Analysis of Similarity/Dissimilarity of DNA Sequences Based on a Class of 2D Graphical Representation

YU-HUA YAO,[1] QI DAI,[2] XU-YING NAN,[1] PING-AN HE,[3] ZUO-MING NIE,[1] SONG-PING ZHOU,[3] YAO-ZHOU ZHANG[1]
[1]*College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China*
[2]*Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China*
[3]*College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China*

Received 12 June 2007; Revised 4 November 2007; Accepted 24 December 2007
DOI 10.1002/jcc.20922
Published online 21 February 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: On the basis of a class of 2D graphical representations of DNA sequences, sensitivity analysis has been performed, showing the high-capability of the proposed representations to take into account small modifications of the DNA sequences. And sensitivity analysis also indicates that the absolute differences of the leading eigenvalues of the L/L matrices associated with DNA increase with the increase of the number of the base mutations. Besides, we conclude that the similarity analysis method based on the correlation angles can better eliminate the effects of the lengths of DNA sequences if compared with the method using the Euclidean distances. As application, the examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of different species has been performed by our method, and the reasonable results verify the validity of our method.

 J Comput Chem 29: 1632–1639, 2008

Key words: DNA; similarity; sensitivity analysis; graphical representation

Introduction

Recently, graphical techniques have emerged as a very powerful tool for the visualization and analysis of long DNA sequences.[1–23] Graphical representation of DNA sequence provides useful insights into local and global characteristics and the occurrences, variations, and repetition of the nucleotides along a sequence, which are not as easily obtainable by other methods.[1] Several authors outlined different 2D graphical representation of DNA sequences based on 2D Cartesian coordinates. The original plot of a DNA sequence as a random walk on a 2D grid by using the four cardinal directions to represent the four bases in DNA sequences was done by Gates,[2] Nandy,[3] and Leong and Mogenthaler.[4] Whereas, such each representation is accompanied by (1) some loss of visual information associated with crossing and overlapping of the resulting curve by itself; (2) an arbitrary decision with respect to the choice of the directions for the four bases; and (3) based on the graphical representation, there is not an effective mathematical evaluation scheme and a very useful classification method so far. To eliminate, or at least reduce the degeneracy of the above 2D rectangular walk methods, this graphical representation technique was modified. Guo et al. allowed the four unit vectors that represent the corresponding bases to be at a small angle to the four axial directions and showed that this reduced the degeneracy[12]; however, it was observed later by Guo and Nandy that such a prescription always in some circumstances exist accidental degeneracy.[13]

A different way to devise 2D graphical representation was proposed by Randić et al., which did not involve the Cartesian coordinate system.[7] In the letter,[15] we propose the concepts of cell and system of graphical representation of DNA sequences, and subsequently, we introduce a class of 2D graphical representations based on different cell designs. The method can completely avoids loss of information accompanying alternative 2D and 3D representations in which the curve standing for DNA overlaps and intersects itself. Several graphical representations introduced by Randić et al.,[7] Liao and Wang,[18] and Song and Tang,[21] could essentially ascribed to this graphical methodology.

In recent years, based on existing graphical representation, several authors have presented various methods to assign mathematical descriptors to DNA sequences in order to quantitatively compare the sequences and determine similarities/dissimilarities between them. In particular, the leading eigenvalues of the *L*/*L* matrices have been considered to be good descriptors of DNA sequences. Following the initial paper of Randić et al., the leading eigenvalues of the *L*/*L* matrices have been used widely in the similarity analysis of the biological (DNA, RNA, or protein)

Correspondence to: Y.-H. Yao; e-mail: yaoyuhua2288@163.com

Table 1. The Coding Sequences of the First Exon of β-Globin Gene of 11 Different Species.

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAG GTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAA GGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAG GTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAG GTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAG GTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAG GTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

sequences and their secondary structures. So a further study for it is necessary and significative. We tried to find out that the leading eigenvalues of *L*/*L* matrices actually reflect some characterizations of biological sequences by means of random simulation. We perform several simulation tests, discuss test results, and find out that the absolute differences of the leading eigenvalues of the *L*/*L* matrices associated with DNA increase with the increase of the number of the base mutations. Our test favors the proposal that the leading eigenvalues of the *L*/*L* matrices can be regarded as an indicator of DNA sequences.

