Internet Electronic Journal of Molecular Design **2002**, *1*, 675–684 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

December 2002, Volume 1, Number 12, Pages 675–684

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 4

Guest Editor: Jun–ichi Aihara

The Numerical Characterization and Similarity Analysis of DNA Primary Sequences

Yachun Liu

Department of Mathematics and Physical Science, Nanhua University, Hengyang 421001, Hunan, P. R. China

Received: July 24, 2002; Revised: August 19, 2002; Accepted: October 29, 2002; Published: December 31, 2002

Citation of the article:
Y. Liu, The Numerical Characterization and Similarity Analysis of DNA Primary Sequences, *Internet Electron. J. Mol. Des.* **2002**, *1*, 675–684, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

The Numerical Characterization and Similarity Analysis of DNA Primary Sequences[#]

Yachun Liu*

Department of Mathematics and Physical Science, Nanhua University, Hengyang 421001, Hunan, P. R. China

Received: July 24, 2002; Revised: August 19, 2002; Accepted: October 29, 2002; Published: December 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (12), 675–684

Abstract

Motivation. DNA sequencing has become routine and has resulted in an abundance of data on primary sequences of DNA for various species. Hence, we faced the task of process such great amount of data, which poses a number of yet unsolved problems. The motivation of this paper is to introduce a new numerical characterization of DNA sequences.

Method. We define a scheme to give a logic order of DNA primary sequences in term of the classification of nucleic acid bases. Using logic sequences we generate a set of 4×6 matrices to represent DNA primary sequences, which are based on counting all (0,1) triplets in the logic sequences. Using the condensed representation of primary DNA primary sequences and the eigenvalues of the corresponding symmetric real matrix a comparison is made between the primary sequences for exon–1 of human β–globin and seven other species.

Results. With this procedure we extend the matrix method to determine new invariants as descriptors for DNA sequences.

Conclusions. On the basis of this new scheme, we find that a new similarity index, the informational compression ratio, can characterize the evolution relationships for different species.

Keywords. DNA sequence; β–globin gene; similarity analysis; condensed matrix; DNA descriptor index; evolutionary rates and gradient.

1 INTRODUCTION

In recent years, effective representation of long DNA sequences has led to several innovative techniques to provide useful ways for viewing, sorting, analyzing, and comparing various sequences. For example, Gate, Nandy, Leong, Mogenthaler, and Randić have defined methods for representing graphically DNA sequences using a two–dimensional Cartesian coordinate system [1–7]. These methods are based on choosing four directions in the *x*,*y* coordinate system to represent

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; phone: 86–0734–8282580; E–mail: Liuyachun65@263.net.

the content of the four bases in DNA sequences. The algorithm essentially consists of plotting a point corresponding to a base by moving a step in a direction depending on the defined association of a base with the direction. The cumulative plot of such points produces a graph that corresponds to the sequence of bases in the gene fragment under consideration.

It is clear that the difference in the base composition and distribution of individual members of a homologous family will induce changes in the plot of the sequences in the graphical representation. Nevertheless, this method has some disadvantages. First, graphical approaches involve to a greater or lesser degree arbitrary conventions when the assignment of the direction in the *x*,*y* plane are selected for the nucleic bases A, C, G and T, where A, C, G, T are the codes respectively for adenine, cytosine, guanine and thymine. Second, DNA primary sequences vary enormously in their length. 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 length.

In 1986, Gate [1] proposed a Manhattan distance approach only for an equal length sequence. In 1996, Nandy [8] estimated the divergence of two graphs by calculating the plot density, *i.e*., the ratio of the number of points and the enclosing area for the points, but this method misses out on difference arising out of shape changes within the same overall distribution. In recent years, it was proposed a numerical characterization of graphical representation of DNA primary sequences, and further proposed a method to compare more graphs in order to provide a quantitative estimate of the divergence of the different sequences. However, because the graphical representation of a shorter DNA sequence may correspond to more DNA sequences (for example, sequences AG, AGAG, AGAGAG, …, have the same graphical representation), much more work needs to be done to forge this technique into a precision tool for characterization of gene sequences.

