Internet Electronic Journal of Molecular Design **2003**, *2*, 000–000
BioChem Press http://www.biochempress.com

A Simple Method for Characterization and Similarity Analysis of DNA Sequences

Chun Li, [1,2,*] and Jun Wang[2,3]

[1] Department of Mathematics, Bohai University, Jinzhou 121000, Liaoning, P. R. China
[2] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P. R. China
[3] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P. R. China

Internet Electron. J. Mol. Des. **2003**, *1*, 000–000

Abstract

Motivation. DNA sequencing has resulted in an abundance of data on DNA sequences for various species. Hence, characterizations and comparisons of sequences become more important but still difficult tasks. The motivation of this paper is to introduce a simple method to characterize and compare the DNA sequences.

Method. First, based on three classifications of the four nucleic acid bases, we convert a DNA sequence into three binary sequences. Then, we associate each binary sequence with a ${}^{b}E/_{b}G$ matrix by giving a 2-D ladder-like graphical representation, and thus obtain a 3-component vector whose entries are the normalized matrix norms. By the introduced vector, similarity analysis is made among the coding sequences of the exon I of beta-globin gene of eleven species.

Results. As a result, we find that the most similar species are associated with the pairs human-Gorilla, human-Chimpanzee, and Gorilla-Chimpanzee, while gallus shows great dissimilarity with others among the eleven species. This is coincident with the results reported in other literature.

Conclusions. From these findings, the main conclusion we can draw is that the 3-component vector has captured important features of the DNA sequences. In addition, our method is so simple that it can be directly extended to deal with long DNA sequences.

Keywords. DNA; Characteristic sequence; Graphical representation; Matrix; Norm.

1 INTRODUCTION

Biologists need the useful features of DNA sequences, especially the long ones including several thousands or several tens of thousands of bases. However, DNA sequences, as strings of four nucleic acid bases A, C, G and T, do not yield an immediately useful or informative characterization. Comparison of DNA sequences even with bases less than a hundred could be quite difficult [1]. Therefore, many kinds of methods have been proposed to characterize DNA sequences.

Among the existing methods of DNA sequence visualization, 2-D graphical representations play important roles in practice. The 2-D graphical representation of DNA sequences was first proposed by Gates [2], and rediscovered independently by Nandy [3, 4] and Leong and Mogenthaler [5]. However, these 2-D graphical representations of DNA sequences have high degeneracy due to overlapping and crossing of the curve representing DNA with itself [1, 6, 7], although [8] has

* Correspondence author E-mail: lchlmb@yahoo.com.cn

demonstrated that such degeneracies exist only in certain restrictive cases. To reduce the degeneracy, Guo et al. [1, 6] improved the representation by a modification of the directions of vectors assigned to the four bases. Moreover, in the DB-Curve (Dual-Base Curve) introduced by Wu et al. [7] and the `four horizontal lines' graph proposed by Randic et al. [9, 10], the problem of degeneracy is totally avoided because the two kinds of 2-D curves both have a monotonic increasing characteristic.

Matrix representation is another useful method for characterization and comparison of DNA sequences. Unlike the geometric graphical representation, which gives a visual and hence qualitative characterization of DNA sequences, the matrix representation allows a quantitative analysis of data on DNA sequences. Examples of the matrix representations of DNA sequences include the frequency matrix [11-14], condensed matrix [11-13, 15], E matrix, D/D matrix, M/M matrix, S/S matrix, L/L matrix and their `higher order' matrices [9-11] etc.. Once a matrix is given, its leading eigenvalue is always calculated and used as an invariant to describe DNA sequences [10-14].

