{45}$
3			0	1	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{6}$	$\sqrt{7}$	$\sqrt{10}$	$\sqrt{15}$	$\sqrt{18}$	$\sqrt{23}$	$\sqrt{26}$	$\sqrt{33}$	$\sqrt{38}$
4				0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	$\sqrt{7}$	$\sqrt{12}$	$\sqrt{15}$	$\sqrt{18}$	$\sqrt{21}$	$\sqrt{28}$	$\sqrt{33}$
5					0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{6}$	$\sqrt{11}$	$\sqrt{12}$	$\sqrt{15}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{28}$
6						0	1	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{10}$	$\sqrt{11}$	$\sqrt{12}$	$\sqrt{15}$	$\sqrt{22}$	$\sqrt{25}$
7							0	1	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{6}$	$\sqrt{7}$	$\sqrt{10}$	$\sqrt{15}$	$\sqrt{18}$
8								0	1	$\sqrt{4}$	$\sqrt{5}$	$\sqrt{6}$	$\sqrt{7}$	$\sqrt{12}$	$\sqrt{15}$
9									0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	$\sqrt{7}$	$\sqrt{10}$
10										0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	$\sqrt{5}$
11											0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$
12												0	1	$\sqrt{2}$	$\sqrt{3}$
13													0	1	$\sqrt{2}$
14														0	1
15															0

Table 4. Expressions for Band Average Widths 1−14 for the 15 × 15 Fragment of the Distance Matrices for the First Exons of the Eight Species of Table 1

band	human	goat	opossum	gallus
1	14/14	14/14	14/14	14/14
2	$(11\sqrt{2} + 2\sqrt{4})/13$	$(13\sqrt{2})/13$	$(11\sqrt{2} + 2\sqrt{4})/13$	$(11\sqrt{2} + 2\sqrt{4})/13$
3	$(6\sqrt{3} + 6\sqrt{5})/ 12$	$(11\sqrt{3} + \sqrt{5})/12$	$(6\sqrt{3} + 6\sqrt{5})/12$	$(6\sqrt{3} + 6\sqrt{65})/12$
4	$(4\sqrt{4} + 4\sqrt{6} + \sqrt{8} + 2\sqrt{10})/11$	$(5\sqrt{4} + 6\sqrt{6})/11$	$(3\sqrt{4}\ 6\sqrt{6} + \sqrt{8} + \sqrt{10})/11$	$(3\sqrt{4} + 6\sqrt{6} + \sqrt{8} + \sqrt{10})/11$
5	$(\sqrt{5} + 5\sqrt{7} + 3\sqrt{11} + \sqrt{13})/10$	$(7\sqrt{7} + 3\sqrt{9})/10$	$(7\sqrt{7} + \sqrt{9} + \sqrt{11} + \sqrt{13})/10$	$(8\sqrt{7} + \sqrt{11} + \sqrt{13})/10$
6	$(3\sqrt{10} + 5\sqrt{12} + \sqrt{14})/9$	$(7\sqrt{10} + 2\sqrt{12})/9$	$(6\sqrt{10} + 2\sqrt{12} + \sqrt{14})/9$	$(7\sqrt{10} + \sqrt{12} + \sqrt{14})/9$
7	$8\sqrt{15}/8$	$(7\sqrt{13} + \sqrt{15})/8$	$(4\sqrt{13} + 4\sqrt{15})/8$	$(5\sqrt{13} + 3\sqrt{15})/8$
8	$(5\sqrt{18} + \sqrt{20} + \sqrt{22})/7$	$(\sqrt{16} + 6\sqrt{18})/7$	$(\sqrt{16} + 6\sqrt{18})/7$	$(\sqrt{16} + 6\sqrt{18})/7$
9	$(\sqrt{21} + 3\sqrt{23} + 2\sqrt{25})/6$	$(2\sqrt{21} + 4\sqrt{23})/6$	$(2\sqrt{21} + 4\sqrt{23})/6$	$(2\sqrt{21} + 2\sqrt{23} + 2\sqrt{25})/6$
10	$(\sqrt{26} + 3\sqrt{28} + \sqrt{30})/5$	$(\sqrt{26} + 4\sqrt{28})/5$	$(3\sqrt{26} + 2\sqrt{30})/5$	$(2\sqrt{26} + \sqrt{28} + \sqrt{30} + \sqrt{34})/5$
11	$(3\sqrt{33} + \sqrt{35})/4$	$(\sqrt{31} + \sqrt{33} + 2\sqrt{35})/4$	$(\sqrt{31} + 3\sqrt{33})/4$	$(\sqrt{31} + \sqrt{33} + \sqrt{38} + \sqrt{39})/4$
12	$(2\sqrt{38} + \sqrt{40})/3$	$(2\sqrt{38} + \sqrt{42})/3$	$(2\sqrt{38} + \sqrt{40})/3$	$(\sqrt{38} + 2\sqrt{42})/3$
13	$(2\sqrt{45})/2$	$(\sqrt{45} + \sqrt{47})/2$	$(2\sqrt{45})/2$	$(2\sqrt{47})/2$
14	$(\sqrt{50})/1$	$(\sqrt{54})/1$	$(\sqrt{54})/1$	$(\sqrt{54})/1$

