Condensed Representation of DNA Primary Sequences

Milan Randić*

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

Received July 26, 1999

With rapid reporting of DNA sequences derived from automated DNA sequencing techniques, the problem of reviewing and ordering such information has become acute. We have introduced a condensed representation of primary sequences of DNA that offers an alternative method of registering DNA. The advantage of the condensed codes for DNA is that it not only offers fast, qualitative comparisons of DNA but also allows quantitative comparisons of DNA from different sources. The approach is outlined for a particular human β globin sequence extract. Using the condensed representation of the primary DNA sequences, comparisons are made between primary sequences for Exon 1 of human β globin and seven other β globins.

INTRODUCTION

DNA primary sequencing, even when considered for restricted relatively short segments such as the segment running from 62 205 to 63 628 (of length 1424) shown in Table 1, does not yield an immediately useful or informative characterization of the object it represents. While detailed comparisons between such different primary sequences are possible, the procedure is convoluted and involves numerous, not yet fully resolved issues, such as how to compare sequences of different length. Determining frequencies of symbols, a time-consuming task, the study of distribution of distances in strings of symbols, and the study of similarity of strings, have received some attention.[1−5] The methods of sequence comparisons are used in molecular biology to solve questions on the homology of macromolecules.[6] The degree of homology is high if the differences between the strings representing two molecules are small, pointing to molecules that have a common ancestor. Such studies included statistical analysis of biological DNA sequences, which were decomposed as text into syllables, words, and group of words, or restricted to consideration of 3-tuples, known as codons, a sequence of three nucleotides of messenger RNA that code for an amino acid or for chain termination.

Sequence comparison involves the search for the optimum correspondence between the sequences and the use of a distance function for measuring similarity or dissimilarity when the exact correspondence between string elements is not known. Differences between sequences can arise due to substitution or transposition. A search for the optimum correspondence between sequences considers permutations and alignments (or matching).[7] All this could be avoided if we could represent sequences by matrices. Matrices can be compared readily by using a set of matrix invariants as a basis for their characterization. The first problem, however, is how to assign to a sequence a numerical matrix, the problem that has only recently been considered.[8]

Recently, several schemes for graphical visualization of the primary DNA sequence were proposed.[9−11] Such schemes have an advantage in that they offer an instant, though visual and qualitative, impression or a summary of the lengthy primary DNA sequences. In Figure 1 we show a graphical representation of the first 92 bases of the β globin gene as obtained from the approach outlined by Nandy.[12−14] It is possible, as has been recently demonstrated,[8] to transform such a geometric representation into an algebraic numerical form, which then allows quantitative analyses of data on primary sequences of DNA.

In this contribution we have taken a different approach. Instead of using a geometric representation of the primary DNA sequence as a basis for construction of a matrix associated with the DNA sequence, we will use the DNA sequence directly. Graphical approaches involve to a greater or lesser degree arbitrary conventions when the assignment of the direction in the x, y plane, or x, y, z space, are selected for the four nucleic bases, A, G, C, and T. Here A, and G are the codes for purine bases adenine and guanine and C, and T are the codes for pyrimidine bases cytosine and thymine.

CONDENSED REPRESENTATION OF DNA

Our nongraphical approach uses as the input information on DNA its primary sequence only. Hence, the approach is devoid of the arbitrary intermediary steps used in construction of a graphical representation of DNA. We will illustrate the approach by considering a portion of the DNA sequence of the human β globin of Table 1 of length 92. The initial portion is printed in bold letters and it represents the first exon of human β globin shown. Its graphical representation has already been illustrated in Figure 1. An exon is a polypeptide encoding portion of structural genes that give rise to amino acid sequences in the protein and are separated from each other by an intervening sequence (intron). The DNA of Table 1 has two more exons, also shown in bold letters, separated by long portions of introns (DNA sequences within a structural gene that are transcribed into RNA but do not give rise to amino acid sequences in the protein). Hence, if one is to fragment a long DNA sequence, it makes sense to separately consider each exon. The β globin

* On leave from the Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311.