A possible strategy to avoid such difficulties is to represent the DNA sequences by suitably constructed matrices. A way to arrive at a numerical matrix for a sequence, or a structure such as a molecule, is by first imbedding the sequence (or a structure) in a 2D or 3D space. For a system of fixed geometry one can consider for any two elements their Euclidean distance (‘through space distance’) and their graph theoretical distance (‘through bond distance’). The D/D matrix [9], the elements of which are given as the quotient of Euclidean and the graph theoretical distance, have been shown to lead to useful structural invariants. In this way one arrives at an index that has been structurally interpreted as measuring the degree of folding of chain molecules, planar or spatial curves, and fractals.

Instead of using a geometrical representation of the primary DNA sequences, Randić [10] constructed the S/S matrix. The entry of the S/S matrix is the quotient of serial distance between selected labels of one kind only and the sequence distance when all labels are counted in the primary DNA sequence. The large S/S matrices belonging to lengthy primary DNA sequences are

reduced to smaller matrices, namely AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC. Using the average matrix element of each sub matrix as its 'representative' in the 4×4 condensed matrix, the large initial matrix is reduced to symmetrical 4×4 matrix with, at most, ten different entries. Such matrices not only offer some instant insight into the nature of the primary sequence of DNA but also allow one to make qualitative and quantitative comparisons between different DNA sequences, whether within the same or between different species.

Randić introduced a 4×4 condensed matrix representation of primary DNA sequences that offers an alternative method of transforming the DNA sequences into a numerical value [11]. Instead of analyzing the primary sequence by constructing a large matrix, one can associate a smaller matrix with the segment of DNA in which the rows and columns are assigned to individual nucleic bases of the same kind. The entry (*X*, *Y*) of the matrix is assigned to the frequency of occurrence of (*X*, *Y*) as adjacent entries in the primary sequence of DNA, where $X, Y \in \{A, C, G, T\}$. Clearly, there is a greater loss of information when one condenses the primary sequence of DNA into a 4×4 matrix by this method. Due to this drawback, Randić has done a lot of work on the recovery of lost information, and constructed the additional 4×4 condensed matrices for nonadjacent pairs of nucleic bases at distance 2, 3, …, 6 and so on. In addition, he did not pay attention to the influence of different length of DNA primary sequences, so, the usefulness and reliability of the present approach would be evaluated in comparison with the of exon–1 of human and seven other species β–globins.

2 METHODS

2.1 The Construction of Logic Sequences Based on the Classification of Nucleic Acids

In this contribution we adopt the procedure proposed by He and Wang for encoding into a numerical form the DNA sequence [12]. Nucleic acids and proteins are all linear macromolecules. Although in this approach we consider only the primary structure of DNA, the 3D structure of DNA should be considered in order to have a comprehensive picture of the DNA similarity.

In DNA sequences, the four bases A, C, G, T can be divided into two classes according to their chemical structures, *i.e.* purine P_1 = {A, G} and pyrimidine P_2 = {C, T}. The bases can also be divided into another two classes, amino group G_1 = {A, C} and keto group G_2 = {G, T}. In addition, the division can also be made according to the strength of the hydrogen bond, *i.e.* weak H–bonds H_1 = {A, T} and strong H–bonds H_2 = {G, C}. For a DNA sequence, using each classification, we can rewrite the sequence into a (0,1) sequence. For example, ATG becomes 101 in the $P_1(P_2)$–schema, 100 in $G_1(G_2)$–schema and 110 in the $H_1(H_2)$–schema.