band	lemur	mouse	rabbit	rat
1	14/14	14/14	14/14	14/14
2	$(11\sqrt{2} + 2\sqrt{4})/13$	$(10\sqrt{2} + 3\sqrt{4})/13$	$(11\sqrt{2} + 2\sqrt{4})/13$	$(10\sqrt{2} + 3\sqrt{4})/13$
3	$(8\sqrt{3} + 3\sqrt{5} + \sqrt{9})/ 12$	$(5\sqrt{3} + 7\ \sqrt{5})/12$	$(6\sqrt{3} + 6\sqrt{5})/12$	$(4\sqrt{3} + 8\sqrt{5})/12$
4	$(3\sqrt{4} + 6\sqrt{6} + 2\sqrt{10})/11$	$(3\sqrt{4} + 4\sqrt{6} + 3\sqrt{8} + \sqrt{10})/11$	$(2\sqrt{4}\ 7\sqrt{6} + \sqrt{8} + \sqrt{10})/11$	$(\sqrt{4} + 7\sqrt{6} + \sqrt{8} + 2\sqrt{10})/11$
5	$(4\sqrt{7} + 2\sqrt{9} +4\ \sqrt{11})/10$	$(5\sqrt{7} + \sqrt{9} + 2\sqrt{11} + 2\sqrt{13})/10$	$(4\sqrt{7} + 3\sqrt{9} + 2\sqrt{11} + \sqrt{13})/10$	$(2\sqrt{7} + 4\sqrt{9} + 3\sqrt{11} + \sqrt{13})/10$
6	$(2\sqrt{10} + 4\sqrt{12} + 2\sqrt{14} + \sqrt{18})/9$	$(3\sqrt{10} + 4\sqrt{12} + \sqrt{14} + \sqrt{18})/9$	$(4\sqrt{10} + 2\sqrt{12} + 3\sqrt{14})/9$	$(\sqrt{10} + 4\sqrt{12} + 4\sqrt{14})/9$
7	$(5\sqrt{15} + \sqrt{19} + 2\sqrt{21})/8$	$(\sqrt{13} + 6\sqrt{15} + \sqrt{19})/8$	$(2\sqrt{13} + 54\sqrt{15} + \sqrt{19})/8$	$(5\sqrt{15} + \sqrt{17} + 2\sqrt{19})/8$
8	$(2\sqrt{10} + 4\sqrt{12} + 2\sqrt{14} + \sqrt{24})/7$	$(4\sqrt{18} + 2\sqrt{20} + \sqrt{22})/7$	$(4\sqrt{18} + 3\sqrt{20})/7$	$(3\sqrt{18} + 2\sqrt{20} + \sqrt{22} + \sqrt{26})/7$
9	$(4\sqrt{25} + 2\sqrt{27})/6$	$(\sqrt{21} + 4\sqrt{23} + \sqrt{25})/6$	$(4\sqrt{23} + \sqrt{25} + \sqrt{27})/6$	$(\sqrt{21} + 3\sqrt{23} + \sqrt{27} + \sqrt{29})/6$
10	$(\sqrt{28} + 2\sqrt{30} + \sqrt{34} + \sqrt{36})/5$	$(\sqrt{26} + 4\sqrt{28})/5$	$(5\sqrt{30})/5$	$(2\sqrt{26} + \sqrt{28} + 2\sqrt{30})/5$
11	$(2\sqrt{33} + 2\sqrt{39})/4$	$(\sqrt{31} + \sqrt{33} + 2\sqrt{35})/4$	$(\sqrt{35} + 3\sqrt{37})/4$	$(2\sqrt{31} + \sqrt{33} + \sqrt{35})/4$
12	$(\sqrt{40} + 2\sqrt{42})/3$	$(2\sqrt{38} + \sqrt{42})/3$	$(\sqrt{42} + \sqrt{43} + 4\sqrt{46})/3$	$(\sqrt{36} + 2\sqrt{38})/3$
13	$(2\sqrt{49})/2$	$(2\sqrt{45})/2$	$(\sqrt{51} + \sqrt{49})/2$	$(2\sqrt{43})/2$
14	$(\sqrt{60})/1$	$(\sqrt{52})/1$	$(\sqrt{58})/1$	$(\sqrt{50})/1$

