On the Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases

Milan Randić,*[†,#] Xiaofeng Guo,[§,#] and Subhash C. Basak[#]

Department of Mathematics & Computer Science, Drake University, Des Moines, Iowa 50311, Institute of Mathematics and Physics, Xinjiang University, Wulumuqi Xinjiang 830046, P. R. China, and Natural Resources Research Institute, University of Minnesota at Duluth, Duluth, Minnesota 55811

Received August 12, 2000

We consider construction of a set of smaller 4 × 4 matrices to represent DNA primary sequences which are based on enumeration of all 64 triplets of nucleic acids bases. The leading eigenvalue from the constructed matrices has been selected as an invariant for construction of a vector to characterize DNA. Additional invariants considered of the derived condensed matrices of DNA include a 64-component vector, the components of which consist of ordered triplets XYZ, with X, Y, Z = A, C, G, T. Construction of similarity/dissimilarity tables based on different invariants for a set of sequences of DNA belonging to the first exon of the *β*-globin gene of eight species illustrates the utility of newly formulated invariants for DNA.

INTRODUCTION

Compilation of DNA primary sequence data continues unabated and tends to overwhelm us with voluminous outputs that increase daily. Comparison of primary sequences of different DNA strands remains one of the important aspect of the analysis of DNA data banks. Considerable attention was given to this problem[1] which has led to the development of suitable software for the comparison and manipulation of sequences.[2] Almost all such comparisons are based on the comparison of strings using ideas initially proposed by Levenshtein,[3] who "presented the earliest known use of a distance function that is appropriate in the presence of insertion and deletions errors. His distance function and generalizations of it play a major role in sequence comparisons."[3] As is well-known, string comparisons are computer intensive, and despite the fact that practical schemes for sequence comparison have been outlined, there are a number of steps in such approaches that involve arbitrary decisions, e.g., decisions on the relative weights of different elementary string operations: deletion−insertions, substitution, and penalties for unacceptable alignments.

Recently we have introduced an alternative approach to DNA sequence comparisons which, instead of using string comparisons, considers a set of invariants of DNA sequence and uses these invariants for establishing the degree of similarity/dissimilarity among DNA sequences. We are at an early stage of the developments of this novel approach which involves a number of as yet unresolved questions. In particular, questions that need our attention are as follows: how to obtain suitable invariants to characterize DNA sequences and how to select invariants suitable for sequence comparisons. In this contribution we have proposed a set of novel invariants for characterization of DNA and provided an illustration of their utility by making a comparison between eight DNA sequences belonging to the same gene in different species.

DNA SEQUENCE INVARIANTS

The basic task is that of associating with a DNA sequence, such as the sequence of Table 1, a set of sequence invariants. A sequence invariant, as considered here, is a number independent of the labels A, C, G, T standing for adenine (A), cytosine (C), guanine (G), and thymine (T). A trivial sequence invariant is the length of a sequence. What else? In recent papers on the characterization of DNA primary sequences[4−9] we have introduced a number of novel sequence invariants. The route to these various invariants involved first the construction of a matrix associated with DNA and then calculation of suitable matrix invariants. Two kinds of matrices have been considered: (1) matrices in which an individual entry corresponds to an individual pair of bases[4,5,8,9] and (2) matrices in which entries summarize information of different X−Y pairs of bases.[6,7,9] Matrices that summarize information of X−Y pairs have been referred to as condensed matrices. They are 4 × 4 size, because there are only 16 possible (ordered) pairs that four letters can generate: AA, AC, AG, AT, CA, CC, etc. This leads to 10 distinct pairs of bases (AA, AC, AG, AT, CC, CG, CT, GG, GT, and TT) to be considered for the construction of invariants which are independent of the ordering of bases. Invariants considered for these matrices included the average matrix element and the leading eigenvalues. The study most closely related to the present work is the characterization of DNA primary sequences by condensed matrices that give the count of the frequency of occurrence of pairs of bases, such as AC, at various distances apart.[6] We want to elaborate on this topic by considering the frequencies of occurrence of triplets of nucleic bases.

THE FREQUENCIES OF TRIPLETS OF NUCLEIC ACID BASES

In a DNA sequence of four letters, there are 64 possible triplets (subsequences of length 3) that can occur, starting

[*] Corresponding author phone: (515)292-7411; fax: (515)292-8629. Present address: 3225 Kingman Road, Ames, IA 50014.
[†] Drake University.
[§] Xinjiang University.
[#] University of Minnesota at Duluth.

10.1021/ci000120q CCC: $20.00 © 2001 American Chemical Society
Published on Web 03/14/2001

Table 1: Human β-Globin Gene of Length 1424[a]