10.1021/ci990084z CCC: $19.00
Published on Web 11/19/1999

Table 1. DNA Primary Sequence for Human β-Globin (Segment from 62205 to 63628)[a]

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAC

GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTT

AAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGG

CACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAG**GCTGCTGGTGGTCTACCC**

TTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG

CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC

TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCT

GCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACCCTTGATGTTTTCTTTCCCC

TTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGAGAAGTAACAGGGTACAGTTTAG

AATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTT

TTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCTTTTTTTTTCT

TCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA

TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTAC

TATTTGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTT

ATTTTTAATTGATACATAATCATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATA

TGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAAAATGCTTT

CTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTT

TCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTG

ATAATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTGCATATAAAT

TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTG

CTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGC

TAATCATGTTCATACCTCTTATCTTCCTCCCACAG**CTCCTGGGCAACGTGCTGGTCTGTG**

TGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAG

TGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

[a] The three exons portions are shown in boldface.

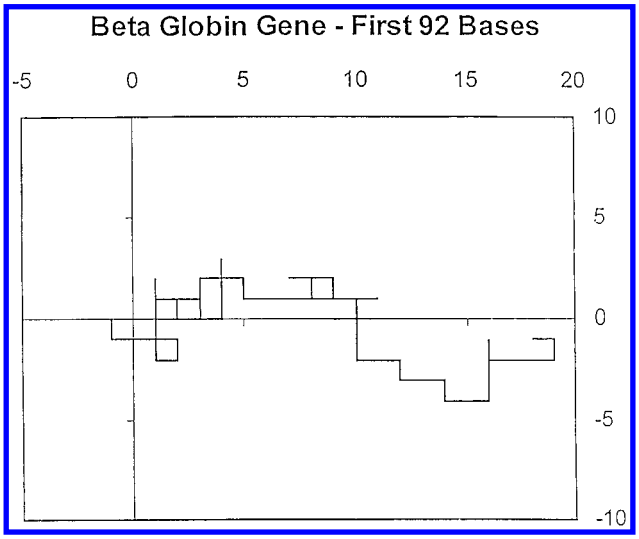

Figure 1. Graphical representation of β globin gene—first 92 bases, according to the scheme of Nandy.

sequence extract of Table 1 has three exons (1−92, 223−445, 1296−1424) and two introns (93−222 and 446−1295).

Instead of analyzing the primary sequence of the first exon by constructing a (92 × 92) matrix that one can associate with this segment of DNA in which the rows and columns are assigned to *individual* nucleic bases of the same kind, we will consider a reduced (4 × 4) matrix in which the rows and columns are assigned to *totality* of nucleic acids of the same kind

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

The ordering of the nucleic bases in such a matrix is not important. We have taken the order of occurrence of these bases in the primary sequence of the exon 1 considered. Each entry in this table is associated with a pair of labels. Thus in the first row we have entries AA, AT, AG, and AC. We will assign to a pair an XY numerical value that gives the frequency of occurrence of XY as adjacent entries in the primary sequence of DNA, which distinguishes it from YX, which is counted separately. Hence, the condensed (4 × 4) matrix will generally be asymmetric. For exon 1 of Table 1 we obtain AA = 4, AT = 2, AG = 7, and AC = 4, which are the entries of the first row of the condensed matrix for

Table 2. Condensed Matrix for Exon 1−92

	A	T	G	C	Row sum
A	4	2	7	4	17
T	1	2	15	2	20
G	8	9	12	6	35
C	3	7	2	6	18
Column sum	16	20	36	18	90

Table 3. Condensed Matrices for the Remaining Exons and Introns of the Human β-Globin

Exon 223-445

	A	T	G	C
A	10	6	14	4
T	4	12	28	12
G	15	12	20	15
C	14	25	2	15

Exon 1296-1424

	A	G	T	C
A	9	5	6	6
G	4	3	17	6
T	2	10	11	12
C	12	12	1	12

Intron 93-222

	A	T	G	C
A	7	6	12	9
T	8	12	9	8
G	11	7	12	3
C	8	11	0	6

Intron 446-1295

	A	T	G	C
A	73	96	36	33
T	90	148	55	53
G	27	42	31	26
C	48	62	3	24

exon 1 of Table 1. By completing the count of successive adjacent nucleotide bases for exon 1 we obtain the 4 × 4 condensed matrix shown in Table 2.