In the following we will assume that all distance matrices considered have already been so ordered. We can now consider the sums of elements in diagonal entries parallel to the main diagonal, which consists of zeroes. In the case of Table 3 we obtain for the first few neighboring diagonal the following sums: 14, $(11\ \sqrt{2} + 2\ \sqrt{4})$, $(6\ \sqrt{3} + 6\ \sqrt{5})$, and so on. We will refer to these sums as band invariants of different width, specifically b-1, b-2, b-3 for 14, $(11\ \sqrt{2} + 2\ \sqrt{4})$, $(6\ \sqrt{3} + 6\ \sqrt{5})$, respectively. One can observe that these quantities are not matrix invariants but can always be extracted from any matrix in whatever form it is presented by considering adjacency between sequence elements. If the distance matrix is already in the canonical form based on assigning labels to nucleic acid bases sequentially, band invariants can readily be obtained by summing elements along each of the lines parallel to the main diagonal.

In Table 4 we show the band invariants based on 15 × 15 fragments of the DNA sequences of Table 1. As one immediately sees, there is considerable variation in the form and numerical values for bandwidth invariants belonging to different DNA sequences.

In Table 5 are listed the numerical values for the average bandwidths, which also show variations in the relative magnitudes of average bandwidth, except for the initial three bands and the terminal bandwidth which, because of truncation, have small number of terms contributing to partial sums,

Table 5. Numerical Values for the Band Average Widths 1−14 for the 15 × 15 Fragment of the Distance Matrices for the First Exons of the Eight Species of Table 1

band	human	goat	opossum	gallus	lemur	mouse	rabbit	rat
0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
2	1.5043	1.4142	1.5043	1.5043	1.5043	1.5494	1.5043	1.5494
3	1.9841	1.7741	1.9841	1.9841	1.9637	2.0261	1.9841	2.0681
4	2.4501	2.2452	2.4261	2.4261	2.4565	2.4950	2.4670	2.5727
5	2.9020	2.7520	2.8442	2.8088	2.9850	2.7955	2.9822	3.0847
6	3.3943	3.2293	3.2937	3.2602	3.5452	3.4669	3.4225	3.5539
7	3.8730	3.6390	3.7393	3.7058	4.1111	3.9003	3.8669	4.0257
8	4.3394	4.2080	4.2080	4.2080	4.5939	4.3722	4.3410	4.4045
9	4.8283	4.7247	4.7247	4.7928	5.0654	4.7943	4.8966	4.9252
10	5.2902	5.2530	5.2503	5.3595	5.6154	5.2530	5.4772	5.2888
11	5.7874	5.7861	5.7004	5.9304	5.9948	5.7861	6.0411	5.6990
12	6.2178	6.2699	6.2178	6.3753	6.4287	6.2699	6.6068	6.1096
13	6.7082	6.7819	6.7082	6.8557	7.0000	6.7082	7.0707	6.5574
14	7.0711	7.3484	7.3485	7.3485	7.7460	7.2111	7.6158	7.0711

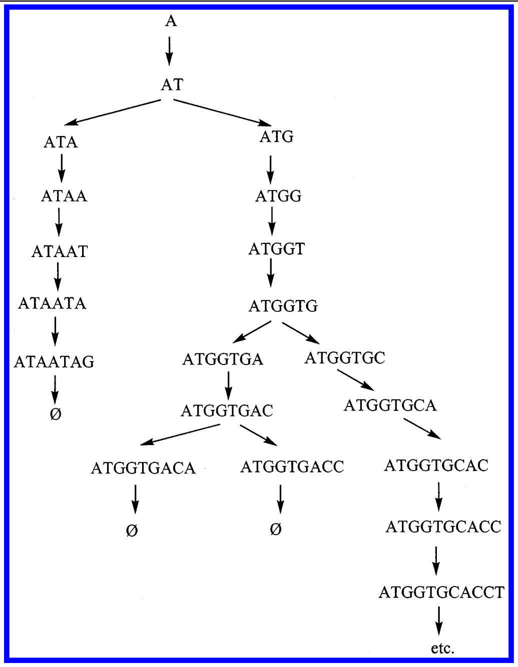
increasing the chance for coincidental values. There is a pair of degenerate values in each of the bands 8−10. If we restrict our attention to the data of Table 5, we are in fact confined to characterization of the *local* sequence structure, a fragment of the initial DNA sequence. We see that band invariants or bandwidth descriptors are sensitive to minor changes in DNA sequence. For example at the bandwidths 5−7 all the eight DNA sequences have different average distance matrix elements.