So many authors have offered the methods of similarity analysis of DNA sequences based on different graphical representations.[7–11,14–26] Main idea is to constructs an x-component vector whose components are the leading eigenvalues of the *L*/*L* matrices associated with a DNA sequence. In general, there are two alternative approaches used for "measure" of the degree of similarity/dissimilarity among DNA sequences: (a) approach based on Euclidean distance of vectors characterizing DNA sequences and (b) approach using the correlation angle of two vectors characterizing DNA sequences. We claim that the method (b) is not sensitive on N, the length of DNA sequences, while the method (a) appears to be influenced by length of the sequence.

Besides, we analyzed the sensitivity of the proposed graphical representations. As application, we would make a comparison for the first exon of β-globin genes belonging to 11 different species in Table 1 based on this kind of graphical representations.

Several 2D Graphical Representations of DNA Sequences Based on the Concepts of System and Cell

In the paper,[15] we introduced the graphical method called cell and system. The cell design was viewed as a unit square in which the four points in the corners are designated as the four bases A, T, C, and G, see Figure 1a. All cells are arranged in horizontal direction, and the adjacent cells are separated by one unit. The corresponding system gotten by this way is showed in Figure 1b, we call it system I. The x-coordinate of the base in the unit cell is obtained by finding which column the individual base is in. By labeling the first column as zero, the even columns are found by the formula $(2(i - 1))$ and the odd columns are found by $(2(i - 1) + 1)$ where i is the base number. Then the y-coordinate is found by whether the base is in the first row or the second row of the cell. In summary, the following designations are given to each base: $(2(i - 1), 0) \rightarrow$ G, $(2(i - 1), 1) \rightarrow$ A, $(2(i - 1) + 1, 0) \rightarrow$ C, and $(2(i - 1) + 1, 1) \rightarrow$ T where i is the position of the base in the sequence. The novel 2D graphical representation of DNA sequence is illustrated in Figure 1c on the segment of DNA consisting of the first 10 bases, ATGGTGCACC, of the coding sequence of the first exon of human β-globin gene.

We called the corresponding plot set be characteristic plot set. The curve connected all plots of the characteristic plot set in turn is called characteristic curve.

Figure 2 II to VI are another kind of 2D systems based on different cell design, Figure 2 VII to IX are the 2D systems corresponding to these graphical representations introduced by Liao and Wang,[18] Song and Tang,[21] and Randić et al.,[7] respectively, the left is corresponding cell designs.

Sensitivity Analysis

To check the sensitivity of the proposed approach, three different simulation tests have been studied:

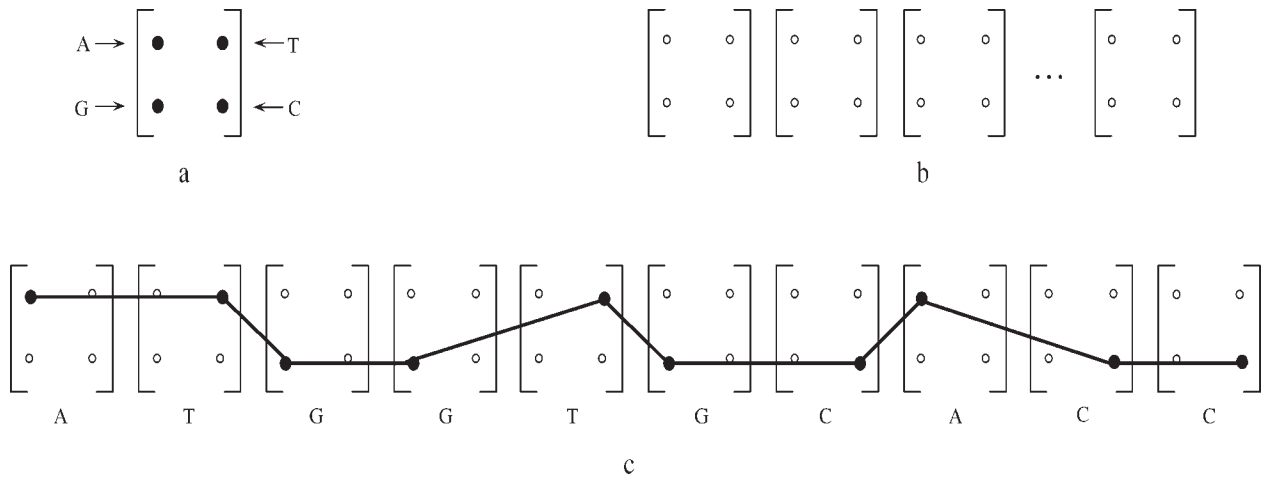

Figure 1. The 2D graphical representation based on the system and the cell of the sequence ATGGTGCACC. (a) Cell, (b) system, (c) zigzag like curve of the sequence ATGGTGCACC. The dots denote the bases making up the sequence.