ATGGTGCACC	TGACTCCTGA	GGAGAAGTCT	GCCGTTACTG
CCCTGTGGGG	CAAGGTGAAC	GTGGATGAAG	TTGGTGGTGA
GGCCCTGGGC	AGGTTGGTAT	CAAGGTTACA	AGACAGGTTT
AAGGAGACCA	ATAGAAACTG	GGCATGTGGA	GACAGAGAAG
ACTCTTGGGT	TTCTGATAGG	CACTGACTCT	CTCTGCCTAT
TGGTCTATTT	TCCCACCCTT	AGGCTGCTGG	TGGTCTACCC
TTGGACCCAG	AGGTTCTTTG	AGTCCTTTGG	GGATCTGTCC
ACTCCTGATG	CTGTTATGGG	CAACCCTAAG	GTGAAGGCTC
ATGGCAAGAA	AGTGCTCGGT	GCCTTTAGTG	ATGGCCTGGC
TCACCTGGAC	AACCTCAAGG	GCACCTTTGC	CACACTGAGT
GAGCTGCACT	GTGACAAGCT	GCACGTGGAT	CCTGAGAACT
TCAGGGTGAG	TCTATGGGAC	CCTTGATGTT	TTCTTTCCCC
TTCTTTTCTA	TGGTTAAGTT	CATGTCATAG	GAAGGGGAGA
AGTAACAGGG	TACAGTTTAG	AATGGGAAAC	AGACGAATGA
TTGCATCAGT	GTGGAAGTCT	CAGGATCGTT	TTAGTTTCTT
TTATTTGCTG	TTCATAACAA	TTGTTTTCTT	TTGTTTAATT
CTTGCTTTCT	TTTTTTTTCT	TCTCCGCAAT	TTTTACTATT
ATACTTAATG	CCTTAACATT	GTGTATAACA	AAAGGAAATA
TCTCTGAGAT	ACATTAAGTA	ACTTAAAAAA	AAACTTTACA
CAGTCTGCCT	AGTACATTAC	TATTTGGAAT	ATATGTGTGC
TTATTTGCAT	ATTCATAATC	TCCCTACTTT	ATTTTCTTTT
ATTTTTAATT	GATACATAAT	CATTATACAT	ATTTATGGGT
TAAAGTGTAA	TGTTTTAATA	TGTGTACACA	TATTGACCAA
ATCAGGGTAA	TTTTGCATTT	GTAATTTTAA	AAAATGCTTT
CTTCTTTTAA	TATACTTTTT	TGTTTATCTT	ATTTCTAATA
CTTTCCCTAA	TCTCTTTCTT	TCAGGGCAAT	AATGATACAA
TGTATCATGC	CTCTTTGCAC	CATTCTAAAG	AATAACAGTG
ATAATTTCTG	GGTTAAGGCA	ATAGCAATAT	TTCTGCATAT
AAATATTTCT	GCATATAAAT	TGTAACTGAT	GTAAGAGGTT
TCATATTGCT	AATAGCAGCT	ACAATCCAGC	TACCATTCTG
CTTTTATTTT	ATGGTTGGGA	TAAGGCTGGA	TTATTCTGAG
TCCAAGCTAG	GCCCTTTTGC	TAATCATGTT	CATACCTCTT
ATCTTCCTCC	CACAGCTCCT	GGGCAACGTG	CTGGTCTGTG
TGCTGGCCCA	TCACTTTGGC	AAAGAATTCA	CCCCACCAGT
GCAGGCTGCC	TATCAGAAAG	TGGTGGCTGG	TGTGGCTAAT
GCCCTGGCCC	ACAAGTATCA	CTAA	

[a] Nucleic acids are grouped in group of tens.

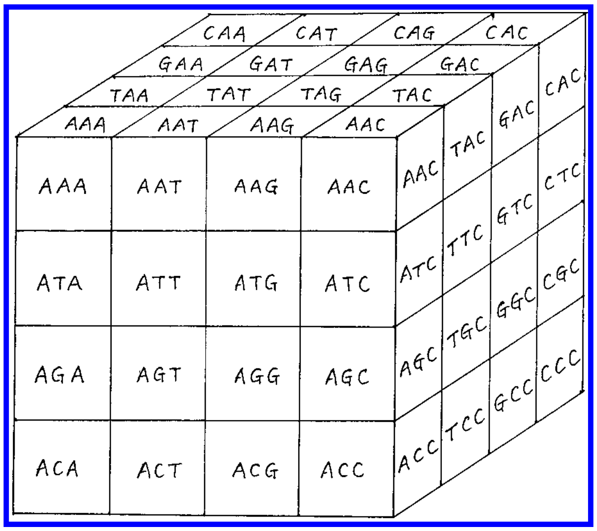

Figure 1. A 4 × 4 × 4 "cubic matrix" with 64 entries which indicate the frequencies of occurrence of all the 64 triplets in a DNA sequence, respectively.

Table 2: Three Groups of Four 4 × 4 Matrices, {M_1, M_2, M_3, M_4}, {M_5, M_6, M_7, M_8}, and {M_9, M_{10}, M_{11}, M_{12}} Listing All 64 Possible XYZ Entries, Where X, Y, Z = A, C, G, T

M_1				M_2			
AAA	AAT	AAG	AAC	TAA	TAT	TAG	TAC
ATA	ATT	ATG	ATC	TTA	TTT	TTG	TTC
AGA	AGT	AGG	AGC	TGA	TGT	TGG	TGC
ACA	ACT	ACG	ACC	TCA	TCT	TCG	TCC
M_3				M_4			
GAA	GAT	GAG	GAC	CAA	CAT	CAG	CAC
GTA	GTT	GTG	GTC	CTA	CTT	CTG	CTC
GGA	GGT	GGG	GGC	CGA	CGT	CGG	CGC
GCA	GCT	GCG	GCC	CCA	CCT	CCG	CCC
M_5				M_6			
AAA	AAT	AAG	AAC	ATA	ATT	ATG	ATC
TAA	TAT	TAG	TAC	TTA	TTT	TTG	TTC
GAA	GAT	GAG	GAC	GTA	GTT	GTG	GTC
CAA	CAT	CAG	CAC	CTA	CTT	CTG	CTC
M_7				M_8			
AGA	AGT	AGG	AGC	ACA	ACT	ACG	ACC
TGA	TGT	TGG	TGC	TCA	TCT	TCG	TCC
GGA	GGT	GGG	GGC	GCA	GCT	GCG	GCC
CGA	CGT	CGG	CGC	CCA	CCT	CCG	CCC
M_9				M_{10}			
AAA	TAA	AAG	CAA	AAT	TAT	GAT	CAG
ATA	TTA	GTA	CTA	ATT	TTT	GTT	CTG
AGA	TGA	GGA	CGA	AGT	TGT	GGT	CGG
ACA	TCA	GCA	CCA	ACT	TCT	GCT	CCG
M_{11}				M_{12}			
AAG	TAG	GAG	CAG	AAC	TAC	GAC	CAC
ATG	TTG	GTG	CTG	ATC	TTC	GTC	CTC
AGG	TGG	GGG	CGG	AGC	TGC	GGC	CGC
ACG	TCG	GCG	CCG	ACC	TCC	GCC	CCC

from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC, etc. We introduce a 4 × 4 × 4 cubic matrix with 64 entries which denote the frequencies of occurrence of all the 64 triplets in a DNA sequence, respectively (see Figure 1). For the cubic matrix, three groups of four 4 × 4 matrices, {M_1, M_2, M_3, M_4}, {M_5, M_6, M_7, M_8}, {M_9, M_{10}, M_{11}, M_{12}}, can be obtained, each group of which contain all entries of the cubic matrix (see Table 2). Usually we take the group of four 4 × 4 matrices {M_1, M_2, M_3, M_4} as the representative of the