By looking at Table 1 we see that the Euclidean 4-D distance matrix is rich in information, but we should be also aware of the fact that besides a loss of information obtained by calculating the average values for matrix elements there is also some loss of information already associated with the distance matrix itself. An apparent loss of information is that of not knowing the initial labels A, T, G, and C. A DNA sequence in which one substitutes any pair of labels at the very beginning (when they appear for the first time) will necessarily generate the same distance matrix. However, if we know in which order the nucleic acid bases appear for the first time, does this suffice to fully reconstruct the sequence from its distance matrix? We believe that it does, as will be demonstrated in the next section on the 15 × 15 fragment of the distance matrix for the first exon of human β-globin.

SEQUENCE RECONSTRUCTION

Because we require information on the first appearance of any of the four nucleic acid bases we will classify all DNA sequences into 16 possible types, each designated by the order in which nucleic bases appear for the first time in a sequence. For example, all the eight DNA sequences of Table 1 are ATGC-type. The DNA primary sequence for human β-globin (shown for example in the Table 1 of ref 7), namely the segment from 62 205 to 63 628 has three exons. The remaining two exons belong to GCTA-class and TCGA-class. Assuming knowledge of the DNA class, we can start reconstruction of DNA sequence from the distance matrix of Table 1. Clearly, any ATGC-type sequence starts with AT, not with AG or AC, all of which are at the same distance of 1. The continuation of the sequence requires us to examine fragments ATA, ATT, and ATG, but not ATC, because the third in appearance is guanine (G) and not cytosine (C). The element (1, 3) of the distance matrix in Table 3 is $\sqrt{2}$, which eliminates sequence ATT for which the element (1, 3) would be $\sqrt{4}$. We continue with fragments ATA and ATG, knowing from Table 3 that we have to satisfy a distance matrix that requires (1, 4) and (2, 4) to be $\sqrt{5}$ and $\sqrt{4}$, respectively. It is not difficult to see that only fragments ATAA and ATG satisfy the two requirements. The process continues as illustrated on chart (Table 6), which after nine steps is reduced to a single possibility: ATGGTGCACC. In other words, among all 4^{10} possible DNA sequences that start with adenine (A), a number that is over 1 000 000, only the sequence ATGGTGCACC satisfies the 10 × 10 submatrix of the distance matrix for the first exon of the human β-globin.

Table 6. Initial Steps in the DNA Sequence Reconstruction from the 4-D Distance Matrix



Although this does not prove that the distance matrix is unique to a DNA sequence, except for the knowledge of the initial order in which adenine, cytosine, guanine, and thymine appear, it seems plausible to conjecture that the Euclidean 4-D distance matrix is unique to a DNA primary sequence, except for short sequences, like those of Table 6, having less than 10 nucleic acid bases or a dozen bases in other

Table 7. Similarity/Dissimilarity between the First Exons of β-Globin among Eight Species Based on Average Bandwidth 5, 10, and 15

width 5	human	goat	opossum	gallus	lemur	mouse	rabbit	rat
human	0	0.3416	0.0626	0.0962	0.0888	0.1310	0.0820	0.2398
goat		0	0.3057	0.2969	0.3884	0.3822	0.3929	0.5680
opossum			0	0.0354	0.1502	0.1045	0.1439	0.2974
gallus				0	0.1838	0.0934	0.1782	0.3267
lemur					0	0.2056	0.0218	0.1804
mouse						0	0.1986	0.3024
rabbit							0	0.1754
rat								0
width 10	human	goat	opossum	gallus	lemur	mouse	rabbit	rat
human	0	0.4776	0.2480	0.2803	0.5605	0.1636	0.2173	0.3738
goat		0	0.3281	0.3310	0.9327	0.5501	0.5841	0.8373
opossum			0	0.1416	0.3281	0.3141	0.1523	0.6016
gallus				0	0.7536	0.3572	0.3555	0.6324
lemur					0	0.5889	0.4321	0.4197
mouse						0	0.3228	0.3848
rabbit							0	0.3649
rat								0
width 15	human	goat	opossum	gallus	lemur	mouse	rabbit	rat
human	0	0.5597	0.3816	0.4718	0.9708	0.2216	0.8313	0.4266
goat		0	0.3708	0.3833	1.0697	0.5718	0.8213	0.9283
opossum			0	0.3476	0.9972	0.3520	0.7879	0.6812
gallus				0	0.8682	0.4474	0.5566	0.8307
lemur					0	0.8870	0.4925	1.0084
mouse						0	0.7602	0.4730
rabbit							0	1.0575
rat								0

cases. Our prime interest in is constructing sequence invariants, and clearly a matrix that is unique, or almost unique, may well serve as a basis for the design and extraction of novel sequence-related invariants, as are the bandwidth averages already outlined in this communication.

