A Naturally Logical Representation for DNA Primary Sequences

Chun Li[*,a,b] and Jun Wang[b,c]

[a]*Department of Mathematics, Bohai University, Jinzhou 121000, P.R. China*

[b]*Department of Applied Mathematics and* [c]*College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P.R. China*

Abstract: In this paper, we (1) introduce a logical representation (LR) for DNA primary sequences; (2) show relations between LR and some other representations including the characteristic sequences of a DNA sequence, Randic's 2-D, 4-D representations, and Z-curve (a 3-D graphical representation); and (3) outline the constructions of the *S/S* matrix specific for a logical sequence and its 2*2 condensed matrix.

Keywords: DNA sequence, logical representation, condensed matrix, invariant.

1. INTRODUCTION

With the completion of the genome projects for human, mouse and rat one after another, and the development of some similar projects for other species, the number of known DNA sequences is increasing at an exponential rate. A DNA sequence is usually expressed in terms of a series of four letters A, G, C, and T, which denote the four nucleic acid bases: adenine, guanine, cytosine, and thymine, respectively. This expression is called a letter sequence representation or a DNA primary sequence. DNA primary sequences vary enormously in their length. They might consist of fewer than a hundred bases but also can extend to over a hundred thousand bases. Even when the long sequences are broken down into segments corresponding to exons or introns, the segments corresponding to the same position within a gene and belonging to different species may have different lengths. Clearly, this means that a base-by-base comparison of genes cannot be used. Therefore, the alignment of sequences has been considered in the literature. The standard procedures of alignment consider differences between strings due to deletion-insertion, compression-expansion, and substitution of the string elements. Finding the smallest number of changes that are necessary to match labels in two sequences is far from trivial.

In order to give a visual and hence qualitative characterization of DNA primary sequences, graphical representations were suggested [1-13], which mainly regard a DNA primary sequence as a curve embedded in 2-D plane or 3-D space. Besides these, Randic and Balaban [14] proposed a 4-D representation of DNA primary sequences by assigning to each of the four bases A, T, G, C directions along the four orthogonal coordinate axes. Although it seemed as if the four bases can be assigned in 4!=24 ways, it is not difficult to find that there is only one essential pattern of the 4-D representation corresponding to the same DNA sequence, because the four directions of 4-D space are fully equivalent. Of course, as pointed out by the authors, the important advantage of graphical visualization of DNA sequences has been lost.

In 2002, He and Wang [15] introduced another representation for DNA sequences, which is based on the idea of the coarse-grained description of the DNA primary sequence. As we know, the four nucleic acid bases A, G, C and T can be divided into two classes according to their chemical structures, *i.e*. purine R = {A, G} and pyrimidine Y = {C, T}. The bases can be also divided into another two classes, amino group M = {A, C} and keto group K = {G, T}. In addition, the division can be made according to the strength of the hydrogen bond, *i.e*. weak H-bonds W ={A, T} and strong H-bonds S ={G, C} [11, 15-17]. By labeling the elements of R, M and W by 1, and that of Y, K and S by 0, respectively, He and Wang transform a DNA primary sequence into three (0,1)-sequences, which are named the (R, Y)-, (M, K)-, and (W, S)-characteristic sequences of the DNA sequence, respectively.

In this paper, we offer a naturally logical representation for DNA primary sequences, which is different from the existing representations of DNA sequences while it is closely related to some of them.

2. THE LOGICAL REPRESENTATION FOR DNA PRIMARY SEQUENCES

It is easy to see that the four bases A, G, C and T can be classified into two classes based on the knowledge of logic: A and not-A (G or C or T). Denoting not-A by $\overline{\mathrm{A}}$, we reduce a DNA primary sequence into a binary sequence: by 1 (0) we denote A ($\overline{\mathrm{A}}$). The obtained (0,1)-sequence is called the A-sequence of the DNA sequence. In this representation, some information of the DNA sequence structure may be lost; however, this will give prominence to local information from adenine.

Besides this we can similarly define other three (0,1)-sequences: G-sequence, C-sequence and T-sequence, respectively. We state the above process in mathematical terms as follows.