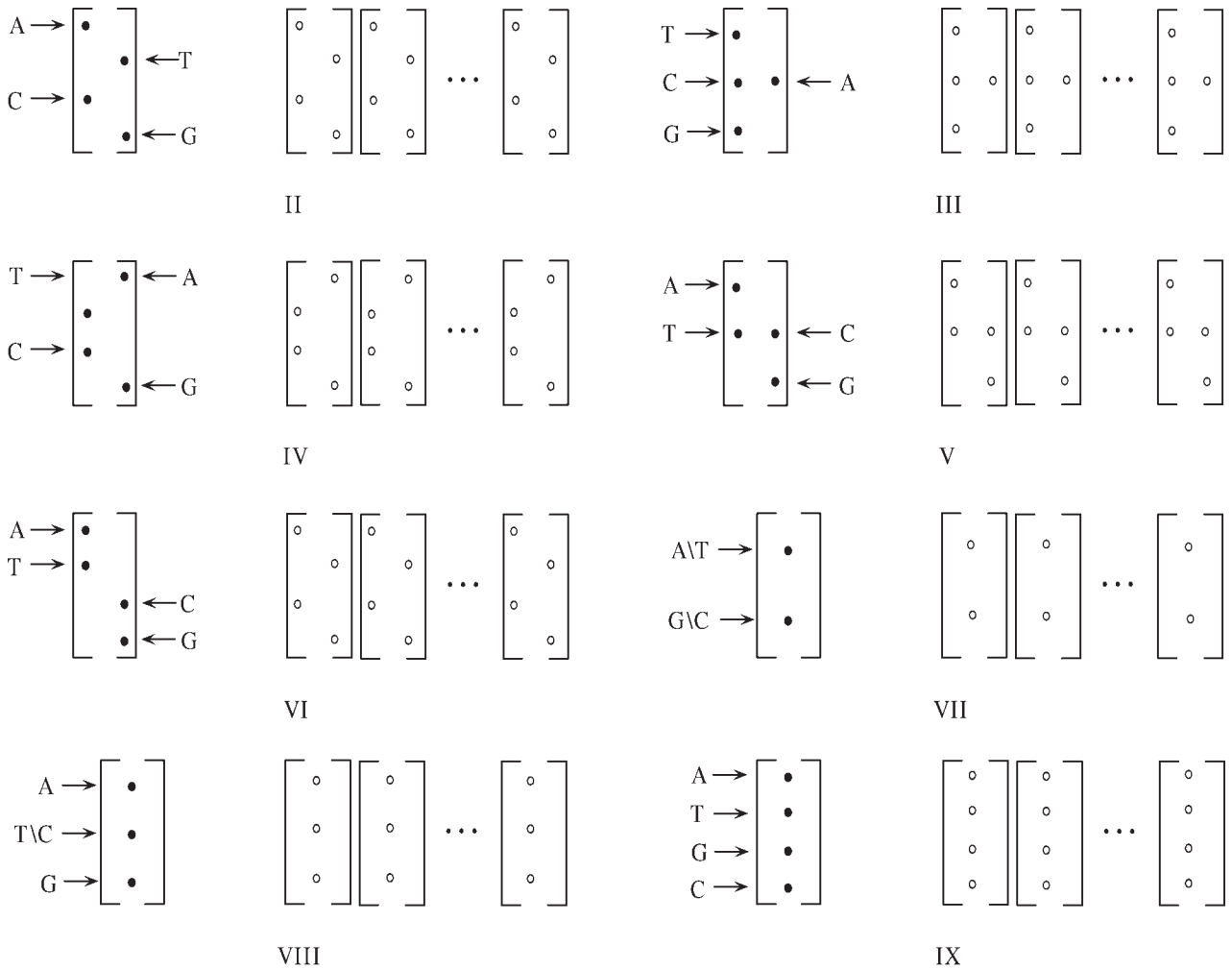

Figure 2. A kind of 2D systems based on different cell designs. The left is the cell designs, the right is the corresponding systems.