Table 1. Exon–1 of the β–Globin Genes for Eight Species

No	Species	Bases	Length
1	Human	ATGGT GCACC TGACT CCTGA GGAGA AGTCT GCCGT TACTG CCCTG TGGGG CAAGG TGAAC GTGGA TGAAG TTGGT GGTGA GGCCC TGGGC AG	92
2	Goat	ATGCT GACTG CTGAG GAGAA GGCTG CCGTC ACCGG CTTCT GGGGC AAGGT GAAAG TGGAT GAAGT TGGTG CTGAG GCCCT GGGCA G	86
3	Opossum	ATGGT GCACT TGACT TCTGA GGAGA AGAAC TGCAT CACTA CCATC TGGTC TAAGG TGCAG GTTGA CCAGA CTGGT GGTGA GGCCC TTGGC AG	92
4	Gallus	ATGGT GCACT GGACT GCTGA GGAGA AGCAG CTCAT CACCG GCCTC TGGGG CAAGG TCAAT GTGGC CGAAT GTGGG GCCGA AGCCC TGGCC AG	92
5	Lemur	ATGAC TTTGC TGAGT GCTGA GGAGA ATGCT CATGT CACCT CTCTG TGGGG CAAGG TGGAT GTAGA GAAAG TTGGT GGCGA GGCCT TGGGC AG	92
6	Mouse	ATGGT TGCAC CTGAC TGATG CTGAG AAGTC TGCTG TCTCT TGCCT GTGGG CAAAG GTGAA CCCCG ATGAA GTTGG TGGTG AGGCC CTGGG CAG	93
7	Rabbit	ATGGT GCATC TGTCC AGTGA GGAGA AGTCT GCGGT CACTG CCCTG TGGGG CAAGG TGAAT GTGGA AGAAG TTGGT GGTGA GGCCC TGGGC	90
8	Rat	ATGGT GCACC TAACT GATGC TGAGA AGGCT ACTGT TAGTG GCCTG TGGGG AAAGG TGAAC CCTGA TAATG TTGGC GCTGA GGCCC TGGGC AG	92

In this representation, some information of the DNA sequence structure may be lost, however, it is easier to compare sequences. We perform similar operations on the sequence according to the second and third classifications so that the loss of information of the sequence can be greatly reduced. Thus, we obtain three (0,1) sequences corresponding to the same DNA primary sequence, and we call them as logic sequences of the DNA primary sequence over (P_1, P_2), (G_1, G_2), and (H_1, H_2), respectively. For example, the logic sequences for exon–1 of the β–globin gene for humans (species 1 in Table 1) are presented in Table 2.

Analogously, we derive the other seven species logic sequences for their β–globin genes, and the *i*–th species logic sequences are listed in Table $i + 1$, $i = 2, 4, \ldots, 7$. In the next subsection, we will generate a type of condensed matrices by considering the frequencies of occurrence of (0, 1) triplets based on the logic sequences for the species. As will be seen, the sequence length will standardize the condensed matrices when different lengths of DNA sequences are compared together.

Table 2. The logic sequences of exon–1 of the β–globin gene for human (species 1)

(P_1, P_2)	10110101000110000011111111100010010010010000101111011110111010111101110011011011110000111011
(G_1, G_2)	10000011110011011001001011001001100011001110000000111000011100001001100000000001001110000110
(H_1, H_2)	11001001001010100101001011010100001110100001010000011001011001001011101100100101000001000010

Table 3. The logic sequences of exon–1 of the β–globin gene for goat (species 2)

(P_1, P_2)	10100110010011111111111001100100100100110000011110111101111101110111100110100111100001110 11
(G_1, G_2)	10010011001001001011001001100111100100100000111000011100001001100000000100100111000011 0
(H_1, H_2)	11001010100101001011000100001010000011010000011001011101001101101100100101000001000010

Table 4. The logic sequences of exon–1 of the β–globin gene for opossum (species 3)

(P_1, P_2)	10110101000110000011111111111001010010010010001100011110101110011001110011011011110000011011
(G_1, G_2)	10000011100011001001001011011100110111011110100001011000011000001111011000000001001110000110
(H_1, H_2)	11001001011010110101001011011010011010110011010010111001001001101001010100100101000001100010

Table 5. The logic sequences of exon–1 of the β–globins gene for gallus (species 4)

(P_1, P_2)	10110101001110010011111111101100010010011000001111011110011010110011101011110011110000110011
(G_1, G_2)	10000011100011001001001011011010110111100110100000111000111000001101100000001101101110001110
(H_1, H_2)	11001001010010100101001011001001011010000001010000011001011101000001110100000001100001000010

Table 6. The logic sequences of exon–1 of the β–globin gene for lemur (species 5)

(P_1, P_2)	10110000100111010011111111101000101001000000010111101111011101011111111100110110111110000111011
(G_1, G_2)	10011000010010001001001011001011000111101010000000111000001000101011100000000101001100000110
(H_1, H_2)	11010111001010100101001011100101101010010101010000011001001101101011101100100001000011000010

Table 7. The logic sequences of exon–1 of the β–globin gene for mouse (species 6)