Clearly there is a loss of information when one condenses the primary sequence of DNA to a (4 × 4) matrix. Nevertheless, as we will see later, different sequences are represented by different matrices. The advantage of such a reduced representation of the primary DNA sequence is that it offers upon inspection useful information that is hidden in the lengthy sequence of the primary DNA. The row sums and the column sums are simply related to the frequency number of the individual nucleic acids, while the sum of all the entries gives the length of the exon, decreased by 2 because the starting and the ending base are counted without the preceding or the following base. For the same reason there is a difference between the row sums and the column sums for A (the initial base) and G (the ending base).

Using the above outlined representation of primary DNA sequences, the β globin sequence extract shown in Table 1 receives an alternative representation shown Table 2 and Table 3 given by five condensed ATGC matrices. Three of the five matrices characterize the three exons and the two last matrices characterize the two introns. The different characters of different exons, and differences between exons and introns, are immediately visible from the very distinct forms of the corresponding condensed (4 × 4) matrices. On the other hand we can also immediately see some common features between different condensed matrices. For example, all three exon matrices show that TG is the most frequently occurring pair of adjacent bases while CG is one of the least frequently occurring base pairs, which is well-known. Intron 446−1295 is characterized by a high frequency of several diagonal entries, in particular the TT pair.

COMPARATIVE STUDY OF EXON 1 OF DIFFERENT SPECIES

In Table 4 we have reproduced exon 1 of β-globin for several different species, including goat, opossum, gallus, mouse, rabbit, and rat, all having between 86 and 93 bases. Table 5 gives the corresponding (4 × 4) condensed DNA matrices for adjacent pairs of nucleic bases. We have also included human β gene matrix for convenience of comparison. β-Globin sequences of Table 4 represent the conservative gene, that is, the gene that changes little from one species to another. Indeed that this is the case can be seen by comparing different (4 × 4) condensed matrices of Table 5.

A visual inspection of Table 5 reveals some common features among the eight cases: in all cases CG is a rather infrequent pair while TG and GG appear quite frequently. Other infrequent pairs in all eight cases appear to be TA and TT. Besides such qualitative observation one can derive quantitative conclusions by considering the similarity/dissimilarity. We can view the matrix entries of the condensed (4 × 4) matrices as components of a 16-component vector: {AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC}. The order in which the components are listed is not important, but the same order should be kept for all eight cases. By using the Euclidean distance as a measure of similarity/dissimilarity we obtained the similarity/dissimilarity between the eight species in Table 6.

Because the condensed matrix is associated with loss of information, all that we can conclude from the similarity/dissimilarity matrix is that DNA sequences that are similar will necessarily have a relatively small number as the entry corresponding to such cases. We expect that exon 1 of mouse and exon 1 of rat will be similar. From Table 6 we see that the corresponding entry in the matrix is relatively small, being 42. The converse, that the small entry signifies similar species, however, need not be true. We can see that this is the case with entries (A, F), (A, G), and (B, E) in Table 2 which correspond to species which are not close in the evolutionary tree. Hence, without further testing we should not conclude that the corresponding DNA sequences are necessarily similar.

While a small entry in the similarity/dissimilarity matrix may indicate species that are similar, a large entry in such a table certainly points to species that are dissimilar. As we can see from Table 6, almost all entries belonging to gallus are large. Hence we may conclude that gallus has little similarity with the remaining species of Table 6, and indeed, it is not a mammal, while other species in Table 6 are mammals.