Let $X = \mathrm{x}_1\mathrm{x}_2\cdots$ be a DNA primary sequence. We define four homomorphism maps ϕ_i (i = 1, 2, 3, 4) by $\phi_i(\mathrm{X}) = \phi_i(\mathrm{x}_1)\phi_i(\mathrm{x}_2)\cdots$, where

*Address correspondence to this author at the Department of Mathematics, Bohai University, Jinzhou 121000, P.R. China;
E-mail: lchlmb@yahoo.com.cn

1386-2073/07 $50.00+.00

$$\phi_1(x_j)=\begin{cases}1 & if\ x_j=A\\ 0 & if\ x_j\neq A\end{cases}\quad (j=1,\ 2,\cdots)$$

$$\phi_2(x_j)=\begin{cases}1 & if\ x_j=C\\ 0 & if\ x_j\neq C\end{cases}\quad (j=1,\ 2,\cdots)$$

$$\phi_3(x_j)=\begin{cases}1 & if\ x_j=G\\ 0 & if\ x_j\neq G\end{cases}\quad (j=1,\ 2,\cdots)$$

$$\phi_4(x_j)=\begin{cases}1 & if\ x_j=T\\ 0 & if\ x_j\neq T\end{cases}\quad (j=1,\ 2,\cdots)$$

We call the four (0,1)-sequences the logical representation (or simply **LR**) of the DNA sequence considered.

Taking the first 60 bases of the exon-1 of the human *beta*-globin gene as an example; we list its **LR** in Table **1**.

Table 1. The Logical Representation (LR) of the DNA Sequence

Sequence (*S*)	ATGGTGCACC TGACTCCTGA GGAGAAGTCT GCCGTTACTG CCCTGTGGGG CAAGGTGAAC
$LR^A(S)$	1000000100001000000100101100000000001000000000000000110000110
$LR^C(S)$	0000001011000101100000000000100110000100111000000010000000001
$LR^G(S)$	0011010000010000001011010010001001000001000010111100011010000
$LR^T(S)$	0100100000100010010000000001010000110010000101000000000010000

According to the definition of **LR** and Table **1**, we have:

(1) The **LR** gives all information of the corresponding DNA primary sequence, because a DNA primary sequence is uniquely determined by any three of its four logical sequences.

(2) The **LR** clearly shows the distribution of the four nucleic acid bases. Therefore, the information on the relative abundance and the location of a certain nucleotide in a DNA sequence and its segments can be obtained directly.

(3) The **LR** can be used to numerically identify the point mutations. In particular, it is easier to identify whether the substitutions are transitions ($T\leftrightarrow C$, $A\leftrightarrow G$) or transversions ($T\leftrightarrow A$, $T\leftrightarrow G$, $C\leftrightarrow A$, $C\leftrightarrow G$). For our purpose, we write the **LR** of a sole nucleotide Y as a 4-D vector $y=(LR^A(Y), LR^G(Y), LR^C(Y), LR^T(Y))$, and denote it by $y=(y_a,y_g,y_c,y_t)$ for brevity. For two nucleotides Y and Z, if $y\cdot z=1$, where "·" is the inner product between two vectors, then they are matched bases. While $y\cdot z=0$ implies that there is a substitution. To further identify the substitution, we only need to investigate $y_a+z_a+y_g+z_g:=yz_{ag}$. Obviously, if $yz_{ag}=1$, it must be a transversion; otherwise, it must be a transition.

3. COMPARISONS WITH OTHER REPRESENTATIONS

3.1. The LR and the Characteristic Sequences

For a given DNA primary sequence *S*, from the definition of **LR** and Table **1**, we see that

(M, K)-characteristic sequence = $LR^A(S) \vee LR^C(S)$,

(R, Y)-characteristic sequence = $LR^A(S) \vee LR^G(S)$,

(W, S)-characteristic sequence = $LR^A(S) \vee LR^T(S)$,

where $\vee$ denotes the logical sum.