Case 1: a base change simulation. The human beta-globin has been considered as the reference sequence (Seq. C) and three other sequences have been artificially produced, changing only position 10 (C in human beta-globin) with G, T, and A, (Seq. G, Seq. T, and Seq. A, respectively). This means that only one base has been changed over a sequence of 92 bases.[27]

We can calculate the similarity between two DNA sequences from the distance of the leading eigenvalues of their *L*/*L* matrices. The similarities among Seq. C, Seq. G, Seq. T, and Seq. A based on different systems are collected in Table 2. Table 2 indicates that Seq. T and Seq. C are more similar in every system. It is worth to note that the first 15 bases of the coding sequence of human beta-globin are ATGGTGCACCTGACT. The changed base locates at the position 10, and the adjacent bases are C and T, which can explain the reason why Seq. T and Seq. C are more similar. We think that the cost of base changing with the adjacent bases is smaller. Take a closer look at the Table 2, we find out that the DNA curve of the coding sequence of human beta-globin does not change in Song's system[21] when the base A change with T and T change with A; the above DNA curve does not change in Liao's system[18] when the base A change with G, T change with C, G change with A, and C change with T. That is to say: S_{Song} (Seq. T, Seq. A) = 0, S_{Liao} (Seq. G, Seq. A) = 0, S_{Liao} (Seq. T, Seq. C) = 0. So, Liao's and Song's systems are less perfect than other systems with regard to the systems' sensitivity.

Case 2: changed base's position simulation. The human beta-globin has still been considered as the reference sequence, denoted as (mod. 0); five other sequences have been artificially generated with one modification in sequence, model 10 (at position 10) with respect to the reference sequence. Then, other four modifications with respect to the reference sequence have been performed at positions 30, 50, 70, and 90. In other words, the sequences mod. 10, mod. 30, . . . , mod. 90 have a base changed at position 10, 30, 50, 70, and 90, respectively. Following the same method, the sequences mod. 6, mod. 26, mod. 46, mod. 66, and mod. 86. can be gotten.

To compare the effect of the modifications at different position, we define average value of similarities, denoted as AV for convenience, of mod. Z as follows:

$$\mathrm{AV}^{Z} = \sum_{XY} \mathrm{Similarity}(\text{mod. } Z(X), \text{mod. } Z(Y))/12,$$

where, mod. Z(X) and mod. Z(Y) denote the new sequences obtained by changing the reference sequence only position Z with X and Y; respectively; $Z \in \{6,10,26,30,46,50,66,70,86,90\}$; $X,Y \in \{A,C,G,T\}$. For example, AV^{10} in system II is 2(0.144 + 0.003 + 0.131 + 0.141 + 0.013 + 0.128)/12 = 0.0933. The average similarities of mod.Z based on different systems are listed in Table 3. The comparison of AV among different mod.Z obtained by the different systems is shown in Figure 3. From the construction of 2D graphical representation of DNA sequences, we know that the effect of changed base's position is symmetrical. Table 3 and Figure 3 verify this proposition. Table 3 and Figure 3 also indicate that the effect is obvious if the changed base's position is in the middle of sequences, it weakens as the changed base's position moves to two sites of sequences. This regular phenomenon can be observed in every system except for system I, Liao's, Song's, and Randić's systems.

Case 3: all-around simulation. In this case, we modify the reference sequences, human beta-globin, considering the number of mutated bases (NMB) and their position. The bases' mutation includes insertion, deletion, and substitution. Here, we assume that the 12 kinds of substitutions {A → T, A → G, . . .} are equiprobable and the bases' mutation positions are also equiprobable. Insertion/Deletion rate is 100/40, it indicates one insertion every 100 bases mutations and one deletion every 40 bases mutations. We focus six kinds of modifications whose NMB are 1, 2, 5, 10, 15, and 20, respectively. For every kind of modifica-

Table 2. The Similarities Among Seq. C, Seq. G, Seq. T, and Seq. A Based on Different Systems.