Table 3: Count of Triplets XYZ for the DNA Sequence of Table 1 Broken Down into Three Exons and Two Introns Segments

Exon 1−92

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	0	0	3	1	A	0	0	0	1	A	3	1	3	1	A	1	0	1	1
T	0	0	2	0	T	1	0	1	0	T	0	2	6	1	T	0	0	6	1
G	1	2	3	0	G	5	1	6	3	G	2	4	3	3	G	0	2	0	0
C	0	1	2	1	C	0	1	0	1	C	3	0	0	3	C	0	4	1	2

Exon 223−445

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	1	0	6	3	A	1	1	1	1	A	3	4	5	3	A	5	1	2	7
T	0	0	4	2	T	2	4	4	2	T	0	2	8	3	T	2	6	12	5
G	3	4	5	2	G	8	3	10	7	G	4	5	4	6	G	0	1	1	0
C	3	4	1	6	C	4	3	1	4	C	5	8	0	3	C	3	10	0	3

Exon 1296−1424

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	2	2	3	1	A	2	2	0	0	A	2	0	0	0	A	3	1	3	5
T	0	1	1	3	T	0	1	1	1	T	1	0	8	1	T	3	1	7	1
G	2	3	1	0	G	0	3	9	5	G	0	3	1	7	G	0	1	0	0
C	1	2	1	2	C	4	1	0	1	C	3	5	0	4	C	4	3	0	5

Intron 93−222

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	1	1	4	1	A	1	3	3	1	A	2	1	3	5	A	3	1	2	2
T	2	2	1	1	T	3	4	3	2	T	1	4	1	1	T	2	2	4	4
G	7	0	4	0	G	2	1	5	1	G	2	5	2	2	G	0	0	0	0
C	3	4	0	2	C	1	6	0	1	C	2	0	0	1	C	2	2	0	2

Intron 446−1295

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	20	32	12	9	A	35	32	7	17	A	9	10	5	3	A	10	22	11	4
T	34	32	18	12	T	32	68	19	29	T	12	16	9	5	T	14	30	10	8
G	6	13	11	5	G	10	20	10	16	G	10	7	10	4	G	1	1	0	1
C	18	9	1	5	C	13	32	1	8	C	11	11	0	4	C	6	10	1	7

Table 4: Exon-1 of the β-Globin Genes for Eight Species Considered, Including Human, Labeled from A (Human) to H (Rat)

A, human β-globin, 92 bases
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCC
GTTACTGCCCTGTGGGGCAAGGTGAACGTGG
ATTAAGTTGGTGGTGAGGCCCTGGGCAG

B, goat alanine β-globin, 86 bases
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACC
GGCTTCTGGGGCAAGGTGAAAGTGGATGAAGT
TGGTGCTGAGGCCCTGGGCAG

C, opossum β-hemoglobin β M-gene, 92 bases
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCA
TCACTACCATCTGGTCTAAGGTGCAGGTTGAC
CAGACTGGTGGTGAGGCCCTTGGCAG

D, gallus gallus β globin, 92 bases
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCAT
CACCGGCCTCTGGGGCAAGGTCAATGTGGCC
GAATGTGGGGCCGAAGCCCTGGCCAG

E, lemur β-globin, 92 bases
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTC
ACCTCTCTGTGGGGCAAGGTGGATGTAGAGAA
AGTTGGTGGCGAGGCCTTGGGCAG

F, mouse β-a-globin, 93 bases
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTG
TCTCTTGCCTGTGGGCAAAGGTGAACCCCGATG
AAGTTGGTGGTGAGGCCCTGGGCAGG

G, rabbit β-globin, 90 bases
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGT
CACTGCCCTGTGGGGCAAGGTGAATGTGGAAG
AAGTTGGTGGTGAGGCCCTGGGC

H, rat β-globin, 92 bases
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTA
GTGGCCTGTGGGGAAAGGTGAACCCTGATAATGT
TGGCGCTGAGGCCCTGGGCAG

cubic matrix. The four matrices contain not only the information about frequencies of occurrence of all triplets of a DNA sequence but also the information about the frequencies of occurrence of pairs and every letter in a DNA sequence. For example, the number of all TG-pairs in a DNA sequence is equal to the row sum of the third row in M_2 plus δ, where $\delta = 0$ if the last two letters of the DNA sequence are not TG and $\delta = 1$ otherwise. The frequency of occurrence of any pair in a DNA sequence can be obtained by the above method. In addition, the frequencies of occurrence of four letters A, T, G, C are respectively equal to the sum of all entries of M_1, M_2, M_3, M_4 plus δ, where δ are respectively equal to the number of A, T, G, C in the last two letters of the DNA sequence. The column sums of M_1, M_2, M_3, M_4 just denote the numbers of pairs of distance two in a DNA sequence.

We will illustrate the novelty of the approach introduced in this article by examining the first exon of human β-globin gene listed in Table 1. In Table 2 we have listed 64 possible XYZ triplets. Here X, Y, and Z stand for any of the four nucleic acid bases. In Table 3 we give the count of all 64 XYZ triplets for the human β-globin gene of Table 1. The first entry i of a triplet i, j, k is represented by the header (the common row and column single base A, T, G, or C) of the four 4 × 4 matrices. The j, k entries of the triplets consist of j (indicating the row) and k (indicating the column) entries of the four 4 × 4 matrices shown in Table 3. For example, AGT is found in the first group of 16 nucleic bases in the third row and second column (the entry AGT = 2). The first three sets of 4 × 4 matrices describe XYZ frequencies in exons 1−92, 223−445, and 1296−1242, respectively, and the last two sets correspond to the introns 93−222 and 446−1295, respectively. A close look at Table 3 shows some common features of exons and introns that are not easily

Table 5: Count of Triplets XYZ for Matrices M_1−M_4 for the Eight Exons of Table 4

Human

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	0	0	3	1	A	0	0	0	1	A	3	1	3	1	A	1	0	1	1
T	0	0	2	0	T	1	0	1	0	T	0	2	6	1	T	0	0	6	1
G	1	2	3	0	G	5	1	6	3	G	2	4	3	3	G	0	2	0	0
C	0	1	2	1	C	0	1	0	1	C	3	0	0	3	C	0	4	1	2