However, for an $n \times n$ -matrix, the higher is the order n, the more difficult is the calculation of its eigenvalues generally. Are there any other suitable descriptors for the DNA sequences? From the view of matrix theory, for any $n \times n$ (0,1) matrix A = (a_{ij}), there are two facts: (1) its leading eigenvalue λ_1 , an invariant of matrix A, is equal to its spectral radius $\rho(A)$; (2) $\min\{r_1, r_2, \ldots, r_n\} \le \rho(A) \le \max\{r_1, r_2, \ldots, r_n\}$, where r_i is the sum of the i-th row elements of A, that is, $r_i = \sum_{j=1}^{n} a_{ij}$ $(i = 1,2,\ldots,n)$. Therefore, one can use a certain linear combination of the row sums $(r_1, r_2, \ldots, r_n)$ to describe a DNA sequence in stead of the leading eigenvalue λ_1. Moreover, the m1-norm of a matrix A=(a_{ij}) usually is defined as follows:

$$\|A\|_{m1} = \sum_i \sum_j |a_{ij}| \tag{1}$$

Obviously, if A is an $n \times n$ (0,1) square matrix, then

$$\|A\|_{m1} = \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} = \sum_{i=1}^{n} r_i \,, \tag{2}$$

and the normalized m1-norm of A, $\frac{\|A\|_{m1}}{n}$ (or simply $\|A\|_n$ if this does not lead to confusion), is just approximate to the leading eigenvalue. These facts prompt us to describe DNA sequences by the normalized matrix norms. It is easy to see that this method is very simple because the calculation of

the normalized m1-norm is very easy. The utility of our approach is showed with the examination of similarities/dissimilarities among the coding sequences of the exon I of beta-globin gene of 11 different species in Table 1.

Table 1: The coding sequences of the exon I of beta-globin gene of 11 species

Species	Coding sequence	Length
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	92
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG	86
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAG GTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG	92
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAG GTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG	92
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAG GTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG	92
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAG GTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG	92
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAG GTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC	90
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAG GTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG	92
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	93
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	86
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG	105

2 MATERIALS AND METHODS

2.1 Ladder-like graphical representation of the characteristic sequences

In DNA sequences, the four nucleic acid bases A, C, G 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}.

For a given DNA sequence $X = \mathrm{x}_1\mathrm{x}_2\cdots$, one can define three homomorphic maps ϕ_i (i = 1; 2; 3) by $\phi_i(\mathrm{X}) = \phi_i(\mathrm{x}_1)\phi_i(\mathrm{x}_2)\cdots$, where

$$\phi_1(x_i) = \begin{cases} 1 & if\ \mathrm{x_i} \in R \\ 0 & if\ \mathrm{x_i} \in Y \end{cases} \tag{3}$$

$$\phi_2(x_i) = \begin{cases} 1 & if\ \mathrm{x_i} \in M \\ 0 & if\ \mathrm{x_i} \in K \end{cases} \quad (4)$$

$$\phi_3(x_i) = \begin{cases} 1 & if\ \mathrm{x_i} \in W \\ 0 & if\ \mathrm{x_i} \in S \end{cases} \quad (5)$$

Thus, three (0,1)-sequences corresponding to the same DNA sequence can be obtained, which are named the (R, Y)-, (M, K)-, and (W, S)-characteristic sequences of the DNA sequence, respectively [13]. (They are called logic sequences of the DNA sequence in [15].) In Table 2, we present the three characteristic sequences of exon I of the beta-globin gene for human (species 1 in Table 1).

Table 2: the three characteristic sequences of exon I of the beta-globin gene for human

(M, K)-	10000011110011011001 00101100100110001100111000 000011100001 11000010011000000000010011100001 10
(R, Y)-	10110101000110000011 11111110001001001001000010 111101111011 10101110111100110110111100001110 11
(W, S)-	11001001001010100101 00101101010000111010000101 0000011001 01100100110110110010010100000100001 0

The 2-D graph of a characteristic sequence can be constructed as follows: starting from point (0,0), move one unit in the positive x-direction for `base' 1, and along the positive y-direction for `base' 0, we thus obtain a ladder-like curve. From such 2-D ladder-like graphical representation, one can directly find some biological and chemical properties of the DNA sequence considered. For example, the `(W, S)-curve' displays the variation of weak H-bonds W ={A, T} against strong H-bonds S ={G, C}, and would be helpful in visualizing the variation in the G+C content along genes, chromosomes and genomes.