(P_1, P_2)	10110010100011001101001111111000100100000010001011101111101110000110111100110110111100001110111
(G_1, G_2)	10000001111001100100100101100100100010100011000000111100001111110100110000000000100111000011100
(H_1, H_2)	11001100100101010110010101101010010101011000101000011100101100000110110110010010100000100001000

Table 8. The logic sequences of exon–1 of the β–globin gene for rabbit (species 7)}

(P_1, P_2)	10110101000100011011111111110001011001001000010111101111011101011111111100110110111100001110
(G_1, G_2)	10000011010001110001001011001001000111001110000000111000011000001101100000000000100111000011100
(H_1, H_2)	11001001101010010101001011010100001010100001010000011001011101001101101100100101000001000010000

Table 9. The logic sequences of exon–1 of the β–globin gene for rat (species 8)

(P_1, P_2)	10110101000110011010011111111001001001101100010111111111101110000110110100110100111100001111011
(G_1, G_2)	10000011110111001001001011001011000010000110000000111000011111001011000000101001001110000110
(H_1, H_2)	11001001001110101100101011000110101110100001010000111001011000101111101100000101000001000010

2.2 The Construction of Novel Condensed Matrices

In each DNA logic sequence, there are eight possible triplets that can occur: 000, 001, 010, 011, 100, 101, 110 and 111. Therefore, for a given exon–1 of the β–globin gene, there exist 24 triplets, *i.e*. $P_2P_2P_2$, $P_2P_2P_1$, $P_2P_1P_2$, $P_2P_1P_1$, $P_1P_2P_2$, $P_1P_2P_1$, $P_1P_1P_2$, $P_1P_1P_1$, $G_2G_2G_2$, $G_2G_2G_1$, $G_2G_1G_2$, $G_2G_1G_1$, $G_1G_2G_2$, $G_1G_2G_1$, $G_1G_1G_2$, $G_1G_1G_1$, and $H_2H_2H_2$, $H_2H_2H_1$, $H_2H_1H_2$, $H_2H_1H_1$, $H_1H_2H_2$, $H_1H_2H_1$, $H_1H_1H_2$, $H_1H_1H_1$. For example, $P_1P_2P_2$ signifies that X, Y, Z belong to P_1, P_2 and P_2 respectively, *i.e*. $P_1P_2P_2$ is equivalent to the logic triplet 100, in which XYZ is one of triplet codons of a DNA primary sequence, X, Y, Z ∈ {A, C, G, T}. We transform the above 24 logic triplets into a 4×6 condensed matrix as follows:

$$M = \frac{1}{l-2} \cdot \begin{bmatrix} P_2P_2P_2 & P_2P_2P_1 & G_1G_2G_2 & G_1G_2G_1 & H_1H_2H_2 & H_1H_2H_1 \\ P_2P_1P_2 & P_2P_1P_1 & G_1G_1G_2 & G_1G_1G_1 & H_1H_1H_2 & H_1H_1H_1 \\ P_1P_2P_2 & P_1P_2P_1 & G_2G_2G_2 & G_2G_2G_1 & H_2H_2H_2 & H_2H_2H_1 \\ P_1P_1P_2 & P_1P_1P_1 & G_2G_1G_2 & G_2G_1G_1 & H_2H_1H_2 & H_2H_1H_1 \end{bmatrix} \quad (1)$$

where l is the length of DNA primary sequence. Counting the enumeration of the frequency of occurrence of the 24 logic (0, 1) triplets, we can obtain the corresponding matrix for each species listed in Table 1. If the condensed matrix of the i–th type of species is denoted by M_i, i = 1, 2, …, 8, then

$$M_1 = \frac{1}{90}\cdot\begin{bmatrix} 9 & 9 & 15 & 2 & 15 & 12 \\ 8 & 13 & 12 & 6 & 7 & 2 \\ 9 & 12 & 23 & 15 & 12 & 15 \\ 12 & 18 & 5 & 12 & 21 & 6 \end{bmatrix} \qquad M_2 = \frac{1}{84}\cdot\begin{bmatrix} 5 & 11 & 18 & 1 & 13 & 11 \\ 6 & 12 & 9 & 5 & 8 & 1 \\ 11 & 7 & 14 & 18 & 14 & 13 \\ 11 & 21 & 10 & 9 & 17 & 7 \end{bmatrix}$$