3.2. The LR and Randic's 2-D Graph

In [8, 9], Randic *et al.* proposed a 'four horizontal lines' graph of DNA primary sequence. They draw four horizontal lines separated by unit distances, on which dots (rectangles) representing the bases constituting the considered sequence are placed. The representation requires first that the four types of bases are associated with the four horizontal lines. They label the lines in the order as the types of bases appear for the first time in the sequence, which is A, T, G, and C. The consecutive bases along the horizontal axes are placed at unit displacement. By connecting adjacent dots, a zigzag like curve is obtained (Fig. **1**).

For the same DNA sequence, ATGGTGCACCTGACTC CTGA, the first 20 bases of the exon-1 of human *beta*-globin

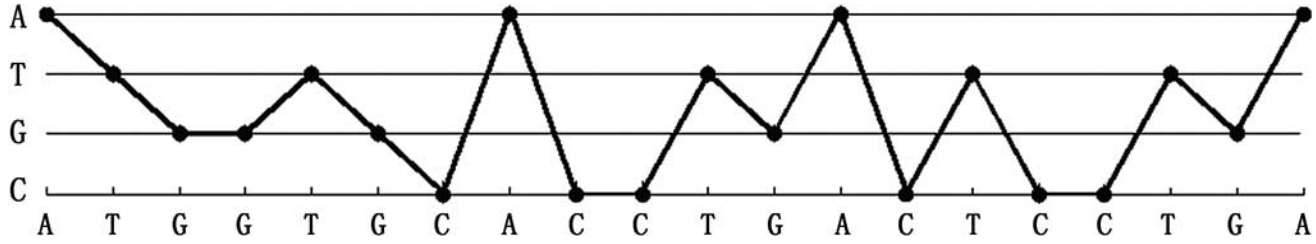

Fig. (1). The 2-D graphical representation of the sequence ATGGTGCACCTGACTCCTGA based on four horizontal lines. The dots denote the bases making up the sequence.

gene, we list its **LR** in the order of A-T-G-C in Table **2**. Suppressing 'zeros' and linking only 'ones' one after another, we immediately obtain the same 2-D graphical representation of the DNA sequence.

Table 2. The LR of the Sequence ATGGTGCACCTGACT CCTGA

$LR^A(S)$	1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1
$LR^T(S)$	0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0
$LR^G(S)$	0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0
$LR^C(S)$	0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 1 1 0 0 0

3.3. The LR and Randic's 4-D Representation

In [14], a 4-D representation of DNA primary sequences was considered by assigning to each of the four bases A, T, G, C directions along the four orthogonal coordinate axes:

A (1, 0, 0, 0)

T (0, 1, 0, 0)

G (0, 0, 1, 0)

C (0, 0, 0, 1)

Taking the first 15 bases of the exon-1 of human *beta*-globin gene as an example, its 4-D coordinates are listed in Table **3**.

Table 3. 4-D Coordinates for the First 15 Bases of the First Exon of Human *Beta*-Globin Gene[a]

Human *Beta*-Globin	
No.	**1 2 3 4 5 6 7 8 9 10 11 12 13 14 15**
Base	**A T G G T G C A C C T G A C T**
e_A	1 1 1 1 1 1 1 2 2 2 2 2 3 3 3
e_T	0 1 1 1 2 2 2 2 2 2 3 3 3 3 4
e_G	0 0 1 2 2 3 3 3 3 3 3 4 4 4 4
e_C	0 0 0 0 0 0 1 1 2 3 3 3 3 4 4

[a]Taken from Table **2** [14].

For our purpose, by $LR_i^Y(S)$ we denote the *i*-th 'base' in the logical sequence $LR^Y(S)$, (*Y*=A, T, G, C). Define

$$\begin{cases} x_{1i} = \sum_{k=1}^{i} LR_k^A(S) \\ x_{2i} = \sum_{k=1}^{i} LR_k^T(S) \\ x_{3i} = \sum_{k=1}^{i} LR_k^G(S) \\ x_{4i} = \sum_{k=1}^{i} LR_k^C(S) \end{cases}$$

then, we immediately obtain the same 4-D coordinates (cf. Table **2**).