	Seq. C	Seq. G	Seq. T	Seq. A	Seq. C	Seq. G	Seq. T	Seq. A	Seq. C	Seq. G	Seq. T	Seq. A
		Sys I				Sys II				Sys III		
Seq. C	0	0.023	0.006	0.037	0	0.144	0.003	0.131	0	0.200	0.008	0.050
Seq. G	0.023	0	0.017	0.060	0.144	0	0.141	0.013	0.200	0	0.208	0.150
Seq. T	0.006	0.017	0	0.043	0.003	0.141	0	0.128	0.008	0.208	0	0.058
Seq. A	0.037	0.060	0.043	0	0.131	0.013	0.128	0	0.050	0.150	0.058	0
		Sys VI				Sys V				Sys VI		
Seq. C	0	0.204	0.002	0.133	0	0.156	0.003	0.150	0	0.179	0.010	0.162
Seq. G	0.204	0	0.206	0.071	0.156	0	0.159	0.006	0.179	0	0.188	0.017
Seq. T	0.002	0.206	0	0.135	0.003	0.159	0	0.153	0.010	0.188	0	0.171
Seq. A	0.133	0.071	0.135	0	0.150	0.006	0.153	0	0.162	0.017	0.171	0
		Sys VII				Sys VIII				Sys XI		
Seq. C	0	0.364	0	0.364	0	0.363	0.001	0.001	0	0.067	0.012	0.180
Seq. G	0.364	0	0.364	0	0.363	0	0.364	0.364	0.067	0	0.056	0.248
Seq. T	0	0.364	0	0.364	0.001	0.364	0	0	0.012	0.056	0	0.192
Seq. A	0.364	0	0.364	0	0.001	0.364	0	0	0.180	0.248	0.192	0

Table 3. The Average Similarities of mod. Z Based on Different Systems.

AV	mod. 10	mod. 30	mod. 50	mod. 70	mod. 90
Sys. I	0.0308	0.2306	0.1554	0.1265	0.0096
Sys. II	0.0933	0.3965	0.5895	0.3765	0.0495
Sys. III	0.1122	0.2103	0.2261	0.2038	0.0394
Sys. IV	0.1253	0.4087	0.5678	0.4791	0.0294
Sys. V	0.1043	0.2001	0.3279	0.1850	0.0406
Sys. VI	0.1210	0.4463	0.5937	0.4265	0.0428
Sys. VII	0.2430	0.0013	0.0026	0.0040	0.0893
Sys. VIII	0.1825	0.1589	0.2016	0.3883	0.0890
Sys. XI	0.1258	0.7057	0.8939	0.8864	0.0778
	mod. 6	mod. 26	mod. 46	mod. 66	mod. 86
Sys. I	0.0197	0.0624	0.2197	0.1358	0.0959
Sys. II	0.0660	0.0518	0.8101	0.3725	0.1671
Sys. III	0.0698	0.2164	0.4402	0.2154	0.0785
Sys. IV	0.0908	0.1039	0.7802	0.4688	0.1680
Sys. V	0.0433	0.0577	0.5116	0.1913	0.0814
Sys. VI	0.0517	0.0820	0.8294	0.4186	0.1957
Sys. VII	0.1652	0.4686	0.5376	0.0001	0.0124
Sys. VIII	0.1492	0.3517	0.9033	0.3861	0.0773
Sys. XI	0.0925	0.3016	0.6930	0.8322	0.4098

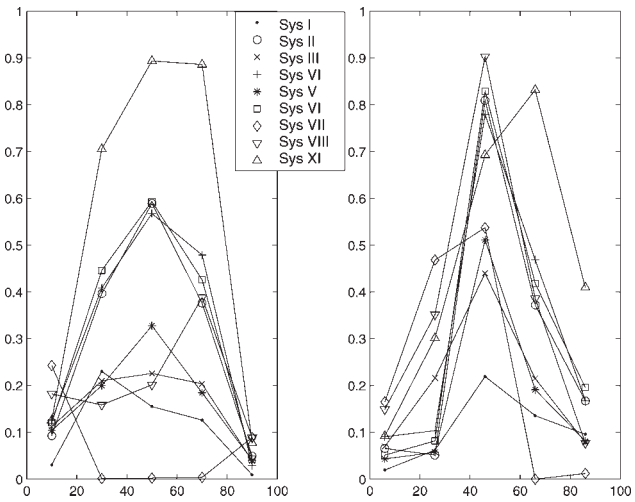

Figure 3. Comparison of the AV among different mod. Z obtained by the different systems. The left is the comparison of mod. 10, mod. 30, mod. 50, mod. 70, and mod. 90 based on the different systems; the right is the comparison of mod. 6, mod. 26, mod. 46, mod. 66, and mod. 86 based on the different systems.