$$M_3 = \frac{1}{90}\cdot\begin{bmatrix} 9 & 12 & 12 & 7 & 14 & 16 \\ 8 & 13 & 13 & 8 & 11 & 1 \\ 12 & 9 & 19 & 12 & 4 & 14 \\ 12 & 15 & 6 & 13 & 20 & 10 \end{bmatrix} \qquad M_4 = \frac{1}{90}\cdot\begin{bmatrix} 6 & 12 & 11 & 9 & 14 & 10 \\ 7 & 14 & 15 & 7 & 7 & 2 \\ 12 & 9 & 17 & 11 & 19 & 14 \\ 13 & 17 & 5 & 15 & 18 & 6 \end{bmatrix}$$

$$M_5 = \frac{1}{90}\cdot\begin{bmatrix} 10 & 8 & 14 & 7 & 12 & 16 \\ 8 & 12 & 8 & 4 & 10 & 3 \\ 8 & 12 & 22 & 14 & 9 & 12 \\ 11 & 21 & 13 & 8 & 19 & 9 \end{bmatrix} \qquad M_6 = \frac{1}{91}\cdot\begin{bmatrix} 11 & 11 & 16 & 3 & 14 & 15 \\ 7 & 14 & 9 & 9 & 10 & 1 \\ 11 & 10 & 22 & 15 & 11 & 13 \\ 13 & 15 & 9 & 9 & 19 & 9 \end{bmatrix}$$

$$M_7 = \frac{1}{88}\cdot\begin{bmatrix} 7 & 8 & 14 & 3 & 13 & 15 \\ 8 & 12 & 10 & 5 & 8 & 1 \\ 8 & 12 & 26 & 14 & 12 & 12 \\ 12 & 21 & 6 & 10 & 20 & 7 \end{bmatrix} \qquad M_8 = \frac{1}{90}\cdot\begin{bmatrix} 6 & 11 & 14 & 5 & 12 & 12 \\ 8 & 14 & 10 & 8 & 10 & 6 \\ 11 & 11 & 20 & 14 & 14 & 12 \\ 13 & 16 & 9 & 10 & 15 & 9 \end{bmatrix}$$

Observing these matrices, we can obtain some common features, which are not easily detected from the DNA primary sequences in Table 1. The triplets $P_1P_1P_1$, $G_2G_2G_2$ and $H_2H_1H_2$ are the most frequent elements of the condensed matrices, while the less frequent triplets are $P_2P_1P_2$, $P_2P_2P_2$, $G_1G_2G_1$, $G_1G_1G_1$ and $H_1H_1H_1$. Some triplets, such as $P_2P_1P_1$, $P_1P_1P_2$, $G_1G_1G_2$, $G_1G_1G_1$, $H_2H_2H_1$, $H_2H_1H_1$, $H_1H_2H_2$ are small variations triplets in the frequency of occurrence, while other elements, for example, $H_2H_2H_2$ etc. show considerable variations. Moreover, we find also that for each condensed matrix, the sum of all elements in the last two rows is greater than one in the first two rows, which means that the values of A+G, T+G and C+G are large, hence, we could conclude G is dominant among all nucleic acid bases for each sequence. These observed results reflect the chemical structure properties implied in exon–1 of β–globin genes. In fact, this phenomenon was also indicated in Table 5 of [11].

3 RESULTS AND DISCUSSION

3.1 Comparative Study of Exon–1 of Different Species

Let M_i and M_j be the condensed matrices of species i and j in Table 1, which appear in the above section. The distance of matrices M_i and M_j is defined as

$$d_{ij} = \left\| M_i - M_j \right\| = \sum_{l=1}^{4}\sum_{m=1}^{6}(a^{(i)}{}_{lm} - a^{(j)}{}_{lm})^2 \quad (i, j = 1, 2, \cdots, 8) \tag{2}$$

where $a^n{}_{lm}$ denotes row l and column m element of matrix M_i, which signifies the frequency of

occurrence of the corresponding (0,1) logic triplet. By using these distance listed in Table 10 as a measure of similarity/dissimilarity, we could investigate the similarities and dissimilarities for the eight exon–1 β–globin genes. The underlying assumption is that the larger the distance, the lesser similar the corresponding DNA sequence. We expected that exon–1 of mouse and exon–1 of rat will be quite similar. From Table 10 we see indeed that the corresponding entry is the smallest number, being 0.113061. However, in [11] the smallest entry is not the entry (mouse, rat), but (human, mouse), (human, rabbit), and (goat, lemur) entries which are much smaller than the entry (mouse, rat), and correspond to species that are not close in the evolutionary tree. In this paper, the magnitudes of numerical values in Table 10 are roughly in proportion to the degrees of similarity. A large entry in such a table certainly points to species that are dissimilar. It is also interesting to observe from Table 10, that gallus (species 4) shows great dissimilarity with other species, because almost all entries belonging to gallus are large. In fact, it is not a mammal, while the other species in Table 10 are mammals.