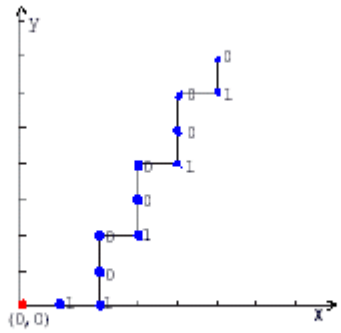

Figure 1: The 2-D ladderlike graphical representation of the sequence 110010010010

Fig.1 shows the 2-D ladder-like graphical representation of the segment consisting of the first 12 `bases' of the (W, S)-characteristic sequence in Table 2.

Obviously, there is only one essential ladder-like curve representing a characteristic sequence. Moreover, as pointed in [13, 14], the three characteristic sequences contain all information of the

DNA sequence. Therefore, a DNA sequence can be represented uniquely by such three 2-D graphs corresponding to the three characteristic sequences.

2.2 Numerical characterization of DNA sequences with 3-component vectors

In order to numerically characterize a DNA sequence denoted by the three 2-D graphs above, we associate a corresponding ladder-like curve with a ${}^{b}E/_{b}G$ matrix, and then calculate its normalized m1-norm. The ${}^{k}E/_{k}G$ matrix is the symmetric matrix whose (i, j) element is defined as follows:

$$\left[{}^{k}E/_{k}G\right]_{ij} = \begin{cases} (\frac{d_{ij}}{\rho_{ij}})^{k} & if\ \mathrm{i} \neq \mathrm{j} \\ 1 & if\ \mathrm{i} = \mathrm{j} \end{cases}, \ (k = 1,2,\ldots), \tag{6}$$

where $d_{ij}(\rho_{ij})$ is the Euclidean (the graph theoretical) distance between vertices i and j on the ladder-like curve. It is easy to see from Fig.1 that for any two vertices i and j, the inequality $0 < d_{ij} \le \rho_{ij}$ always holds, and hence $0 < \left[{}^{k}E/_{k}G\right]_{ij} \le 1$. Clearly, as k approaches infinity the matrix sequences $\left\{{}^{k}E/_{k}G\right\}$ converge to the binary matrix ${}^{b}E/_{b}G$, which is obtained from the E/G matrix by substitution of all elements <1 with zero.

For instance, the ${}^{b}E/_{b}G$ matrix associated with Fig.1 is

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}_{13\times 13} .$$

It normalized m1-norm is 45/13 = 3.461538.

Since each above 2-D graph corresponds to a binary matrix ${}^{b}E/_{b}G$ uniquely, while each DNA sequence can be represented uniquely by such three 2-D graphs, a DNA sequence can be characterized by a 3-component vector in which the normalized matrix norms are individual

components. In Table 3 we list the normalized m1-norms of the ${}^{b}E/_{b}G$ matrices associated with three 2-D ladder-like curves, which correspond to the three characteristic sequences, representing each of the coding sequences of the exon I of beta-globin gene of 11 species in Table 1.

Table 3: The normalized m1-norms of the ${}^{b}E/_{b}G$ matrices associated with three 2-D ladder-like curves for the coding sequences of Table 1

Species	n-(M,K)	n-(R,Y)	n-(W,S)
Human	5.774194	5.258065	4.225806
Goat	4.862069	5.597701	4.402299
Opossum	5.236559	5.494624	3.795699
Gallus	5.000000	5.150538	4.849462
Lemur	5.279570	5.709677	4.096774
Mouse	5.623656	5.559140	4.053763
Rabbit	5.879121	5.549451	4.208791
Rat	5.279570	5.193548	4.569892
Gorilla	5.765957	5.276596	4.234043
Bovine	5.781609	5.689655	4.517241
Chimpanzee	5.905660	5.188679	4.226415