tions, we perform 1000 simulation tests and calculate the averages of absolute differences between the leading eigenvalues of the modified sequences' *L*/*L* matrices and that of reference sequence's *L*/*L* matrix, i.e.

$$A_{\mathrm{NMB}=j} = \frac{1}{1000}\sum_{i=1}^{1000} \mathrm{abs}(\lambda^i_{\mathrm{NMB}=j} - \lambda_{\mathrm{Human}}) \quad j = 1, 2, 5, 10, 15, 20,$$

which are listed in Table 4, where the $\lambda^i_{\mathrm{NMB}=j}$ represent the leading eigenvalues of *i*-th random simulation sequence obtained *j* bases mutation for Human beta-globin. Table 4 shows that the averages of absolute differences of the leading eigenvalues of the *L*/*L* matrices associated with DNA increase with the increase of the NMB. Our test favors the proposal that the leading eigenvalues of the *L*/*L* matrices can be regarded as an indicator of DNA sequences.

Application

Comparison of Two Similarity Measures of DNA Sequences

The general methods introduced by Randić[6–12] are that the DNA sequences are transformed to matrix, and the invariants of the matrix are regarded as the descriptors of the sequence. The normalized leading eigenvalue of *L*/*L* matrix is generally used as the descriptor.[6–10,12–22]

Give two arbitrary DNA sequences, $S^1 = s^1_1 s^1_2 \cdots s^1_{N_1}$, $S^2 = s^2_1 s^2_2 \cdots s^2_{N_2}$, whose lengths are N_i and N_j, respectively. In the graphical approaches where more than one graph was indicated to completely represent a sequence, a set of leading eigenvalues was generated. $(\lambda^1_1, \lambda^1_2, \ldots, \lambda^1_k)$ and $(\lambda^2_1, \lambda^2_2, \ldots, \lambda^2_k)$ are the respective *k*-dimensions vectors composed of the leading eigenvalues of characteristic curves of *k* different patterns of the sequence S^1 and S^2. And then, the similarities of sequence S^1 and S^2 can be computed in two ways: (1) calculate the Euclidean distance between two vectors; (2) calculate the correlation angle of two vectors. The smaller is the Euclidean distance between the end points of two vectors, the more similar are the DNA sequences. On the other hand, the smaller is the correlation angle between two vectors, the more similar are the DNA sequences. The Euclidian distance $D(S^1, S^2)$ between the two vectors is

$$D(S^1, S^2) = \sqrt{\sum_{i=1}^{k}\left(\frac{\lambda^1_i}{N_1} - \frac{\lambda^2_i}{N_2}\right)^2} \qquad (1)$$

Table 4. The Averages of Absolute Differences Between the Leading Eigenvalues of the Modified Sequences' *L*/*L* Matrices and that of Reference Sequence's *L*/*L* Matrix.

NMB	1	2	5	10	15	20
Sys. I	0.1371	0.2238	0.3842	0.6031	0.7699	0.8881
Sys. II	0.2877	0.4492	0.7302	1.0184	1.2879	1.4234
Sys. III	0.2234	0.3489	0.5586	0.8422	1.0702	1.2130
Sys. VI	0.2866	0.4417	0.7148	0.9889	1.2469	1.3873
Sys. V	0.2054	0.3352	0.5602	0.8008	0.9810	1.0870
Sys. VI	0.2962	0.4566	0.7402	1.0238	1.2964	1.4359
Liao	0.2148	0.3831	0.6868	0.9918	1.1931	1.3306
Song	0.3518	0.5604	0.8641	1.1941	1.4336	1.6244
Randić	0.4501	0.6502	1.0205	1.3556	1.7296	2.0324

Table 5. The Similarity/Dissimilarity Matrix for the Coding Sequences of Table 1 Based on the Correlation Angles Between the 12-Component Vectors of the Leading Eigenvalues of the *L*/*L* Matrices for the System I.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.00556	0.00547	0.00721	0.00603	0.00662	0.00452	0.00464	0.00162	0.00319	0.00143
Goat		0.00592	0.00417	0.00283	0.00451	0.00709	0.00506	0.00591	0.00284	0.00638
Opossum			0.00950	0.00393	0.00379	0.00364	0.00850	0.00440	0.00582	0.00527
Gallus				0.00638	0.00845	0.00935	0.00544	0.00812	0.00427	0.00804
Lemur					0.00336	0.00564	0.00730	0.00573	0.00417	0.00638
Mouse						0.00658	0.00751	0.00587	0.00569	0.00688
Rabbit							0.00857	0.00346	0.00558	0.00356
Rat								0.00568	0.00385	0.00572
Gorilla									0.00389	0.00114
Bovine										0.00398