Table 10. Similarity/Dissimilarity Table for the Eight Exons in Table 1

Species	1. Human	2. Goat	3. Opossum	4. Gallus	5. Lemmur	6. Mouse	7. Rabbit	8. Rat
1 Human	0	0.17295	0.164054	0.172133	0.160247	0.11737	0.0905605	0.142292
2 Goat		0	0.232425	0.21019	0.192083	0.16794	0.19071	0.163029
3 Opossum			0	0.203063	0.176383	0.137324	0.186579	0.169967
4 Gallus				0	0.239341	0.2022	0.208684	0.156347
5 Lemmur					0	0.133222	0.127072	0.149897
6 Mouse						0	0.134051	0.113061
7 Rabbit							0	0.154569
8 Rat								0

3.2 Evolutionary Rates and Gradient Analysis for Various Species

The eigenvalues of a matrix are one of the best–known matrix invariants [13,14]. If a matrix is symmetric, then the eigenvalues are real. A set of eigenvalues can be viewed as a numerical characteristic of a structure. In the previous section, we generated 4×6 matrix M_i, i = 1, 2, …, 8. In order to obtain a symmetric matrix, we define $S_i = M_i \cdot M_i^t$, and denote $R_i = \lambda_{i1}/\lambda_{i4}$, where M_i^t denotes the transpose of M_i, λ_{i1} and λ_{i4} are the maximum and minimum eigenvalue of S_i respectively. With R_i we denote the informational compression ratio of DNA primary sequence. The magnitude of R_i gives expression to the oscillatory degree of elements in matrix S_i, i = 1, 2, …, 8, as will be seen. Using a program encode into Mathematica 4.0, we obtain the symmetric matrices S_i as follows:

$$S_1 = \frac{1}{8100}\begin{bmatrix} 760 & 510 & 924 & 756 \\ 510 & 466 & 708 & 621 \\ 924 & 708 & 1348 & 961 \\ 756 & 621 & 961 & 1114 \end{bmatrix} \qquad S_2 = \frac{1}{7056}\begin{bmatrix} 761 & 444 & 727 & 773 \\ 444 & 351 & 491 & 596 \\ 727 & 491 & 1055 & 899 \\ 773 & 596 & 899 & 1081 \end{bmatrix}$$

$$S_3 = \frac{1}{8100}\begin{bmatrix} 870 & 610 & 808 & 891 \\ 610 & 588 & 614 & 703 \\ 808 & 614 & 942 & 769 \\ 890 & 703 & 769 & 1074 \end{bmatrix} \qquad S_4 = \frac{1}{8100}\begin{bmatrix} 678 & 556 & 872 & 784 \\ 556 & 572 & 703 & 647 \\ 872 & 703 & 1192 & 985 \\ 784 & 647 & 985 & 10681 \end{bmatrix}$$

$$S_5 = \frac{1}{8100}\begin{bmatrix} 809 & 484 & 882 & 888 \\ 484 & 397 & 566 & 693 \\ 882 & 566 & 1113 & 1017 \\ 888 & 693 & 1017 & 1237 \end{bmatrix} \qquad S_6 = \frac{1}{8281}\begin{bmatrix} 948 & 557 & 977 & 880 \\ 557 & 508 & 673 & 662 \\ 977 & 673 & 1220 & 952 \\ 880 & 662 & 952 & 998 \end{bmatrix}$$

$$S_7 = \frac{1}{7744}\begin{bmatrix} 712 & 426 & 894 & 731 \\ 426 & 398 & 646 & 625 \\ 894 & 946 & 1368 & 968 \\ 731 & 625 & 968 & 1170 \end{bmatrix} \qquad S_8 = \frac{1}{8100}\begin{bmatrix} 666 & 574 & 849 & 718 \\ 574 & 560 & 766 & 702 \\ 849 & 766 & 1178 & 957 \\ 718 & 702 & 957 & 912 \end{bmatrix}$$

From the above matrices, $R_1 = 67.1821$, $R_2 = 184.101$, $R_3 = 116.669$, $R_4 = 211.175$, $R_5 = 636.007$, $R_6 = 183.414$, $R_7 = 176.324$, $R_8 = 308.009$, and their ordinal relation is: $R_1 < R_3 < R_7 < R_6 < R_2 < R_4 < R_8 < R_5$.