From Table 3, the smallest magnitude of the normalized m1-norm is found for the (W, S)-characteristic sequences. This indicates that the alteration between the weak and strong H-bonds in every coding sequence is relatively frequent. Moreover, in this column, gallus, the only non-mammal among them, is different from other species by the largest value 4.849462. While opossum, the most remote species from the remaining mammals, corresponds to the smallest value 3.795699. This result might imply that the chemical structures and, especially, the hydrogen bond might play special roles in the classification of mammal and non-mammal.

3 RESULTS AND DISCUSSION

3.1 Similarities/dissimilarities among the coding sequences of the exon-I of beta-globin gene of different species

A direct comparison of sequences using computer codes is somewhat less straightforward due to the fact that the sequences have different lengths. If we represent the sequences with the corresponding 3-component vectors then the above question will be avoided totally. In this section, we use these vectors to investigate similarities and dissimilarities for 11 coding sequences of Table 1. The underlying assumption is that if two vectors point to a similar direction in the 3-dimensional space and have similar magnitudes, then the two DNA sequences represented by the two 3-component vectors are similar. Therefore, the similarity between any such two vectors ***a*** and ***b*** can be examined by the formula below:

$$\mathbf{D}\boldsymbol{c} = \mathrm{d}(\boldsymbol{a}, \boldsymbol{b})/\cos(\boldsymbol{a}, \boldsymbol{b}), \tag{7}$$

where d(***a, b***) is the Euclidean distance between the end-points of vectors ***a*** and ***b***, cos(***a, b***) the cosine of the correlation angle of vectors ***a*** and ***b***. Clearly, The smaller the quotient **D*c***, the more similar the two DNA sequences.

In Table 4, the similarities and dissimilarities for 11 coding sequences that based on the quotient **D*c*** of the 3-component vectors are listed.

Table 4: The similarity/dissimilarity matrix for the 11 coding sequences of Table 1 based on the quotient **D*c*** of the 3-component vectors

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.995109	0.729832	1.006175	0.684100	0.378373	0.310248	0.607298	0.021888	0.521164	0.148673
Goat		0	0.722770	0.649065	0.530214	0.841920	1.041939	0.606211	0.979482	0.933750	1.143483
Opossum			0	1.143225	0.372601	0.469874	0.766809	0.835438	0.722786	0.926347	0.855551
Gallus				0	0.984520	1.098560	1.167053	0.398118	0.996636	1.009873	1.108113
Lemur					0	0.378353	0.631772	0.702380	0.667413	0.655902	0.828257
Mouse						0	0.299050	0.722292	0.364424	0.507051	0.497333
Rabbit							0	0.787406	0.296536	0.352771	0.362388
Rat								0	0.598121	0.708479	0.716257
Gorilla									0	0.501350	0.165268
Bovine										0	0.593200
Chimpanzee											0

Observing Table 4, we find that the smallest entries are associated with the pairs: human-Gorilla, human-Chimpanzee, and Gorilla-Chimpanzee. It is also interesting to observe from Table 4, that gallus shows great dissimilarity with others among the eleven species, because almost all entries belonging to gallus are large. This is coincident with the results reported in [10, 12, 13, 15, 16]. On the basis of these findings we conclude that the constructed 3-component vectors have apparently captured important features of the DNA sequences considered.

4 CONCLUDING REMARKS

In this paper, we first represent a DNA sequence by three 2-D ladder-like graphs corresponding to the three characteristic sequences, and then construct a 3-component vector, in which the normalized m1-norms extracted from such three 2-D ladder-like graphs via ${}^{b}E/_{b}G$ matrices are individual components, to characterize and compare the coding sequences. Our method is so simple that it can be directly extended to deal with long DNA sequences. Take the whole human beta-globin gene (see Table 5) as an example, although its length is 1424 much larger than 92 (the length of its exon I), its three corresponding normalized m1-norms are easily calculated as 5.412632,

6.027368 and 5.447719.