SIMILARITY/DISSIMILARITY BASED ON BANDWIDTH AVERAGES

It is possible to measure quantitatively the similarities and dissimilarities among the eight exons of Table 1 by considering the 15 × 15 fragment of the 92 × 92 distance matrix that was displayed in Table 3. Taking the average *n*th (with $n = 5$, 10, or 15) bandwidth from the main diagonal, one obtains a vector in *n*-space. The differences Δx_i between two vectors x_a and x_b (expressed as differences between Euclidean distances) represent measures of similarity/dissimilarity

$$x_{a,i} - x_{b,i} = \Delta x_i = (\Delta {x_1}^2 + \Delta {x_2}^2 + ... \Delta {x_n}^2)^{1/2}$$

where $i = 1, 2, ..., n$.

To demonstrate the utility of novel sequence descriptors we will consider similarities and dissimilarities among the eight exons of Table 1. In Table 7 we have listed the similarity/dissimilarity table between the first exons of β-globin among the eight species based on cumulative bandwidths of order 5, 10, and 15, that is based on vectors having 5, 10, and 15 components (listed in Table 7). The three different bandwidths that we consider correspond to invariants that consider local DNA fragments of length 5, 10, and 15, all confined to the first 15 DNA nucleic acid bases. A comparison of the three separate similarity/dissimilarity tables shows that local similarities may vary. For the shortest segments, two pairs {lemur, rabbit} and {opossum, gallus} show the largest similarities, but as we increase the length of the segments on which the similarity is based, only the pair {opossum, gallus} continues to remain similar, while the dissimilarity for the pair {lemur, rabbit} increases considerably. On the other hand, the similarity between the first exons of the pair {human, mouse} (that initially was not among the most similar ones) is only slightly affected by the increase of the length of the fragment considered and corresponds to the smallest numerical entry for the cumulative bandwidths based on all 15 components of the vectors in Table 5.

Table 8 presents the initial band average width for the eight species of Table 1.

In Table 9 we show the similarity/dissimilarity table for the first exons of β-globin among the eight species based on cumulative bandwidths of order 5 but using the full length of the exons listed in Table 1, that is, based on vectors having from 86 to 92 components. First, we observe that even though we are comparing DNA sequences of somewhat different lengths the methodology of computing the similarity/dissimilarity indices is the same. This may be contrasted to comparisons based on DNA codes, where already such considerations introduce difficulties due to different degrees of deletions or insertions required optimizing sequence alignments.

One can compare the similarity/dissimilarity between the first exon of β-globin among eight species based on the average bandwidth 5 using either the first 15 bases (the first part of Table 7) or the full DNA sequences indicated in Table 9. Interestingly, with the full DNA sequence, the most similar are rabbits and rats with the lowest value of 0.0243 followed by humans and rats with a value of 0.0303 and by mice and rats with a value of 0.0311.

The similarities based on the full DNA sequence are somewhat altered when the full DNA is compared with the initial part of the sequence, as illustrated in Figure 1. Entries that remain close to the line $y = x$ indicate pairs of DNA

Table 8. Initial Band Average Width for the Distance Matrices of the Eight Species of Table 1