Table 11. Calculation of the informational compression ratios

k	R_1^k	R_2^k	R_3^k	R_4^k	R_5^k	R_6^k	R_7^k	R_8^k
1	67.18	184.10	111.66	211.17	636.007	183.414	176.324	308.009
4	171.416	627.769	31.9482	118.355	1759.46	139.748	458.466	280.103
8	6431.42	23419.7	160.357	529.672	60795.2	1812.03	29313.1	2011.58
10	48292.2	174678	474.916	1865.72	455983.	8802.88	299019.	6916.08
15	8.71793×10^{6}	3.44121×10^{7}	8692.03	63377.5	9.01129×10^{7}	580424.	1.212×10^{8}	190618.
25	3.42369×10^{11}	2.02269×10^{12}	3.4814×10^{6}	9.41062×10^{7}	5.11373×10^{12}	3.34297×10^{9}	2.57659×10^{13}	2.11115×10^{8}
30	6.91354×10^{13}	5.24597×10^{14}	7.06502×10^{7}	3.68903×10^{9}	1.32015×10^{15}	2.63301×10^{11}	1.226×10^{16}	7.66256×10^{9}

However, we expected that mouse and rats are the most similar, but the corresponding R_6 and R_8 segregate farther, these phenomena demonstrate that the loss of information also perhaps accompanies the condensation of the DNA sequence into the 4×4 matrix S_i. A way to recover some of the lost information associated with the condensation of the DNA sequence to a single 4×4 symmetric matrix S_i is to introduce the k–th power of S_i in which one can generate the closely related matrices S_i^k, obtained from the S_i by raising each element separately to the k–th power for i = 1, 2, …, 8. After several iterative computations, $R_i^j = \lambda_{i1}^j/\lambda_{i4}^j$ is obtained as presented in Table 11, where λ_{i1}^j and λ_{i4}^j denote the maximum and the minimum eigenvalues of the k–th power of S_i.

Obviously, R_3^k and R_4^k are almost the smallest numbers among R_1^k, R_2^k, …, R_8^k, and R_2^k, R_5^k, R_7^k are almost the larger ones. These phenomena make us think of the results of Randić and Vračko in [15] (page 603, paragraph 2): "The leading eigenvalue of D/D matrix, as has been mentioned earlier, give a measure of the degree of folding of long chains. The smaller the value of λ_1, the more folded the corresponding graphical representation of DNA. It follows therefore that among the eight β–globins the opossum β–hemoglobin, sequence C (here is sequence 3), is the most folded. On the other hand, the graphical representation of rabbit β–globins and the goat β–globins are the least

folded of the eight sequences considered". Hence, the quotient of the maximum and minimum eigenvalues can also be regarded as a good index of the degree of folding even for a structure, such as the sequences of DNA considered. Now, we arrange R_1^k, R_2^k, …, R_8^k in term of their magnitude for different k, and list all results in Table 12, then, some features come into light: R_3^k, R_4^k, R_6^k, R_8^k run ahead along with increase of k, while for advanced species (human, goat, lemur and rabbit) the corresponding informational compression ratios are greater than the formers out and away. Opossum, gallus, mouse and rat are always ahead because of their evolutionary relationship of belonging to murine. At the same time we see that opossum and gallus β–globins show greater differences with all other species, including human. The situation has not much changed when k increases. In fact, gallus species has little similarity with the remaining species of Table 1, as is said above. The condensed information suggests that this species would be separated from the rest at an earlier stage of evolutionary development. The former species (opossum) is very similar to gallus, the former belonging to mammal and the later not. This suggests that opossum would also be separated from the rest at an earlier stage too. Moreover, the magnitude of R_3^k, R_4^k, R_6^k and R_8^k first decrease and then enlarge at lower speed by degrees as k grows, while the others ascend at high speed all through. The growth rate of R_4^k is the least, and the growth rates R_1^k and R_7^k is the biggest. The interval between R_6^k and R_8^k become shorten as k grows, the ordinal relation runs to stabilization although individual R_i^k change endlessly. It is obvious that this stable ordinal relation is in accordance with the species' evolutionary gradient on the whole.