Moreover, as we know, the human beta-globin gene can be broken down into three exons and two introns. Their order is: 5'- exon I - intron I - exon II - intron II - exon III - 3', corresponding to the segments 1-92 (92 bases), 93-222 (130 bases), 223-445 (223 bases), 446-1295 (850 bases) and 1296-1424 (129 bases), respectively. In Table 6 we list the corresponding normalized m1-norms of the five segments of the human beta-globin gene. From it, one can find that: (1) the (M, K)-norms of the introns are roughly smaller than that of the exons, but contrary to this for the (R, Y)- and (W, S)-norms. (2) For exons, the magnitudes is in the order: n-(M,K) > n-(R,Y) > n-(W,S), while for introns n-(M,K), n-(R,Y) and n-(W,S) form the shape of a `^', but there is no clear order between n-(M,K) and n-(W,S). These common features of exons and introns might provide biologists with some useful information on the whole human beta-globin gene. Further, by them, could we discriminate between exon and intron segments and thus identify the protein-coding genes in some genomes?

Table 5: Human beta-globin gene of length 1424[a]

1	ATGGTGCACC	TGACTCCTGA	GGAGAAGTCT	GCCGTTACTG
2	CCCTGTGGGG	CAAGGTGAAC	GTGGATGAAG	TTGGTGGTGA
3	GGCCCTGGGC	AG GTTGGTAT	CAAGGTTACA	AGACAGGTTT
4	AAGGAGACCA	ATAGAAACTG	GGCATGTGGA	GACAGAGAAG
5	ACTCTTGGGT	TTCTGATAGG	CACTGACTCT	CTCTGCCTAT
6	TGGTCTATTT	TCCCACCCTT	AG GCTGCTGG	TGGTCTACCC
7	TTGGACCCAG	AGGTTCTTTG	AGTCCTTTGG	GGATCTGTCC
8	ACTCCTGATG	CTGTTATGGG	CAACCCTAAG	GTGAAGGCTC
9	ATGGCAAGAA	AGTGCTCGGT	GCCTTTAGTG	ATGGCCTGGC
10	TCACCTGGAC	AACCTCAAGG	GCACCTTTGC	CACACTGAGT
11	GAGCTGCACT	GTGACAAGCT	GCACGTGGAT	CCTGAGAACT
12	TCAGG GTGAG	TCTATGGGAC	CCTTGATGTT	TTCTTTCCCC
13	TTCTTTTCTA	TGGTTAAGTT	CATGTCATAG	GAAGGGGAGA
14	AGTAACAGGG	TACAGTTTAG	AATGGGAAAC	AGACGAATGA
15	TTGCATCAGT	GTGGAAGTCT	CAGGATCGTT	TTAGTTTCTT
16	TTATTTGCTG	TTCATAACAA	TTGTTTTCTT	TTGTTTAATT
17	CTTGCTTTCT	TTTTTTTTCT	TCTCCGCAAT	TTTTACTATT
18	ATACTTAATG	CCTTAACATT	GTGTATAACA	AAAGGAAATA
19	TCTCTGAGAT	ACATTAAGTA	ACTTAAAAAA	AAACTTTACA
20	CAGTCTGCCT	AGTACATTAC	TATTTGGAAT	ATATGTGTGC
21	TTATTTGCAT	ATTCATAATC	TCCCTACTTT	ATTTTCTTTT
22	ATTTTTAATT	GATACATAAT	CATTATACAT	ATTTATGGGT
23	TAAAGTGTAA	TGTTTTAATA	TGTGTACACA	TATTGACCAA
24	ATCAGGGTAA	TTTTGCATTT	GTAATTTTAA	AAAATGCTTT
25	CTTCTTTTAA	TATACTTTTT	TGTTTATCTT	ATTTCTAATA
26	CTTTCCCTAA	TCTCTTTCTT	TCAGGGCAAT	AATGATACAA
27	TGTATCATGC	CTCTTTGCAC	CATTCTAAAG	AATAACAGTG
28	ATAATTTCTG	GGTTAAGGCA	ATAGCAATAT	TTCTGCATAT
29	AAATATTTCT	GCATATAAAT	TGTAACTGAT	GTAAGAGGTT
30	TCATATTGCT	AATAGCAGCT	ACAATCCAGC	TACCATTCTG
31	CTTTTATTTT	ATGGTTGGGA	TAAGGCTGGA	TTATTCTGAG
32	TCCAAGCTAG	GCCCTTTTGC	TAATCATGTT	CATACCTCTT
33	ATCTTCCTCC	CACAG CTCCT	GGGCAACGTG	CTGGTCTGTG
34	TGCTGGCCCA	TCACTTTGGC	AAAGAATTCA	CCCCACCAGT
35	GCAGGCTGCC	TATCAGAAAG	TGGTGGCTGG	TGTGGCTAAT
36	GCCCTGGCCC	ACAAGTATCA	CTAA	