band	human	total	goat	total
1	1	1	1	1
2	$(64\sqrt{2} + 26\sqrt{4})/90$	1.5834	$(61\sqrt{2} + 23\sqrt{4})/84$	1.5444
3	$(30\sqrt{3} + 54\sqrt{5} + 5\sqrt{9})/89$	2.1091	$(33\sqrt{3} + 45\sqrt{5} + 5\sqrt{9})/83$	2.0817
4	$(8\sqrt{4} + 43\sqrt{6} + 15\sqrt{8} + 21\sqrt{10} + \sqrt{16})/88$	2.6609	$(8\sqrt{4} + 42\sqrt{6} + 14\sqrt{8} + 17\sqrt{10} + \sqrt{16})/82$	2.6370
5	$(18\sqrt{7} + 24\sqrt{9} + 26\sqrt{11} + 15\sqrt{13} + 4\sqrt{17})/87$	3.1774	$(14\sqrt{7} + 25\sqrt{9} + 25\sqrt{11} + 14\sqrt{13} + 3\sqrt{17})/81$	3.1828
band	opossum	total	gallus	total
1	1	1	1	1
2	$(70\sqrt{2} + 20\sqrt{4})/90$	1.5444	$(65\sqrt{2} + 25\sqrt{4})/90$	1.5769
3	$(38\sqrt{3} + 50\sqrt{5} + \sqrt{9})/89$	2.0294	$(31\sqrt{3} + 53\sqrt{5} + 5\sqrt{9})/89$	2.1034
4	$(11\sqrt{4} + 51\sqrt{6} + 15\sqrt{8} + 11\sqrt{10})/88$	2.5470	$(7\sqrt{4} + 48\sqrt{6} + 17\sqrt{8} + 14\sqrt{10} +2\ \sqrt{16})/88$	2.6356
5	$(29\sqrt{7} + 30\sqrt{9} + 6\sqrt{11} + 11\sqrt{13} + \sqrt{17})/87$	3.0296	$(20\sqrt{7} + 36\sqrt{9} + 12\sqrt{11} + 14\sqrt{13} + 5\ \sqrt{17})/87$	3.1242
band	lemur	total	mouse	total
1	1	1	1	1
2	$(69\sqrt{2} + 21\sqrt{4})/90$	1.5509	$(66\sqrt{2} + 25\sqrt{4})/91$	1.5751
3	$(31\sqrt{3} + 53\sqrt{5} + 5\sqrt{9})/89$	2.1034	$(34\sqrt{3} + 50\sqrt{5} + 6\sqrt{9})/90$	2.0966
4	$(8\sqrt{4} + 44\sqrt{6} + 14\sqrt{8} + 21\sqrt{10} + \sqrt{16})/88$	2.6566	$(10\sqrt{4} + 43\sqrt{6} + 16\sqrt{8} + 19\sqrt{10} + \sqrt{16})/89$	2.6367
5	$(17\sqrt{7} + 25\sqrt{9} + 24\sqrt{11} + 15\sqrt{13} + 6\sqrt{17})/87$	3.2000	$(18\sqrt{7} + 29\sqrt{9} + 22\sqrt{11} + 15\sqrt{13} + 4\sqrt{17})/88$	3.1610
band	rabbit	total	rat	total
1	1	1	1	1
2	$(64\sqrt{2} + 24\sqrt{4})/88$	1.5740	$(65\sqrt{2} + 25\sqrt{4})/90$	1.5769
3	$(25\sqrt{3} + 57\sqrt{5} + 5\sqrt{9})/87$	2.1351	$(31\sqrt{3} + 52\sqrt{5} + 6\sqrt{9})/89$	2.1120
4	$(6\sqrt{4} + 42\sqrt{6} + 15\sqrt{8} + 22\sqrt{10} + \sqrt{16})/86$	2.6846	$(9\sqrt{4} + 43\sqrt{6} + 14\sqrt{8} + 21\sqrt{10} + \sqrt{16})/88$	2.6515
5	$(14\sqrt{7} + 27\sqrt{9} + 22\sqrt{11} + 17\sqrt{13} + 5\sqrt{17})/87$	3.2108	$(19\sqrt{7} + 27\sqrt{9} + 23\sqrt{11} + 15\sqrt{13} + 3\sqrt{17})/87$	3.1495

Table 9. Similarity/Dissimilarity between the First Exons of β-Globin among Eight Species Based on the Average Bandwidth 5 Using the Full DNA Sequences

width 5	human	goat	opossum	gallus	lemur	mouse	rabbit	rat
human	0	0.2066	0.0402	0.0494	0.0536	0.0595	0.0329	0.0303
goat		0	0.2158	0.2526	0.1852	0.1527	0.1754	0.1821
opossum			0	0.0494	0.0345	0.0828	0.0505	0.0577
gallus				0	0.0823	0.1045	0.0791	0.0735
lemur					0	0.0704	0.0405	0.574
mouse						0	0.0375	0.0311
rabbit							0	0.0243
rat								0

sequences which have not varied much after the initial changes. Such are for instance the pairs {lemur, rabbit}, {opossum, gallus}, {human, opossum}, {gallus, mouse}, and {opossum, mouse}. Entries that are at greater distance from the line $y = x$ indicate DNA sequences that show considerable variations not only at the initial fragments, but also throughout their length. Such is for instance the pair {goat, rat} and several other ones including {human, rat}. In Figure 1 we show a plot of bandwidths 1−5 for the initial fragment of length 15 of human DNA against the bandwidths 1−5 for the full length of human DNA. There is a very strong correlation of corresponding bandwidths suggesting that in the case of the exons of Table 1 the initial 15-length segments already have captured most of the variations in DNA basis composition. The same is true for the first exon of other species of Table 1. Each species is characterized by a similar correlation that can be approximated either by a straight line or a parabola having different coefficients and different constant terms. We present in Table 10 the statistical parameters (correlation coefficient r, standard deviation s, and Fisher factor F) for the eight species using both types of correlations.