RECOVERY OF "LOST" INFORMATION

A way to recover some of the lost information associated with the condensation of the DNA sequence to a single (4

Table 4. First Exon (of Length between 86 and 93) for a Selection of Species

A	human	92 bases
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAC		
GTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG		
B	goat	86 bases
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGAT		
GAAGTTGGTGCTGAGGCCCTGGGCAG		
C	opossum	92 bases
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAG		
GTTGACCAGACTGGTGGTGAGGCCCTTGGC AG		
D	gallus	92 bases
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAAT		
GTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG		
E	lemur	92 bases
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGAT		
GTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG		
F	mouse	93 bases
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAA		
CCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG		
G	rabbit	90 bases
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAAT		
GTGGAAGAAGTTGGTGGTGAGGCCCTGGGC		
H	rat	92 bases
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAAC		
CCTGATAATGTTGGCGCTGAGGCCCTGGGCAG		

$\times$ 4) matrix is to introduce additional (4 $\times$ 4) matrices in which information on the number of pair of bases at greater distance is stored. For example, the beginning of exon 1 (shown in Table 1 in bold letters) start with ATGGTGCACC. Besides considering the count of the adjacent pairs: AT, TG, GG, GT, TG, GC, CA, AC, CC, which were used to form the (4 $\times$ 4) matrix of Table 2, we could count nonadjacent pairs. For example, we may consider pairs of bases separated by *one* base. Then for the beginning portion of exon 1 we obtain: AG, TG, GT, GG, TC, GA, CC, AC. If we count pairs separated by two bases for the same initial segment of exon 1 we obtain: AG, TT, GG, GC, TA, GC, CC. In Table 7 we collected for the first exon of the human β globin of Table 1 the frequency tables counting pairs of bases at greater separation (for distances from $d = 2$ to $d = 6$). They are analogous to Table 2, which gives frequencies for adjacent pairs (i.e., pairs at distance 1), which is included in Table 7 for easier comparison. Information on the frequency at increasing distances corresponds to information on longer range interactions of physical models.

We can use the additional (4 $\times$ 4) matrices and extend the similarity/dissimilarity testing between exons belonging to different species. In Table 8 and Table 9 we show the (4 $\times$ 4) condensed DNA matrices for nonadjacent pairs of nucleic bases. Table 8 gives frequencies for pairs separated by single nucleic base and Table 9 for pairs separated by two nucleic bases, corresponding to distances $d = 2$ and $d = 3$ respectively, between the nucleic bases considered, the adjacent bases being at distance of 1 link ($d = 1$).

Table 10 shows the similarity/dissimilarity tables obtained by considering vectors obtained from (4 $\times$ 4) matrices belonging to cases $d = 2$ and $d = 3$. These tables show similar features as the already discussed similarity/dissimilarity table derived for the case of $d = 1$, except that the occurrence of small entries within the table is in some cases redistributed. Again we see that mouse and rat exons are quite similar, being associated with the small entries in cases $d = 2$ and $d = 3$. Table 10 reflects the already stated observation that among the cases considered the primary sequences of mouse and rat are the most similar. On the

Table 5. Condensed Matrices for the First Exon of β-Globin of Table 4, Giving the Frequency for Adjacent Pairs of Bases ($d = 1$)

A

	A	T	G	C
A	4	2	7	4
T	2	3	14	2
G	7	9	12	6
C	3	7	2	7

B

	A	T	G	C
A	5	2	8	2
T	0	2	13	2
G	8	5	12	9
C	3	8	2	4

C

	A	T	G	C
A	3	3	8	7
T	2	4	12	4
G	8	6	9	5
C	7	9	0	4

D

	A	T	G	C
A	5	4	7	3
T	0	0	11	4
G	6	4	13	10
C	7	7	3	7

E

	A	T	G	C
A	4	4	9	2
T	1	4	14	4
G	9	7	11	7
C	4	8	1	2

F

	A	T	G	C
A	5	3	6	3
T	0	3	17	3
G	8	8	10	7
C	3	9	1	7

G

	A	T	G	C
A	5	3	8	1
T	0	2	15	4
G	7	11	13	6
C	4	5	1	5

H

	A	T	G	C
A	6	4	6	4
T	4	2	15	0
G	7	6	11	8
C	2	9	1	6

Table 6. Similarity/Dissimilarity Table for the Eight Exons of Table 5 for Adjacent Bases ($d = 1$)