Goat

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	1	0	4	0	A	0	0	0	0	A	3	1	3	1	A	1	0	1	1
T	0	0	2	0	T	0	0	1	1	T	0	1	3	1	T	0	1	7	0
G	1	2	4	0	G	5	0	4	4	G	2	2	3	5	G	0	1	1	0
C	0	1	0	1	C	1	1	0	0	C	2	5	0	2	C	0	1	2	1

Opossum

A	A	T	G	C	T	A	T	G	C	G	A	T	G	C	C	A	T	G	C
A	0	0	2	1	A	1	0	0	1	A	2	0	3	3	A	0	2	3	2
T	0	0	1	2	T	0	0	3	1	T	0	1	4	1	T	2	3	4	0
G	3	0	4	0	G	4	0	5	3	G	1	6	0	2	G	0	0	0	0
C	0	5	0	2	C	1	3	0	0	C	4	0	0	1	C	2	1	0	1

Gallus

A	A	T	G	C	G	A	T	G	C	T	A	T	G	C	C	A	T	G	C
A	0	2	3	0	A	0	0	0	0	A	3	0	2	1	A	2	1	2	2
T	0	0	3	1	T	0	0	0	0	T	0	0	3	1	T	0	0	5	2
G	1	0	2	3	G	1	2	6	2	G	2	2	4	5	G	2	0	1	0
C	0	2	0	1	C	3	1	0	0	C	3	2	0	5	C	1	2	3	1

Lemur

A	A	T	G	C	G	A	T	G	C	T	A	T	G	C	C	A	T	G	C
A	1	1	2	0	A	0	0	1	0	A	2	1	5	1	A	1	1	1	1
T	0	0	4	0	T	0	1	3	0	T	1	1	4	1	T	0	2	3	3
G	3	2	3	0	G	3	3	5	3	G	2	2	3	4	G	1	0	0	0
C	0	1	0	1	C	2	2	0	0	C	2	3	1	1	C	0	2	0	0

Mouse

A	A	T	G	C	G	A	T	G	C	T	A	T	G	C	C	A	T	G	C
A	1	0	3	1	A	0	0	0	0	A	3	2	2	1	A	1	0	1	1
T	0	0	3	0	T	0	0	3	0	T	0	2	4	2	T	0	1	7	1
G	1	2	3	0	G	6	2	5	1	G	0	4	2	3	G	1	0	0	0
C	0	1	0	2	C	0	3	0	0	C	3	2	0	2	C	0	3	1	3

Rabbit

A	A	T	G	C	G	A	T	G	C	T	A	T	G	C	C	A	T	G	C
A	0	1	4	0	A	0	0	0	0	A	4	0	3	0	A	1	1	1	1
T	0	0	2	1	T	0	0	1	0	T	0	1	7	3	T	0	0	5	0
G	2	3	3	0	G	3	3	6	3	G	2	5	3	3	G	0	0	1	0
C	0	1	0	0	C	1	2	0	1	C	2	0	1	2	C	1	2	0	2

Rat

A	A	T	G	C	G	A	T	G	C	T	A	T	G	C	C	A	T	G	C
A	1	1	2	2	A	2	0	1	1	A	3	2	2	0	A	0	0	1	1
T	1	0	3	0	T	1	0	1	0	T	0	2	4	0	T	2	0	7	0
G	1	1	3	0	G	5	3	5	2	G	1	2	3	5	G	0	0	0	1
C	0	2	0	2	C	0	0	0	0	C	2	3	1	2	C	0	4	0	2

visible from visual inspection of Table 1: Certain triplets occur more often than others in the exons, like the triplet TGG, GTG, CTG, while some triplets are absent: ATA, GCG, CGA, CGC. Some triplets, like GTG and TGG, show small variations in the frequency of occurrence, while others show considerable variations, like CCT or TGA. The triplet CGC is the only triplet that did not occur in the whole DNA sequence (exons and introns included). From the corresponding 4 × 4 matrix of intron 446−1295 we see high frequency of occurrence of several triplets, the most pronounced being TTT.

In Table 4 we have listed the first exon of the β-globin gene for the eight species including the human β-globin gene already portrayed, and in Table 5 we collected the 4 × 4 matrices for the first exons of Table 4. Again a close look at the corresponding entries for the same triplet in the eight species shows visible parallelism for many triplets. In particular the occurrence of zeros in similar locations is noticeable.

Table 6: All 12 Matrices M_1−M_{12} for the First Exon of the Human β-Globin Gene

A_1					A_2					A_3					A_4				
	A	T	G	C		A	T	G	C		A	T	G	C		A	T	G	C
A	0	0	3	1	A	0	0	0	1	A	3	1	3	1	A	1	0	1	1
T	0	0	2	0	T	1	0	1	0	T	0	2	6	1	T	0	0	6	1
G	1	2	3	0	G	5	1	6	3	G	2	4	3	3	G	0	2	0	0
C	0	1	2	1	C	0	1	0	1	C	3	0	0	3	C	0	4	1	2
A_5					A_6					A_7					A_8				
	A	T	G	C		A	T	G	C		A	T	G	C		A	T	G	C
A	0	0	3	1	A	0	0	2	0	A	1	2	3	0	A	0	1	2	1
T	0	0	0	1	T	1	0	1	0	T	5	1	6	3	T	0	1	0	1
G	3	1	3	1	G	0	2	6	1	G	2	4	3	3	G	3	0	0	3
C	1	0	1	1	C	0	0	6	1	C	0	2	0	0	C	0	4	1	2
A_9					A_{10}					A_{11}					A_{12}				
	A	T	G	C		A	T	G	C		A	T	G	C		A	T	G	C
A	0	0	3	1	A	0	0	1	0	A	3	0	3	1	A	1	1	1	1
T	0	1	0	0	T	0	0	2	0	T	2	1	6	6	T	0	0	1	1
G	1	5	2	0	G	2	1	4	2	G	3	6	3	0	G	0	3	3	0
C	0	0	3	0	C	1	1	0	4	C	2	0	0	1	C	1	1	3	2

SEQUENCE VECTORS AS SOURCE OF INVARIANTS

Our prime interest is to extract a set of invariants from the information collected in Tables 3 and 5. Let us again consider the first exon of human β-globin gene and the four 4 × 4 condensed matrices listing the occurrence of all XYZ triplets.