The correlation angle θ (S^1,S^2) between the two vectors is

$$\theta(S^1,S^2) = \arccos\frac{\sum_{i=1}^{k}\lambda_i^1\lambda_i^2}{\sqrt{\sum_{i=1}^{k}\left(\lambda_i^1\right)^2}\sqrt{\sum_{i=1}^{k}\left(\lambda_i^2\right)^2}} \qquad 0 \le \theta \le \pi/4 \tag{2}$$

In the earlier method, it supposes that the leading eigenvalue of the matrix *L*/*L* are linearly relative to the DNA length, and thus, the probabilistic average of λ/N is approximately constant. However, the assumptions are difficult to be convinced in most cases because in almost all the existed models, the leading eigenvalues of *L*/*L* matrix are nonlinearly relative to the length of the sequences. Thus, the better normalization is λ/N^l, $l > 0$, because,

$$\begin{aligned}\theta(S^1,S^2) &= \arccos\frac{\sum_{i=1}^{k}\lambda_i^1/(N_1)^l\ \lambda_i^2/(N_2)^l}{\sqrt{\sum_{i=1}^{k}\left(\lambda_i^1/(N_1)^l\right)^2}\sqrt{\sum_{i=1}^{k}\left(\lambda_i^2/(N_2)^l\right)^2}} \\ &= \arccos\frac{\sum_{i=1}^{k}\lambda_i^1\lambda_i^2(N_1N_2)^l}{\sqrt{\sum_{i=1}^{k}(\lambda_i^1)^2}/(N_1)^l\sqrt{\sum_{i=1}^{k}(\lambda_i^2)^2}/(N_2)^l} \\ &= \arccos\frac{\sum_{i=1}^{k}\lambda_i^1\lambda_i^2}{\sqrt{\sum_{i=1}^{k}(\lambda_i^1)^2}\sqrt{\sum_{i=1}^{k}(\lambda_i^2)^2}},\end{aligned}$$

we have the proposition: the similarity analysis based on the correlation angles can better eliminate the effects of the lengths of DNA sequences.

Similarities/Dissimilarities Among the Coding Sequences of the First Exon of β-Globin Gene of 11 Species

We take system I as a example, to illustrate the use of the quantitative characterization of DNA sequences with an examination of similarities/dissimilarities among the 11 coding sequences of Table 1. We construct a 12-component vectors consisting of the leading eigenvalues of the *L*/*L* matrix corresponding to DNA sequence. The analysis of similarity/dissimilarity among DNA sequences represented by the 12-component vectors is based on the assumption that two DNA sequences are similar if the corresponding 12-component vectors point to a similar direction in the 12D space and have similar magnitudes. The methods based on the correlation angles can eliminate the effects of the lengths of DNA sequences, since the similarity for DNA sequences be measured by calculating the correlation angles between these vectors. Clearly, the smaller is the correlation angles the more similar are the two DNA sequences.

In Table 5, we give the similarities/dissimilarities for the coding sequences of Table 1 based on the correlation angles between the 12-component vectors of the leading eigenvalues of the *L*/*L* matrices. In Tables 6–10, we listed the similarity/dissim-

Table 6. The Similarity/Dissimilarity Matrix of the System II.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.01548	0.02957	0.02864	0.01223	0.01595	0.01059	0.01137	0.00456	0.01349	0.00433
Goat		0.03087	0.02137	0.01520	0.01795	0.01955	0.02043	0.01528	0.00933	0.01696
Opossum			0.02654	0.02519	0.03387	0.03444	0.03049	0.03150	0.03311	0.03208
Gallus				0.02759	0.03291	0.03621	0.02997	0.03091	0.02645	0.03176
Lemur					0.01290	0.01510	0.01351	0.01348	0.01742	0.01430
Mouse						0.01633	0.01920	0.01539	0.02207	0.01695
Rabbit							0.01904	0.00740	0.01637	0.00730
Rat								0.01481	0.01963	0.01421
Gorilla									0.01241	0.00254
Bovine										0.01332