	A	B	C	D	E	F	G	H	
A	0	48	74	92	56	30	37	40	human
B		0	90	52	32	55	69	48	goat
C			0	114	58	90	127	100	opossum
D				0	102	110	113	102	gallus
E					0	57	53	74	lemur
F						0	55	42	mouse
G							0	101	rabbit
H								0	rat

other hand, the accidental similarity of (B, E) when $d = 1$ is now shown to be false, because this entry in case $d = 2$ is quite large, which definitely points to dissimilarity. Several other smaller entries showing similarity based on $d = 1$ have somewhat increased, thus weakening arguments for considering them as belonging to closely related species. However, more conclusive arguments, when no prior information on species is available, should be based on the use of still more information, either using (4×4) matrices giving frequencies at still greater separations, using other sources and alternative comparability studies based on different considerations, or both.

Table 7. Condensed Matrices for the First Exon of Human β-Globin, Giving the Frequency for Nonadjacent Pairs of Bases[a]

d=1

	A	T	G	C
A	4	2	7	4
T	1	2	15	2
G	8	9	12	6
C	3	7	2	6

d=2

	A	T	G	C
A	1	5	8	2
T	8	2	7	5
G	10	7	9	8
C	1	6	8	4

d=3

	A	T	G	C
A	3	6	4	3
T	5	4	8	4
G	8	4	14	8
C	2	5	7	4

d=4

	A	T	G	C
A	1	5	6	4
T	2	3	11	15
G	13	6	7	7
C	3	4	8	3

d=5

	A	T	G	C
A	3	3	8	2
T	6	3	5	7
G	8	8	10	6
C	2	3	9	4

d=6

	A	T	G	C
A	4	3	4	5
T	3	5	9	4
G	5	6	13	7
C	4	5	6	3

[a] The case $d = 1$ corresponds to adjacent bases.

It is interesting to observe from Table 10, that gallus (species D) also shows great dissimilarity with other species when similarity is based on frequencies other than those of adjacent pairs of bases. The persisting uniqueness of its primary DNA sequence (among the species considered) clearly suggests that clustering of DNA information would separate this species from the rest at an earlier stage of evolutionary development. The other species that show several large entries is opossum, the only mammal among those considered belonging to distinctive superorder *marsupialia*, hence, again somewhat removed from others in the evolutionary tree.

CONCLUDING REMARKS

In this report we introduced a novel representation for DNA using condensed matrices that count the frequency of occurrence of adjacent pairs of bases. Such matrices offer some insight into the nature of the primary sequence of DNA. They allow one, when making comparisons, not only to recognize qualitative differences between sequences of DNA, whether within the same or between different species, but also to characterize them quantitatively. The loss of information that accompanies the condensation of the DNA sequence into a (4×4) matrix can be partially recovered by

Table 8. Condensed (4 × 4) Matrices for the First Exon of β-Globin of Table 4, Giving the Frequency for Nonadjacent Pairs of Bases at Distance $d = 2$