3.4. The LR and the Z-Curve

Consider a DNA sequence read from the 5′to the 3'-end with *N* bases. Inspect it by stepping one base at a time. In the step *n* (*n* =1, 2, ..., *N*), count the *cumulative* occurrence numbers of A, C, G and T, respectively, in the subsequence from the 1st base to the *n*-th base in the sequence inspected, and denote them by *An*, *Cn*, *Gn* and *Tn*, respectively. Then Zhang and Zhang [11] proposed a 3-D curve which consists of a series of nodes P_n, whose coordinates x_n, y_n and z_n are constructed as follows:

$$\begin{cases} x_n = 2(A_n + G_n) - n, \\ y_n = 2(A_n + C_n) - n, \\ z_n = 2(A_n + T_n) - n, \end{cases}$$
$$x_n, y_n, z_n \in [-N, N], \quad n = 0,1,2,\cdots,N,$$

where $A_0 = C_0 = G_0 = T_0 = 0$ and hence P_0=0. The Z curve is defined as the connection of the nodes P_0, P_1, P_2, ..., P_N one by one sequentially with straight lines.

From the definition of **LR** and Table **1**, it is easy to see that $Y_n = \sum_{k=1}^{n} LR_k^Y(S)$, (*Y*=A, T, G, C). So, we can obtain the above nodes P_n's from the logical sequences directly, and then the Z curve.

4. THE *S/S* MATRIX REPRESENTATION FOR LOGICAL SEQUENCES

For a DNA primary sequence, Randic [18] introduced a *S/S* matrix by regarding the sequence itself as an input. Similarly, we consider the *S/S* matrix for the logical sequences.

Suppose that $b_1b_2b_3\ldots b_n$ is a (0,1)-sequence, we construct its $S/S=(s_{ij})$ matrix as follows:

$$s_{ji} = s_{ij} = \frac{n_{ij}}{j-i}, \ (j>i); \ s_{ii} = 0,$$

where n_{ij} is the number of the 'logical bases' which are identical to the one corresponding to b_j in the sub-string $b_{i+1}b_{i+2}\ldots b_j$.

Note that the **LR** of a DNA sequence includes four logical sequences. Therefore, from a given DNA sequence, one can obtain four *S/S* matrices following the method above. Consider the first 20 bases of the exon-1 of human *beta*-globin gene:

$A_1\ T_2\ G_3\ G_4\ T_5\ G_6\ C_7\ A_8\ C_9\ C_{10}\ T_{11}$
$G_{12}\ A_{13}\ C_{14}\ T_{15}\ C_{16}\ C_{17}\ T_{18}\ G_{19}\ A_{20}$

In Table **4**, we show the *S/S* matrix corresponding to its G-sequence (see $LR^G(S)$ of Table **2**).

Observing Table **4**, we see that the entries next to the main diagonal in the *S/S* matrix are always equal to 1. In fact, by the definition of *S/S*, the matrix elements for all adjacent pairs of bases must equal 1 and for nonadjacent ones, the matrix elements are less than or equal to 1. Such a matrix has an interesting advantage, that is, from which one can construct a convergent sequence of matrices $^kS/^kS$, (*k*=1, 2, 3, …), whose (*i*, *j*)-element is s_{ij}^k. Clearly, as *k* trends to in-