In Table 6, the 12 matrices for the human β-globin gene are fully listed. Similar matrices can be constructed for the other seven β-globin genes of Table 4. To arrive at more compact information of the 12 matrices we consider the 12 leading eigenvalues of the twelve 4 × 4 matrices. These are listed in Table 7 for all eight species of Table 4. Each column in Table 7 can be viewed as a 12-component vector representing one of the DNA sequence of Table 4. That eigenvalues in each column, when looked in isolation, show considerable variation is not so surprising. The large eigenvalues indicate large average row/column sum for the 4 × 4 submatrix of XYZ triplets, and the small eigenvalues indicate small average row/column sum for the 4 × 4 submatrix of XYZ triplets. The average row/column sum varies from 4 to 9.25 in the case of the 12 XYZ matrices of Table 6, which is reflected in the variations of the leading eigenvalues from 3.22069 to 9.55589. In Figure 2 we show correlations between the leading eigenvalue and the average row/column matrix element for the eigenvalues of the 12 submatrices of Table 6. A quadratic correlation has the coefficient of regression $r = 0.9745$, the standard error $s = 0.502$, and the Fisher ration $F = 84.7$. This suggests that similar comparisons can be expected if DNA sequences are represented by the vector of the leading eigenvalue or the vector of average row/column values.

As we see from Table 7 there is also a visible variation in the magnitudes of the leading eigenvalues along each row, even though a large eigenvalue in one of the column seems to be accompanied with large eigenvalues in other columns and a small eigenvalue in one of the column seems to be accompanied with small eigenvalues in other columns. The

Table 7: Leading Eigenvalues of the 12 Matrices M_1-M_{12} for the Eight DNA Sequences of Table 4 as Components of a 12-D Vectors

	A	B	C	D	E	F	G	H
1	4.64897	5.58774	5.42702	4.41308	5.9587	5	5.6444	4.86215
2	6.27537	4.23607	5.98753	6	7.33048	6.69817	6.61505	6.6301
3	9.05045	8.47278	7	8.69813	8.8305	8	9.51808	8.24599
4	4.4764	3.64575	4.51882	5.40252	3.946	4.39787	2.61803	4.17513
5	5.34203	5.85028	6.27397	4.92646	6.27537	5.03404	5.772	5.18029
6	7.30117	4	5.9406	4.79129	6.39364	7.21699	8.81507	4.75632
7	9.55178	8.26794	7.58613	8.07405	8.65542	8.54848	10.49908	7.57073
8	4.42795	3.27732	4.59535	5.20261	3.07395	3.90547	3.65109	2
9	3.22069	3.45668	5.0101	4.64575	4.14743	3	4.20147	4.70825
10	5.47419	4.26982	6	3.41421	5.60462	5.8564	5.604	4.10548
11	9.65589	8.25349	7.39963	7.97123	10.23594	8.83772	10.1669	8.48982
12	4.69226	5.82843	5	7.07318	5.03791	5.40798	4.9348	5.64575

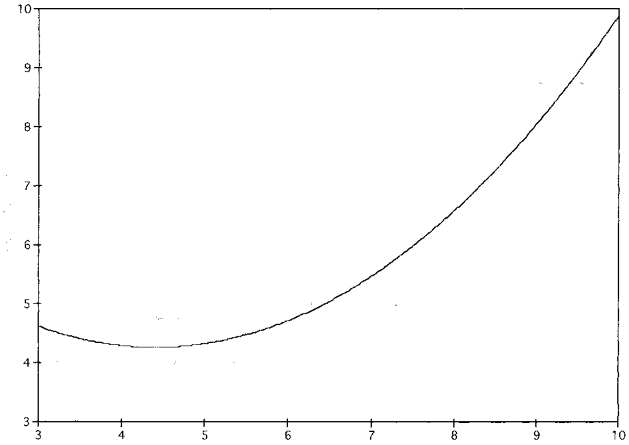

Figure 2. Correlation between the leading eigenvalue and the average row/column matrix element for the eigenvalues of the 12 submatrices of exon-1 of human β-globin gene.

Table 8: Magnitudes of the 64-Components Vectors for the Eight DNA Sequences of Table 4

species	magnitude	numerical	species	magnitude	numerical
human	$\sqrt{292}$	17.0880	lemur	$\sqrt{246}$	15.6844
goat	$\sqrt{272}$	16.4924	mouse	$\sqrt{284}$	16.8523
opossum	$\sqrt{289}$	16.9115	rabbit	$\sqrt{294}$	17.1464
gallus	$\sqrt{269}$	16.4012	rat	$\sqrt{280}$	16.7332

variations along the rows represent the variations in the XYZ triplet composition between different species.

SIMILARITIES AND DISSIMILARITIES AMONG XYZ TRIPLETS

Rather then elaborating on the point-by-point similarities of occurrence of the 64 XYZ triplets and differences in their frequencies, which in itself is of some interest, we will examine similarities and dissimilarities among the first exon of the eight species by constructing a vector having 64 components consisting of the frequency occurrence of all possible triplets. The triplets can be listed alphabetically or in any prescribed way. The underlying assumption is that if two vectors point to a similar direction in the 64-dimensional space and have similar magnitudes, then the two DNA sequences represented by the two vectors are similar. In Table 8 we have listed the magnitudes of the eight 64-components vectors, and as we see they are all of a similar magnitude. The similarity among such vectors can be measured in two ways: either we calculate the Euclidean distance between the end points of the vectors, or we can compute the scalar product (the inner or dot product) of vectors. The smaller is the Euclidean distance between the end point of two vectors the more similar are the corresponding DNA sequences. On the other hand the larger is the scalar product between two vectors (for vectors of a similar magnitude) the more similar are the corresponding DNA sequences.