CONCLUDING REMARKS

The aim of this contribution was to outline a novel numerical characterization of DNA sequences, which is based on a unique 4-D representation of DNA that is free of arbitrary conventions accompanying the 2-D graphical representation of DNA regarding the selection of the four directions for the four nucleic acid bases. Additionally, in contrast to the 2-D and 3-D graphical representations, both of which represent projections of the “walk” along the DNA sequence and are thus accompanied with some loss of information, this appears not to be the case with the matrix based on 4-D representation considered in the present contribution. As we have shown, particularly for longer DNA sequences, the sequence can be reconstructed from the distance/distance matrix, provided that we know the order in which A, G, C, T bases appear for the first time.

The bandwidth invariants, which can be viewed as components of a vector representing a DNA sequence, are easy to calculate and compare. An additional advantage of bandwidths is that they can be selected for a particular fragment of DNA of interest, as illustrated on the initial fragment of 15 nucleic acids of the first exon for several

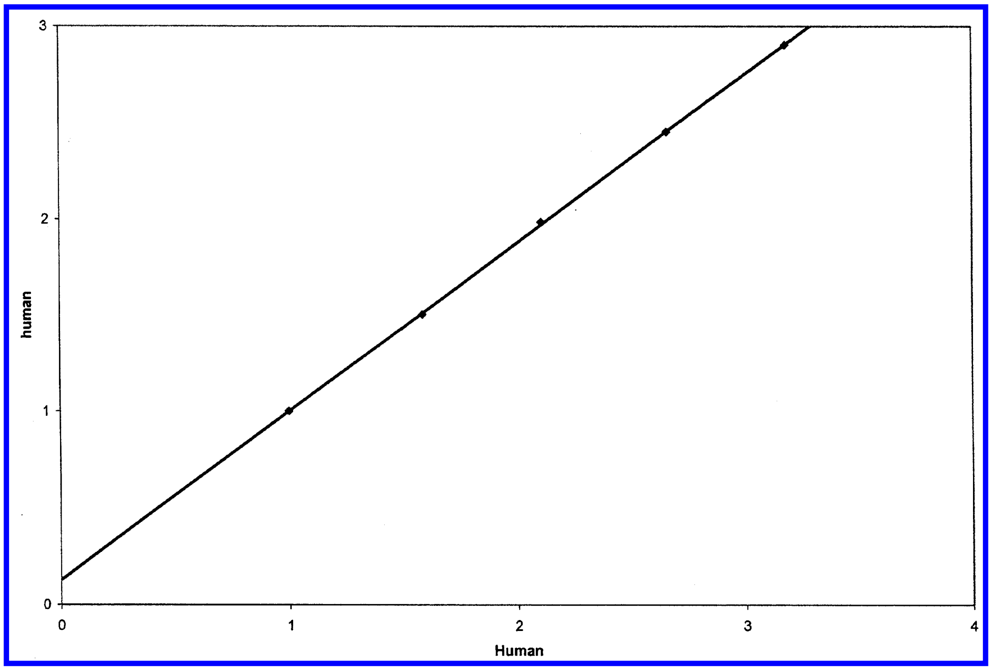


Figure 1. Plot of initial average bandwidths 1−5 as computed from the *complete DNA sequence* of the first exon for human β-globin (of Table 1) versus the corresponding bandwidths of the *truncated sequence to initial 15 nucleic acid bases*.

Table 10. Regression of Bandwidths 5 for Initial Fragment and Full DNA Sequence

	a (coeff)	*b* (const)	*r*	*s*	*F*
	Linear Equation: $y = ax + b$				
human	0.8744	0.1266	0.99995	0.0090	28022
goat	0.7944	0.1775	0.99814	0.0483	805
opossum	0.9109	0.1025	0.99962	0.0233	3915
gallus	0.8559	0.1576	0.99962	0.0228	3975
lemur	0.8940	0.1026	0.99968	0.0226	4755
mouse	0.8435	0.2071	0.99628	0.0718	401
rabbit	0.8903	0.0993	0.99970	0.02216	4940
rat	0.9661	0.0282	0.99990	0.01317	15552

	a (lin)	*b* (quad.)	*c* (const)	*r*	*s*	*F*
	Quadratic Equation: $y = ax + bx^2 + c$					
human	0.8984	−0.0058	0.1048	0.99996	0.0100	11249
goat	0.5132	0.0672	0.4316	0.99963	0.0265	1341
opossum	1.0511	−0.0348	−0.0210	0.99988	0.0158	4252
gallus	0.8798	0.0179	0.1051	0.99998	0.0069	26834
lemur	0.8004	0.0223	0.1874	0.99982	0.0212	2708
mouse	1.2805	−0.1050	−0.1868	0.99948	0.0329	960
rabbit	0.7740	0.0276	0.2048	0.99990	0.0155	5066