Table 4. The Upper Triangles of the Matrix *S/S* Corresponding to $LR^G(S)$

S/S	0_1	0_2	1_3	1_4	0_5	1_6	0_7	0_8	0_9	0_{10}	0_{11}	1_{12}	0_{13}	0_{14}	0_{15}	0_{16}	0_{17}	0_{18}	1_{19}	0_{20}
0_1	0	1/1	1/2	2/3	2/4	3/5	3/6	4/7	5/8	6/9	7/10	4/11	8/12	9/13	10/14	11/15	12/16	13/17	5/18	14/19
0_2		0	1/1	2/2	1/3	3/4	2/5	3/6	4/7	5/8	6/9	4/10	7/11	8/12	9/13	10/14	11/15	12/16	5/17	13/18
1_3			0	1/1	1/2	2/3	2/4	3/5	4/6	5/7	6/8	3/9	7/10	8/11	9/12	10/13	11/14	12/15	4/16	13/17
1_4				0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	2/8	7/9	8/10	9/11	10/12	11/13	12/14	3/15	13/16
0_5					0	1/1	1/2	2/3	3/4	4/5	5/6	2/7	6/8	7/9	8/10	9/11	10/12	11/13	3/14	12/15
1_6						0	1/1	2/2	3/3	4/4	5/5	1/6	6/7	7/8	8/9	9/10	10/11	11/12	2/13	12/14
0_7							0	1/1	2/2	3/3	4/4	1/5	5/6	6/7	7/8	8/9	9/10	10/11	2/12	11/13
0_8								0	1/1	2/2	3/3	1/4	4/5	5/6	6/7	7/8	8/9	9/10	2/11	10/12
0_9									0	1/1	2/2	1/3	3/4	4/5	5/6	6/7	7/8	8/9	2/10	9/11
0_{10}										0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	2/9	8/10
0_{11}											0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	2/8	7/9
1_{12}												0	1/1	2/2	3/3	4/4	5/5	6/6	1/7	7/8
0_{13}													0	1/1	2/2	3/3	4/4	5/5	1/6	6/7
0_{14}														0	1/1	2/2	3/3	4/4	1/5	5/6
0_{15}															0	1/1	2/2	3/3	1/4	4/5
0_{16}																0	1/1	2/2	1/3	3/4
0_{17}																	0	1/1	1/2	2/3
0_{18}																		0	1/1	1/2
1_{19}																			0	1/1
0_{20}																				0

finity, the limit of the matrices sequence $\{ {}^kS/{}^kS \}$ turns into a (0,1)-matrix, which is denoted by ${}^bS/{}^bS$.

However, as we have seen, the orders of *S/S* and its 'higher order' matrices are always equal to the lengths of the sequences considered. Therefore, lengthy sequences will produce such very 'large' matrices that they will have to be processed somehow if useful information is to be gained. To deal with this problem, one can consider the following two aspects.

4.1. The Invariant

Once a real symmetric matrix representation of a sequence is given, some of invariants extracted from the matrix, such as the average matrix element, the average row sum, the leading eigenvalue, and the `**ALE**-index', can be used as descriptors of the sequence [8, 9, 12-15,19-22]. For instance, the `**ALE**-index' corresponding to the *S/S* matrix above is easily calculated as 14.4126, and its normalized value is 0.72063.

4.2. The Condensed Matrix

Similar to the 4×4 condensed matrix specific for the DNA primary sequence [18], we propose a 2×2 condensed matrix the elements of which are all extracted from the matrix *S/S* associated with a logical sequence. We present briefly the method for constructing a condensed matrix as follows.

Suppose that the logical sequence considered consists of *m* 1's and *n* 0's. Thus, by gathering all elements of *S/S* corresponding to bases 1's, one can obtain a partitioned matrix:

	0…..0	1…1
0 …	$00_{n\times n}$	$01_{n\times m}$
1 …	$10_{m\times n}$	$11_{m\times m}$

replacing each partition by a suitable number, a very condensed 2×2 matrix with pertinent information is formed. Various sub-matrices invariants can be selected as the 'representative'. Of course, generally speaking, $m \neq n$, therefore the (leading) eigenvalue is usually avoided.

As an example, in Table **5** we show the partitioned matrix corresponding to the *S/S* matrix above.

If we use the average matrix element of each sub-matrix as its 'representative', then a 2×2 condensed matrix can be obtained:

$$\begin{pmatrix} 0.75469 & 0.69154 \\ 0.69154 & 0.29307 \end{pmatrix}_{2\times2}$$

5. CONCLUSION

According to a classification of four nucleic acid bases, we introduce a naturally logical representation (**LR**) for DNA primary sequences. Then we compare it with some other representations. They are the characteristic sequences proposed by He and Wang [15], the 2-D'four horizontal lines'graph and the 4-D representation of DNA sequences proposed by Randic *et al.* [8, 9, 14], and the Z-curve proposed by Zhang and Zhang [11]. The results show that the