A

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

B

	A	T	G	C
A	2	2	10	1
T	6	1	5	5
G	7	9	9	10
C	1	3	11	12

C

	A	T	G	C
A	3	5	7	5
T	6	3	8	5
G	7	7	7	7
C	4	6	7	3

D

	A	T	G	C
A	1	3	8	6
T	4	3	6	2
G	8	4	9	12
C	5	3	11	5

E

	A	T	G	C
A	4	4	9	1
T	5	6	9	3
G	7	7	13	7
C	2	5	4	4

F

	A	T	G	C
A	2	3	9	3
T	6	5	8	6
G	6	10	8	8
C	2	4	9	5

G

	A	T	G	C
A	2	3	9	3
T	4	5	7	4
G	8	6	14	8
C	2	3	7	3

H

	A	T	G	C
A	3	4	8	4
T	8	3	6	3
G	7	8	10	7
C	2	4	8	4

Table 9. Condensed (4 × 4) Matrices for the First Exon of β-Globin of Table 4, Giving the Frequency for Nonadjacent Pairs of Bases at Distance $d = 3$

A

	A	T	G	C
A	3	6	4	3
T	4	4	9	4
G	8	3	14	8
C	2	5	7	4

B

	A	T	G	C
A	3	4	6	3
T	5	3	6	3
G	7	5	16	5
C	3	2	5	6

C

	A	T	G	C
A	8	6	4	2
T	6	5	6	5
G	3	5	12	8
C	3	5	6	5

D

	A	T	G	C
A	4	4	7	3
T	3	3	6	3
G	5	2	13	11
C	6	3	13	6

E

	A	T	G	C
A	4	7	5	2
T	3	5	9	6
G	6	5	16	7
C	3	5	4	2

F

	A	T	G	C
A	2	5	4	5
T	3	7	7	6
G	5	5	15	6
C	5	5	7	3

G

	A	T	G	C
A	1	5	4	5
T	5	3	10	3
G	8	3	16	8
C	0	7	5	3

H

	A	T	G	C
A	2	4	8	5
T	2	8	6	3
G	8	2	13	8
C	6	5	4	2

considering additional (4 × 4) matrices which count pairs of bases at increasing separation.

It should be added that the idea of condensed matrices that reduce data of a sequence, or a structure, to matrices of much smaller size than the size of the initial system has been outlined before.[15,16] Thus in the case of fullerenes the rows and the columns of reduced matrices were associated with twelve pentagons that are present in each fullerene. In this way, for instance, the (60 × 60) distance matrix of fullerene is reduced to a (12 × 12) matrix distance matrix for pentagons. In another report (4 × 4) matrices of DNA analogous to those considered here were outlined. However, instead of using the count of the occurrence of adjacent bases, the next adjacent bases, etc., for construction of the matrix elements the average matrix elements of different XY submatrixes of the initial ($n \times n$) matrix were used to obtain the corresponding matrix elements of the reduced matrix.[16] Here X and Y stand for the four nucleic bases A, G, C, and T.

The concept of condensed matrices can be extended to proteins.[17] In this case, instead of the count of the occurrence of adjacent bases, next adjacent bases, etc., one counts the occurrence of pairs of adjacent amino acids, next adjacent amino acids, etc. Instead of (4 × 4) matrices now we will have (20 × 20) matrices, the elements of which will be associated with various combinations of major pairs of amino acids.

What can the usefulness and reliability of the present approach be in comparison to analyses of the frequencies of single nucleic bases and triplets? Comparison of the frequencies of single nucleic bases is what lead Crick and Watson to establish the pairing of nucleic bases: adenine and cytosine (A−C), and guanine and thymine (G−T). Triplets, the codons, contain all the information necessary to initiate polypeptide synthesis, designate the specific sequences of amino acids, and terminate the polypeptide synthesis. Often, though not always, the first two nucleic bases fully specify the protein to be synthesized. Hence the use of pairs of adjacent bases offers some indication of the protein production, though not complete information. If one is interested in the role of the *very short* segments of DNA, frequencies for the pairs of nonadjacent bases are even less useful, particularly for the pairs of bases at greater separations. However, the use of the frequencies for the pairs of bases at different distances offers a tool for comparative study of *long* DNA segments. That indeed this appears a promising application we have seen from the fact that among the eight species considered we found gallus, and to lesser degree opossum, to be the most different and the least similar to the remaining species when comparing exon 1 of a DNA sequence of β globin.