Table 9: Similarity/Dissimilarity Table for the Eight DNA Sequences of Table 4 Based on Distances between the End Points of the 64-Component Vectors

		A	B	C	D	E	F	G	H
human	A	0	$\sqrt{80}$	$\sqrt{130}$	$\sqrt{113}$	$\sqrt{102}$	$\sqrt{51}$	$\sqrt{45}$	$\sqrt{68}$
goat	B		0	$\sqrt{158}$	$\sqrt{104}$	$\sqrt{88}$	$\sqrt{71}$	$\sqrt{104}$	$\sqrt{74}$
opossum	C			0	$\sqrt{184}$	$\sqrt{138}$	$\sqrt{121}$	$\sqrt{112}$	$\sqrt{174}$
gallus	D				0	$\sqrt{104}$	$\sqrt{129}$	$\sqrt{112}$	$\sqrt{122}$
lemur	E					0	$\sqrt{91}$	$\sqrt{90}$	$\sqrt{100}$
mouse	F						0	$\sqrt{77}$	$\sqrt{58}$
rabbit	G							0	$\sqrt{105}$
rat	H								0

Table 10: End Point Distances and the Scalar Product (Dot Product) of the 64-Component Vectors for the Eight DNA Sequences of Table 4

DNA pair	end point distance	scalar product	DNA pair	end point distance	scalar product
AB	8.94427	244	CE	11.74734	197
AC	11.40175	224	CF	11.00000	225
AD	10.63015	224	CG	10.58301	226
AE	10.09950	218	CH	13.19091	201
AF	7.14143	263	DE	10.19804	206
AG	6.70820	268	DF	11.35782	213
AH	8.24621	252	DG	10.58301	226
BC	12.56981	200	DH	11.04536	214
BD	10.19804	219	EF	9.53939	220
BE	9.38083	215	EG	9.48683	225
BF	8.42615	243	EH	10.00000	213
BG	10.19804	231	FG	8.77496	251
BH	8.60233	239	FH	7.61577	253
CD	13.56466	186	GH	10.24695	234

In Table 9 we give the similarity/dissimilarity table for the eight DNA sequences of Table 4 based on the Euclidean distance between the end points of the 64-component vectors. In Table 10 in the central column are listed the numerical values for the entries of Table 9, which vary between 6.71 and 13.57, corresponding to $\sqrt{45}$ and $\sqrt{184}$ for (A, G) and (C, D), respectively. Observe that in a few instances there is a degeneracy, that is, for two different pairs we find the same distance between their end points. This is the case with the pairs (B, D), (D, E); (C, G), (D, G). In the last column of Table 10 we give the scalar products between the 64-component vectors for all possible pairs of the eight DNA

Table 11: Similarity/Dissimilarity Table for the Eight DNA Sequences of Table 4 Based on the Quotient of the Corresponding Entries of Table 10

		A	B	C	D	E	F	G	H
human	A	0	0.0367	0.0509	0.0475	0.0463	0.0272	0.0250	0.0327
goat	B		0	0.0628	0.0466	0.0374	0.0347	0.0441	0.0360
opossum	C			0	0.0729	0.0596	0.0489	0.0468	0.0656
gallus	D				0	0.0495	0.0533	0.0468	0.0516
lemur	E					0	0.0434	0.0422	0.0469
mouse	F						0	0.0350	0.0301
rabbit	G							0	0.0438
rat	H								0

sequences of Table 4. Again we find degeneracy for the following pairs of species: (A, C), (A, D); (C, G), (D, G). In comparing the end point distances (Table 9 and the central columns of Table 10) with scalar products (the last column of Table 10) one should remember that the smaller entries for the end points the more similar the sequences are, while the larger the entries of the scalar product the more similar the sequences are. We can combine the information on the end points and the scalar product into a single similarity/dissimilarity table by considering the quotients of the corresponding entries. In Table 11 we collected such information. In this way one may reduce the degeneracy, since the coincidental entries in the two numerical columns of Table 10, in general, need not occur for the same pair of entries. Indeed, as we see from Table 11 there is now only one degenerate pair, (C, G), (D, G), which was simultaneously present in both columns of Table 10.

As we see from Table 9 opossum and gallus are the most dissimilar species among the eight considered because their corresponding rows have large entries. The small entries are found for human-mouse, human-rabbit, human-rat, goat-mouse, mouse-rat, and mouse-rabbit pairs. Similar results have been reported previously based on similarity when only the occurrence of the pairs of nucleic acid bases were considered. That mouse-rat and mouse-rabbit pairs show similarity is not surprising, but the apparent similarity of the DNA sequences of human-mouse, human-rabbit, human-rat, and goat-mouse pairs, as discussed in ref 6, is either an artifact reflecting deficiency of the sequence invariants, or indeed the sequences may have visible similarity even though the corresponding species are not closely related in the evolutionary sense. It is known that widely different species may have very similar DNA sequences, as was the case reported for Catalase gene of the yeast Candida tropicalis.[10]

There is some parallelism between the magnitudes of the similarity/dissimilarity entries of Table 10, which are evident in Figure 3 in which we plotted the scalar products of two 64-component vectors against the distance between their end points. A linear regression gives the following statistical parameters: the regression coefficient $r = 0.9000$, standard error $s = 8.94$, and the Fisher ratio $F = 110.9$. As we see from Figure 3 the two measures of similarity are somewhat different, but they both predict the same pairs of DNA sequences as the most similar and the least similar.

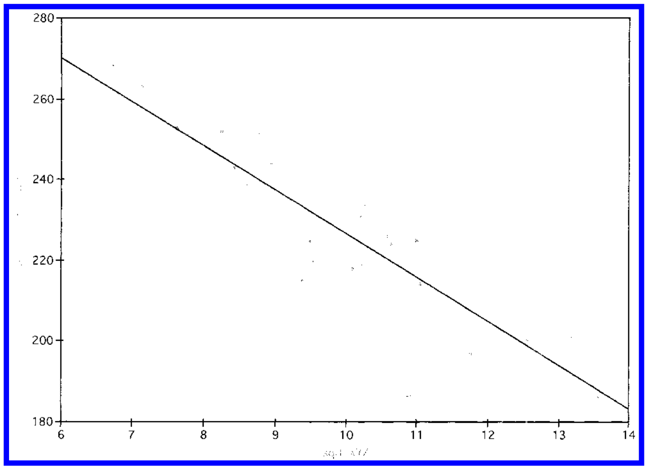

Figure 3. The plot of the scalar products of two 64-component vectors against the distance between their end points.