Table 5. The Partitioned Matrix

0_1	0_2	0_5	0_7	0_8	0_9	0_{10}	0_{11}	0_{13}	0_{14}	0_{15}	0_{16}	0_{17}	0_{18}	0_{20}	l_3	l_4	l_6	l_{12}	l_{19}
0	**1/1**	**2/4**	**3/6**	**4/7**	**5/8**	**6/9**	**7/10**	**8/12**	**9/13**	**10/14**	**11/15**	**12/16**	**13/17**	**14/19**	1/2	2/3	3/5	4/11	5/18
1/1	**0**	**1/3**	**2/5**	**3/6**	**4/7**	**5/8**	**6/9**	**7/11**	**8/12**	**9/13**	**10/14**	**11/15**	**12/16**	**13/18**	1/1	2/2	3/4	4/10	5/17
2/4	**1/3**	**0**	**1/2**	**2/3**	**3/4**	**4/5**	**5/6**	**6/8**	**7/9**	**8/10**	**9/11**	**10/12**	**11/13**	**12/15**	1/2	1/1	1/1	2/7	3/14
3/6	**2/5**	**1/2**	**0**	**1/1**	**2/2**	**3/3**	**4/4**	**5/6**	**6/7**	**7/8**	**8/9**	**9/10**	**10/11**	**11/13**	2/4	2/3	1/1	1/5	2/12
4/7	**3/6**	**2/3**	**1/1**	**0**	**1/1**	**2/2**	**3/3**	**4/5**	**5/6**	**6/7**	**7/8**	**8/9**	**9/10**	**10/12**	3/5	3/4	2/2	1/4	2/11
5/8	**4/7**	**3/4**	**2/2**	**1/1**	**0**	**1/1**	**2/2**	**3/4**	**4/5**	**5/6**	**6/7**	**7/8**	**8/9**	**9/11**	4/6	4/5	3/3	1/3	2/10
6/9	**5/8**	**4/5**	**3/3**	**2/2**	**1/1**	**0**	**1/1**	**2/3**	**3/4**	**4/5**	**5/6**	**6/7**	**7/8**	**8/10**	5/7	5/6	4/4	1/2	2/9
7/10	**6/9**	**5/6**	**4/4**	**3/3**	**2/2**	**1/1**	**0**	**1/2**	**2/3**	**3/4**	**4/5**	**5/6**	**6/7**	**7/9**	6/8	6/7	5/5	1/1	2/8
8/12	**7/11**	**6/8**	**5/6**	**4/5**	**3/4**	**2/3**	**1/2**	**0**	**1/1**	**2/2**	**3/3**	**4/4**	**5/5**	**6/7**	7/10	7/9	6/7	1/1	1/6
9/13	**8/12**	**7/9**	**6/7**	**5/6**	**4/5**	**3/4**	**2/3**	**1/1**	**0**	**1/1**	**2/2**	**3/3**	**4/4**	**5/6**	8/11	8/10	7/8	2/2	1/5
10/14	**9/13**	**8/10**	**7/8**	**6/7**	**5/6**	**4/5**	**3/4**	**2/2**	**1/1**	**0**	**1/1**	**2/2**	**3/3**	**4/5**	9/12	9/11	8/9	3/3	1/4
11/15	**10/14**	**9/11**	**8/9**	**7/8**	**6/7**	**5/6**	**4/5**	**3/3**	**2/2**	**1/1**	**0**	**1/1**	**2/2**	**3/4**	10/13	10/12	9/10	4/4	1/3
12/16	**11/15**	**10/12**	**9/10**	**8/9**	**7/8**	**6/7**	**5/6**	**4/4**	**3/3**	**2/2**	**1/1**	**0**	**1/1**	**2/3**	11/14	11/13	10/11	5/5	1/2
13/17	**12/16**	**11/13**	**10/11**	**9/10**	**8/9**	**7/8**	**6/7**	**5/5**	**4/4**	**3/3**	**2/2**	**1/1**	**0**	**1/2**	12/15	12/14	11/12	6/6	1/1
14/19	**13/18**	**12/15**	**11/13**	**10/12**	**9/11**	**8/10**	**7/9**	**6/7**	**5/6**	**4/5**	**3/4**	**2/3**	**1/2**	**0**	13/17	13/16	12/14	7/8	1/1
1/2	1/1	1/2	2/4	3/5	4/6	5/7	6/8	7/10	8/11	9/12	10/13	11/14	12/15	13/17	**0**	**1/1**	**2/3**	**3/9**	**4/16**
2/3	2/2	1/1	2/3	3/4	4/5	5/6	6/7	7/9	8/10	9/11	10/12	11/13	12/14	13/16	**1/1**	**0**	**1/2**	**2/8**	**3/15**
3/5	3/4	1/1	1/1	2/2	3/3	4/4	5/5	6/7	7/8	8/9	9/10	10/11	11/12	12/14	**2/3**	**1/2**	**0**	**1/6**	**2/13**
4/11	4/10	2/7	1/5	1/4	1/3	1/2	1/1	1/1	2/2	3/3	4/4	5/5	6/6	7/8	**3/9**	**2/8**	**1/6**	**0**	**1/7**
5/18	5/17	3/14	2/12	2/11	2/10	2/9	2/8	1/6	1/5	1/4	1/3	1/2	1/1	1/1	**4/16**	**3/15**	**2/13**	**1/7**	**0**