Table 12. Recovering information and finding the complexity about time

k	Ordinal relation of R_1^k, R_2^k, R_3^k, R_4^k, R_5^k, R_6^k, R_7^k, R_8^k
1	$R_1 < R_3 < R_7 < R_6 < R_2 < R_4 < R_8 < R_5$
4	$R_3^4 < R_4^4 < R_6^4 < R_1^4 < R_8^4 < R_7^4 < R_2^4 < R_5^4$
8	$R_3^8 < R_4^8 < R_6^8 < R_8^8 < R_1^8 < R_2^8 < R_7^8 < R_5^8$
10	$R_3^{10} < R_4^{10} < R_8^{10} < R_6^{10} < R_1^{10} < R_2^{10} < R_7^{10} < R_5^{10}$
15	$R_3^{15} < R_4^{15} < R_8^{15} < R_6^{15} < R_1^{15} < R_2^{15} < R_5^{15} < R_7^{15}$
25	$R_3^{25} < R_4^{25} < R_8^{25} < R_6^{25} < R_1^{25} < R_2^{25} < R_5^{25} < R_7^{25}$
30	$R_3^{30} < R_4^{30} < R_8^{30} < R_6^{30} < R_1^{30} < R_2^{30} < R_5^{30} < R_7^{30}$

4 CONCLUSIONS

We use an intensive approach, which consider not only sequence structures but also chemical structure for DNA primary sequences. The invariant of sequences is applied to the comparison of DNA primary sequences, rather than sequence themselves. Such scheme extends the matrix methods and improves previously obtained results. The method is suitable for application to whole genes. In addition, we find that the informational compression ratio possibly indicates the evolutionary gradient, which is influenced by base composition and distribution.

Acknowledgment

This work is partially supported by a grant from the National Science Foundation of China. The author would like to thank the reviewers of this paper for several helpful suggestions and an independent check of the numerical results

presented here.

5 REFERENCES

[1] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **1986**, *119*, 319–328.
[2] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Current Sci.* **1994**, *66*, 309–314.
[3] A. Nandy, Graphical representation of long DNA sequences, *Curr.Sci.*, **1994**, *66*, 821.
[4] P. M. Leong and S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Applic. Biosc.* **1995**, *11*, 503–507.
[5] A. Nandy and P. Nandy, Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication, *Current Sci.* **1995**, *68*, 75–85.
[6] X. Guo, M. Randić, and S. C. Basak, A novel 2D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.* **2002**, *350*, 106–112.
[7] Y. Liu, X. Guo, J. Xu, L. Pan, and S. Wang, Some notes on 2–D graphical representation of DNA sequences, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 529–533.
[8] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci.* **1996**, *70*, 661–668.
[9] M. Randić, A. F. Kleiner, L. M. DeAlba, Distance/distance matrices, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277–286.
[10] M. Randić, On characterization of DNA primary sequences by a condensed matrix, *Chem. Phys. Lett.* **2000**, *317*, 29–34.
[11] M. Randić, Condensed Representation of DNA Primary Sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50–56.
[12] P. He and J. Wang, Characteristic sequences for DNA primary sequence, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1080–1085.
[13] Gantmacher. F. *Theory of Matrices*, Chelsea Publishers: New York, 1959; Vol. II, Chapter 13.
[14] M. Randić, X. Guo, T. Oxley, H. Krishnapryan, and L. Naylor, Wiener matrix invariant. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361–367.
[15] M. Randić and M. Vračko, On similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606.

Biographies

Yachun Liu is associate professor of applied mathematics at the Nanhua University of China. His main research direction includes discrete mathematics, computer information processing and some mathematical modeling problem in life science. He was invited to undertake scientific research for advanced visiting scholar by Professor Jin Xu at the Huazhong University of Science and Technology in 2000. He is the author of more than 20 technical publications in algorithmic, complexity, neural network, and molecular computing.