a: Nucleic acids are grouped in group of tens.

Table 6: The normalized m1-norms of the ${}^{b}E/_{b}G$ matrices associated with three 2-D ladder-like curves for exons and introns of the human *beta*-globin gene

		n-(M,K)	n-(R,Y)	n-(W,S)
exon I	(92)	5.774194	5.258065	4.225806
exon II	(223)	5.267857	5.187500	4.044643
exon III	(129)	6.153846	4.584615	4.446154
average		5.731966	5.010060	4.238868
intron I	(130)	5.396947	5.946565	4.343511
intron II	(850)	5.249119	6.525264	6.238543
average		5.323033	6.235915	5.291027

Acknowledgment

This work was supported in part by the Young Foundation of Bohai University.

5 REFERENCES

[1] X. Guo, M. Randic, S.C. Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.* **2001,** 350, 106.

[2] M.A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **1986**, 119,319.

[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **1994**, 66, 309.

[4] A. Nandy, Graphical representation of long DNA sequences, *Curr. Sci*. **1994**, 66, 821.

[5] P.M. Leong, S. Mogenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci*. **1995**, 12, 503.

[6] X. Guo, A. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, *Chem. Phys. Lett*. **2003**, 369, 361.

[7] Y. Wu, A.W. Liew, H. Yan, M. Yang, DB-Curve: a novel 2D method of DNA sequence visualization and representation*, Chem. Phys. Lett*. **2003,** 367, 170.

[8] A. Nandy, P. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, Chem. Phys. Lett., 2003, 368, 102.

[9] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **2003**, 368, 1.

[10] M. Randic, M. Vracko, N. Lers, D. Plavsic, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **2003,** 371, 202.

[11] M. Randic, On characterization of DNA primary sequences by a condensed matrix, *Chem. Phys. Lett*. **2000**, 317, 29.

[12] M. Randic, X. Guo, S.C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **2001,** 41, 619.

[13] P. He, J. Wang, Characteristic sequences of DNA primary sequence, *J. Chem. Inf. Comput. Sci.* **2002,** 42, 1080.

[14] P. He, J. Wang, Numerical characterization of DNA primary sequence, *Internet Electron. J. Mol. Des.* **2002,** 1, 668. http://www.biochempress.com

[15] Y. Liu, The numerical characterization and similarity analysis of DNA primary sequences, *Internet Electron. J. Mol. Des.* **2002**, 1, 675. http://www.biochempress.com

[16] C. Li, J. Wang, Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences, *Comb. Chem. High T. Scr.,* (accepted).

Biographies

Chun Li is a PhD student of Applied Mathematics at the Dalian University of Technology. His main research interests include combinatorics, information theory and bioinformatics.

Jun Wang is a Professor of Applied Mathematics at the Dalian University of Technology, the advisor of the first author.