Table 12: Similarity/Dissimilarity Table for the Eight DNA Sequences of Table 4 Based on the 12-Component Vectors of the Leading Eigenvalues of M_1-M_{12} Matrices

		A	B	C	D	E	F	G	H
human	A	0	4.996	4.491	5.015	2.970	2.042	3.171	4.857
goat	B		0	4.358	4.078	4.780	4.683	6.551	3.378
opossum	C			0	4.459	4.287	3.545	4.126	5.466
gallus	D				0	4.723	7.064	3.959	2.934
lemur	E					0	3.566	3.779	4.045
mouse	F						0	4.118	6.213
rabbit	G							0	6.063
rat	H								0

SIMILARITY BASED ON LEADING EIGENVALUES OF XYZ MATRICES

Vectors made by using 12 leading eigenvalues of the 4 × 4 matrices as components (listed in Table 7) offer an alternative similarity analysis for all eight DNA sequences of Table 4. Using the Euclidean distances based on the components of the vectors of the leading eigenvalues we obtain the similarity/dissimilarity values shown in Table 12. Again we can observe large entries for gallus, the only nonmammalian species among those considered, and again we can observe small entries for the pair human-mouse. In contrast to other previously considered similarities somewhat unexpectedly we find a relatively large entry for mouse-rat. While this may appear disappointing one should recall that similarity/dissimilarity among species should not rest on the comparison of a single descriptor or on a single gene. The "disappointing" result of a similarity/dissimilarity based on the leading eigenvalue vectors only indicates, and it indicates very clearly, that different descriptors and different invariants encode *different* structural properties of a DNA sequence. In that respect the finding is beneficial, as it ensures us that with distinct approaches we are summarizing different features of DNA sequence.

SEQUENCE INVARIANTS

We have outlined construction of 64-component vectors and a 12-component vector to represent DNA sequences, but

Table 13: "Power" Matrices for Powers 1−6 for the First of the 12 Matrices for Human β-Globin Gene

$^1A_1$					$^2A_1$					$^3A_1$				
	A	T	G	C		A	T	G	C		A	T	G	C
A	0	0	3	1	A	0	0	9	1	A	0	0	27	1
T	0	0	2	0	T	0	0	4	0	T	0	0	8	0
G	1	2	3	0	G	1	4	9	0	G	1	8	27	0
C	0	1	2	1	C	0	1	4	1	C	0	1	8	1
$^4A_1$					$^5A_1$					$^6A_1$				
	A	T	G	C		A	T	G	C		A	T	G	C
A	0	0	81	1	A	0	0	243	0	A	0	0	729	1
T	0	0	16	0	T	0	0	32	0	T	0	0	64	0
G	1	16	81	0	G	1	32	243	0	G	1	64	729	0
C	0	1	16	1	C	0	1	16	1	C	0	1	32	1

we have come across but a single sequence invariant, the magnitude of the 64-component vectors (Table 8). The components of vectors do not represent invariants, but the magnitude of a vector is invariant, as its value does not depend on ordering of the components. We can similarly derive magnitudes of the 12-component vector, both when leading eigenvalues are used or the average row/columns are used. For example, for the human β-globin gene they are 22.61023 and 20.60643, respectively. Are there other sequence invariants that can be formulated for sequence comparisons?

The difficulty of the task of finding or constructing novel invariants for DNA sequences based on properties of XYZ triplets lies in the fact that we have to condense large amounts of information into a single number. The task was somewhat less difficult when we considered only pairs of nucleic acid bases, there being 16 such partial data, but now we have already 64 partial data to combine into a single invariant. One way of arriving at a set of invariants is to follow the "recipe" of construction of molecular profiles[11] that we have already exploited when considering DNA sequence "profiles" based on frequencies of XY pairs of nucleic bases. We will briefly outline the approach for human β-globin gene that is represented by the twelve 4 × 4 matrices of Table 6.

Consider the first submatrix A_1 and matrices derived from it by raising the individual matrix entries to higher powers. In Table 13 we show this for the first half a dozen powers of A_1 which are indicated as matrices $^2A_1$, $^3A_1$, $^4A_1$, etc. The leading eigenvalue of each matrix $^mA_1$ has been calculated and is shown in Table 14 for m = 1 to m = 20. As we see the leading eigenvalues increase fast so we introduced normalizing factor $1/m!$ in order to obtain convergent sequence of leading eigenvalues as m increases. These normalized leading eigenvalues define the "profile" for object represented by matrix A_1. In case of small molecules in this way the so-called "molecular profiles",[11] which represent an ordered set of molecular invariants, have been obtained in this manner. In the case of representation of DNA by "cubic matrices" of triplets XYZ, we have 12 matrices for each DNA sequence, and thus the "DNA profile" has 12 components, only one of which has been shown in Table 14 and illustrated in Figures 4 and 5 in two different fashions, as an (x, y) curve and by vertical bars, respectively. It is this last pictorial representation of the leading eigenvalue sequence that has led to the label "profile" for the set of invariants so constructed. The distinction between 2-D profiles of molecules and DNA profiles outlined here is that the latter represents a component of a 12-dimensional (12-D) profile that accompanies each DNA sequence. It would be of interest to see how 12-D profiles characterize individual DNA and whether they lead to similarity/dissimilarity studies that show parallelism with simple characterization of DNA based on limited number of invariants.