logical representation may be a useful tool for characterizing and comparing DNA sequences. To give a quantitative analysis of data on DNA sequences, we construct a *S/S* matrix specific for a logical sequence and its family of associated 'higher order' matrices ${}^{k}S/{}^{k}S$ (k=2,3,…). From these matrices, one can derive some structurally related descriptors and the corresponding 2×2 condensed matrices.

ACKNOWLEDGEMENTS

This work was partially supported by the Science Research Project of Educational Department of Liaoning Province and the National Natural Science Foundation of China.

REFERENCES

[1] Nandy, A. *Curr. Sci.*, **1994**, *66*, 309.
[2] Nandy, A. *Curr. Sci.*, **1994**, *66*, 821.
[3] Nandy, A.; Nandy, P. *Curr. Sci.*, **1995**, *68*, 75.
[4] Nandy, A. *Curr. Sci.*, **1996**, *70*, 661
[5] Nandy, A.; Nandy, P. *Chem. Phys. Lett.*, **2003**, *368*, 102.
[6] Guo, X.F.; Randic, M.; Basak, S.C. *Chem. Phys. Lett.*, **2001**, *350*, 106.
[7] Yau, S.S.T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y.K. *Nucleic Acids Res.*, **2003**, *31*, 12.
[8] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. *Chem. Phys. Lett.*, **2003**, *368*, 1.
[9] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. *Chem. Phys. Lett.*, **2003**, *371*, 202.
[10] Hamori, E.; Ruskin, J. *J. Biol. Chem.*, **1983**, *258*, 1318.
[11] Zhang, R.; Zhang, C.T. *J. Biomol. Struc. Dyn.*, **1994**, *11*, 767.
[12] Randic, M.; Vracko, M.; Nandy, A.; Basak, S.C. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1235.
[13] Li, C.; Wang, J. *Comb. Chem. High Throughput Screen.*, **2004**, *7*, 23.
[14] Randic, M.; Balaban, A.T. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 532.
[15] He, P-An; Wang, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1080.
[16] Zhang, C.T. *J. Theor. Biol.*, **1997**, *187*, 297-306.
[17] Cornish-Bowden, A. *Nucleic Acids Res.*, **1985**, 13, 3021-3030.
[18] Randic, M. *Chem. Phys. Lett.*, **2000**, 317, 29.
[19] Randic, M.; Vracko, M. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 599-606.
[20] Randic, M.; Guo, X.F.; Basak, S.C. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 619.
[21] Li, C.; Wang, J. *Comb. Chem. High Throughput Screen.*, **2003**, *6*, 795.
[22] Li, C.; Wang, J. *J. Chem. Inf. Model.*, **2005**, *45*, 115-120.

Received: August 31, 2005 **Revised: October 27, 2005** **Accepted: December 13, 2006**