Table 10. Similarity/Dissimilarity Table between the Eight Exons of Table 5 for Characterization Using Pairs of Nonadjacent Bases at Distance $d = 2$ (Top) and $d = 3$ (Bottom)

	A	B	C	D	E	F	G	H	
A	0	61	55	113	61	51	47	28	human
B		0	100	104	142	54	88	57	goat
C			0	96	88	46	88	33	opossum
D				0	170	110	86	89	gallus
E					0	84	34	65	lemur
F						0	74	40	mouse
G							0	51	rabbit
H								0	rat

	A	B	C	D	E	F	G	H	
A	0	56	77	114	37	52	30	81	human
B		0	88	138	69	72	86	106	goat
C			0	133	76	87	147	151	opossum
D				0	181	132	196	159	gallus
E					0	43	65	91	lemur
F						0	90	66	mouse
G							0	113	rabbit
H								0	rat

ACKNOWLEDGMENT

This work was supported by Grant J1-8901 of the Ministry of Science and Technology of Slovenia. I would also like to thank E. S. Wilks (Wilmington, DE), M. Kunz (Brno), M. Vracko (Ljubljana), and A. Nandy (Calcutta) for useful comments that improved the manuscript considerably.

REFERENCES AND NOTES

(1) Yule, G. U. *The Statistical Study of Literary Vocabulary*; Cambridge University Press: Cambridge, U.K., 1944.
(2) Kunz, M.; Radl, Z. Distribution of Distances in Information Strings. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 374−378.
(3) Harary, F.; Paper, H. H. Toward a General Calculus of Phonemic Distribution. *Language* **1957**, *33*, 143−169.
(4) Schmitt, A. O.; Ebeling, W.; Herzel, H. The Modular Structure of Informational Sequences. *Biosystems* **1996**, *37*, 199−210.
(5) Jerman-Blažič, B.; Fabič, I.; and Randić, M. Comparison of Sequences as a Method for Evaluation of Molecular Similarity. *J. Comput. Chem.* **1986**, *7*, 176−188.
(6) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(7) Kruskal, J. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Rev.* **1983**, *25*, 201−237.
(8) Randić, M.; Nandy, A.; Basak, S. C. On Numerical Characterization of DNA Primary Sequences. *J. Math. Chem.* (submitted).
(9) Roy, A.; Raychaudhury, C.; Nandy, A. Novel Techniques of Graphical Representation and Analysis of DNA Sequence−A Review. *J. Biosci.* **1998**, *23*, 55.
(10) Ninio, J.; Mizraji, E. Perceptible Features in Graphical Representations of Nucleic Acid Sequences. In *Visualizing Biological Information*; Pickover, C. A. Ed.; Word Scientific: Singapore, 1995; pp 33−42.
(11) Huen, Y. K. Representation of Biological Sequences Using Point Geometry Analysis. In *Visualizing Biological Information*; Pickover, C. A. Ed.; Word Scientific: Singapore, 1995; pp 165−182.
(12) Nandy, A. A New Graphical Representation and Analysis of DNA Sequence Structure. I. Methodology and Application to Globin Genes. *Curr. Sci.* **1994**, *66*, 309−313.
(13) Nandy, A. Graphical Analysis of DNA Sequence Structure: III. Indications of Evolutionary Distinctions and Characteristics of Introns and Exons. *Curr. Sci.* **1996**, *70*, 661−668.
(14) Raychaudhury, C.; Nandy, A. Indexing Scheme and Similarity Measures for Macromolecular Sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.
(15) Randić, M. On Characterization of Fullerenes. *Fullerene Sci. Technol.* **1994**, *2*, 427−444.
(16) Randić, M. On Characterization of DNA Primary Sequences by a Condensed Matrix. *Chem. Phys. Lett.* (in press).
(17) Randić, M. (work in progress).

CI990084Z