Table 14: First Component of a 12-Component DNA Profile for Exon-1 of the Human β-Globin Gene

power	leading eigenvalue	normalized leading eigenvalue
1	1.186383 e+2	118.6383
2	3.577666 e+3	894.4165
3	1.087522 e+5	3020.8950
4	3.320808 e+6	5765.2917
5	1.017051 e+8	7062.8533
6	3.121852 e+9	6022.0905
7	9.600520 e+10	3779.4943
8	2.957423 e+12	1819.1663
9	9.124986 e +13	692.9568
10	2.819911 e+15	214.1457
11	8.728054 e+16	54.777979
12	2.705684 e+18	11.792425
13	8.400690 e+19	2.166478
14	2.612354 e+21	0.343728
15	8.136323 e+22	0.047580
16	2.538064 e+24	0.005798
17	7.929629 e+25	0.000627
18	2.481275 e+27	0.000060
19	7.776116 e+28	0.000005
20	2.440670 e+30	0.000000

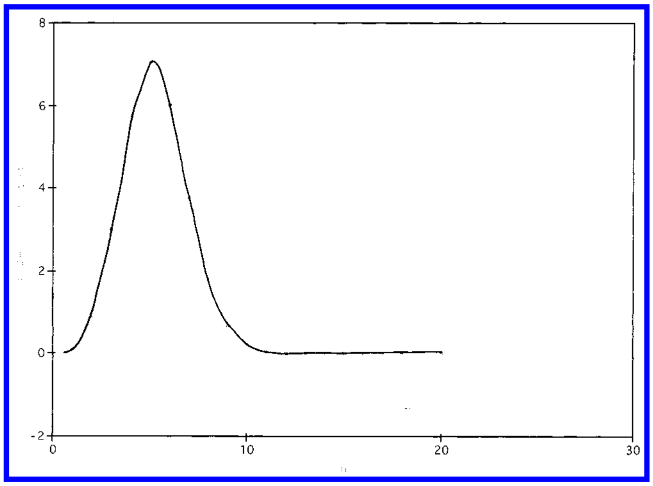

Figure 4. A single component of "DNA profile" (that has 12 components), illustrated as an (x, y) curve.

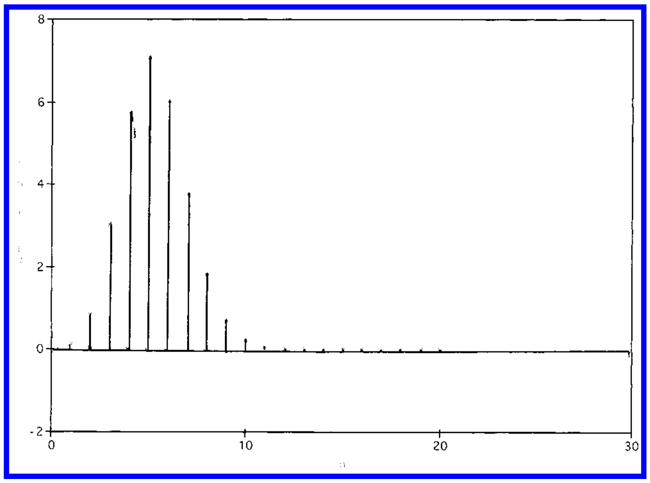

Figure 5. A single component of "DNA profile" (that has 12 components), illustrated by the vertical bars.

CONCLUDING REMARKS

A direct comparison of DNA sequences (even those of a restricted length) is, to say the least, computer intensive. Here we illustrate an alternative approach which is based on comparison of sequence invariants, rather than sequences themselves. In this particular contribution we focused attention to XYZ triplets and the frequencies of their occurrence in DNA primary sequences. We have shown how information in a "cubic" 4 × 4 × 4 supermatrix of all 64 triplets can be used to extract traditional matrix invariants, like the leading eigenvalue of 12 component matrices M_1-M_{12} (Table 2). The same component matrices can also be used to build matrices of the "higher" order (by raising matrix elements to higher powers), which offer a more complete characterization of individual sequences.

Future studies and future applications may lead to additional invariants and may show which set of invariants is more suitable for which kind of applications. Novelty in this kind of approach, in comparison with the traditional direct use of DNA sequences, is that here we have arrived at the possibility of generating numerous numerical attributes for individual DNA sequences. This contribution perhaps raises more questions than offers answers, but that only signifies that the approach may have promise. Here are a few of the questions that ought to be considered.

(1) We mentioned two kind of matrices: (i) matrices in which an individual entry corresponds to an individual pair of bases and (ii) matrices in which entries summarize information of different X−Y pairs of bases (referred to as condensed matrices). Is there any order of preference or obvious choice in characterization of DNA sequence in terms of mono-, di-, and trinucleotides?

(2) Should the characterization on triplets of DNA, particularly in the case of exons associated with protein production, be based on nonoverlapping triplets, rather than on sequential triplets?

(3) We observed similarities based on the leading eigenvalue of the matrices M_1-M_{12} (the individual rows of Table 7). Is this a characteristic typical of all conserved genes where homologies are strong and triplet frequencies would be expected to be fairly well conserved? Could this similarity between the rows be used to identify conservative genes?

(4) How sensitive is the "DNA profile" on permutation, deletion, or insertion of a single or a pair of nucleic acid bases?

We will continue to develop the characterization of DNA by invariants and hope to resolve some of the pending questions soon. We would also welcome comments, inputs, modifications, etc. of other interested scientists which may facilitate further refinements of the outlined methodology for numerical characterizations of DNA.

ACKNOWLEDGMENT

We thank Professor A. T. Balaban and Dr. A. Nandy for critical review of this manuscript and their many comment that improved the manuscript. This is contribution number 282 from the Center for Water and the Environment of the Natural Resources Research Institute Research reported in this paper was supported in part by Grants F49620-98-1-0015 and F49620-01-1-0098 from the United States Airforce.

REFERENCES AND NOTES

(1) *Time Wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison;* Sankofff, D., Kruskal, J. B., Eds.; Addison-Wesley Publ. Co.: Reading, MA, 1983.

(2) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85,* 2444−2448.

(3) Kruskal, J. B. An overview of sequence comparison. In *Time Wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison;* Sankofff, D., Kruskal, J. B., Eds.; Addison-Wesley Publ. Co.: Reading, MA, 1983; pp 1−44; p 5.

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

(5) Randić, M.; Vracko, 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.

(6) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 50−56.

(7) Randić, M. On characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34

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

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

(10) Okada, H.; Ueda, M.; Sygaya, T.; Atomi, H.; Mozaffar, S.; Hishida, T.; Teranishi, Y.; Okazaki, K.; Takechi, T.; Kamiryo, T, et. Catalase gene of the yeast Candida tropicalis. Sequence analysis and comparison with peroxisomal and cytosolic catalases from other sources. *Eur. J. Biochem.* **1987**, *170,* 105−110.

(11) 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.

CI000120Q