species. Comparisons of similarity/dissimilarly based on bandwidths show some differences with similar tables based on the leading eigenvector,[7] the count of bases at different distances,[8] or the use of average distances between bases,[9] thus indicating that our novel invariants have captured distinct structural features of DNA sequences in comparison with sequence invariants considered earlier. It is premature to speculate which set of invariants will show better use in different applications, just as it is not possible to foresee the best selection of molecular descriptors, such as topological indices, in QSAR (Quantitative Structure−Activity Relationship) studies. However, without having several sequence descriptors at our disposal, we would be deprived of considering quantitative sequence similarity and other applications that may emerge in comparative DNA studies. The situation is well illustrated by some recent applications of molecular descriptors in drug design, clustering of, and searching combinatorial libraries—all illustrations of applications that have not been anticipated when novel topological indices were introduced.[10] As outlined by Lahana and coworkers,[11] topological indices have recently found novel uses in drug design. Using a selection of topological indices they successfully screened a combinatorial library of over 280 000 virtual compounds to select two dozen compounds for further analysis. After further considerations five of the compounds were synthesized and tested for immunosuppressive activity. One of these compounds was found to be almost 100 times more active than the starting lead compound. In another study, Andrade and co-workers[12] used the connectivity index[13] and Balaban's distance index[14] to cluster transfer RNAs. Two main groups of t-RNAs that correspond to the biosynthetic amino acid pathways that correspond to the coevolution theory of the genetic code with new biosynthetic pathways for amino acids[15] were obtained, illustrating thus yet another novel use of weighted molecular descriptors. In a third publication, Flower[16] developed a computer program for the analysis of chemical diversity based on molecular descriptors. It selects diverse subsets from a larger collection of chemicals combining a maximum dissimilarity search and a general multidimensional measure of similarity based on combination of different molecular descriptors, including various connectivity indices.

Hence, it depends on users how to use mathematical descriptors of molecular structure and with recent contributions to construction of sequence invariants, future applications of molecular descriptors need not be limited to small and medium chemical structures but can be extended to DNA sequences and larger molecular systems of interest in molecular biology.

REFERENCES AND NOTES

(1) For reviews of graphical representations and analysis of DNA, see: Leong, P. M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Cabios* **1995**, *11*, 503−507.

(2) For reviews of graphical representations and analysis of DNA, see: Roy, A.; Raychaudhury, C.; Nandy, A. Novel techniques of graphical

representations and analysis of DNA sequences − A review. *J. Biosci.* **1998**, *23*, 55−71.

(3) Randić, M.; Nandy, A.; Basak, S. C. On numerical characterization of DNA primary sequences. *J. Math. Chem*. Submitted for publication.

(4) Becker, R. A.; Chambers, J. M.; Wilks, A. R. *The New S Language*; Wadsworth & Brooks/Cole: Pacific Grove, CA, 1988.

(5) 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.* **2001**, *41*, 1235−1244.

(6) Guo, X.; Randić, M.; Basak, S. C. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.* **2001**, *350*, 106−112.

(7) Randić, M.; Vračko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 599−606.

(8) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−46.

(9) Randić, M.; Basak, S. C. Characterization of DNA primary sequences based on the average distance between bases. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 561−568.

(10) Randić, M. Topological indices. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Editor-in-chief; Wiley: London, 1998; pp 3018−3032. For a summary on recent computations of molecular descriptors please consult the following: Katritizky, A. R.; Lobanov, V.; Karelson, M. *CODESSA* (Comprehensive Descriptors for Structural and Statistical Analysis); University of Florida, Gainesville, FL. *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Sci. Publ.: The Netherlands, 1999. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors* (*Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Vol. 11.

(11) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h, R.; Buelow, R. Computer-assisted rational design of immuno-suppressive compounds. *Nat. Biotech.* **1998**, *16*, 748−752.

(12) Bermudez, C. I.; Daza, E. E.; Andrade, E. Characterization and comparison of *Escherichia coli* transfer RNAs by graph theory based on secondary structure. *J. Theor. Biol.* **197**, 193−205.

(13) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976. Randic, M. The connectivity index 25 years after. *J. Mol. Graphics Modelling* **2001**, *1*, 19−35.

(14) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(15) Wong, J. T. A co-evolution theory of the genetic code. *Proc. Nat. Acad. Sci. U.S.A.* **1975**, *72*, 1909−1912.

(16) Flower, D. R. DISSIM: A program for the analysis of chemical diversity. *J. Mol. Graphics Modelling* **1998**, *16*, 239−253.

CI020051A