Table 7. The Similarity/Dissimilarity Matrix of the System III.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.00803	0.01892	0.01767	0.00444	0.00747	0.00532	0.00560	0.00252	0.00846	0.00241
Goat		0.01962	0.01373	0.00977	0.01128	0.01029	0.01232	0.00756	0.00505	0.00848
Opossum			0.01359	0.01646	0.02261	0.02313	0.01807	0.02076	0.02128	0.02080
Gallus				0.01714	0.02019	0.02197	0.01955	0.01890	0.01726	0.01945
Lemur					0.00824	0.00766	0.00458	0.00617	0.01068	0.00591
Mouse						0.00851	0.00934	0.00768	0.01355	0.00802
Rabbit							0.00873	0.00369	0.00909	0.00321
Rat								0.00739	0.01250	0.00679
Gorilla									0.00723	0.00118
Bovine										0.00782

Table 8. The Similarity/Dissimilarity Matrix of the System IV.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.01506	0.02871	0.02759	0.01191	0.01533	0.01036	0.01081	0.00445	0.01317	0.00435
Goat		0.02997	0.02069	0.01478	0.01755	0.01917	0.01964	0.01497	0.00911	0.01657
Opossum			0.02598	0.02432	0.03284	0.03356	0.02958	0.03055	0.03218	0.03126
Gallus				0.02684	0.03206	0.03542	0.02873	0.02990	0.02557	0.03077
Lemur					0.01242	0.01508	0.01276	0.01323	0.01711	0.01423
Mouse						0.01583	0.01848	0.01486	0.02155	0.01642
Rabbit							0.01822	0.00716	0.01600	0.00711
Rat								0.01419	0.01884	0.01372
Gorilla									0.01220	0.00240
Bovine										0.01299

Table 9. The Similarity/Dissimilarity Matrix of the System V.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.00924	0.02091	0.01977	0.00581	0.00894	0.00654	0.00684	0.00287	0.00949	0.00278
Goat		0.02179	0.01523	0.01092	0.01266	0.01172	0.01396	0.00870	0.00568	0.00966
Opossum			0.01568	0.01824	0.02501	0.02548	0.02028	0.02285	0.02355	0.02294
Gallus				0.01902	0.02263	0.02435	0.02196	0.02108	0.01903	0.02163
Lemur					0.00951	0.00869	0.00656	0.00736	0.01193	0.00712
Mouse						0.01025	0.01084	0.00893	0.01522	0.00947
Rabbit							0.01086	0.00488	0.01028	0.00420
Rat								0.00869	0.01412	0.00822
Gorilla									0.00818	0.00145
Bovine										0.00873

Table 10. The Similarity/Dissimilarity Matrix of the System VI.

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.01633	0.02999	0.02916	0.01330	0.01701	0.01088	0.01173	0.00478	0.01359	0.00449
Goat		0.03134	0.02168	0.01535	0.01827	0.02080	0.02066	0.01637	0.00985	0.01823
Opossum			0.02800	0.02535	0.03390	0.03450	0.03140	0.03173	0.03353	0.03247
Gallus				0.02832	0.03373	0.03752	0.02966	0.03174	0.02676	0.03265
Lemur					0.01271	0.01626	0.01441	0.01444	0.01790	0.01567
Mouse						0.01692	0.02054	0.01637	0.02258	0.01817
Rabbit							0.01990	0.00723	0.01720	0.00756
Rat								0.01554	0.01946	0.01488
Gorilla									0.01282	0.00268
Bovine										0.01385

ilarity matrices for the coding sequences of Table 1 based on the correlation angles between the x-component vectors of the leading eigenvalues of the L/L matrices of the system II to VI, respectively. We believe that it is not accidental that the smallest entries in Tables 5–10 are associated with the pairs (Human, Chimpanzee), (Human, Gorilla), and (Gorilla, Chimpanzee), and these results mean that the more similar species pairs are Gorilla-Chimpanzee, Human-Chimpanzee, and Human-Gorilla. On the other hand, the largest entries in the similarity/dissimilarity matrix appear in the rows belonging to opossum (the most remote species from the remaining mammals) and gallus (the only non-mammalian representative).

Conclusion

Comparison between different DNA sequences is a key step in bioinformatics when analyzing similarities of DNA sequences and phylogenetic relationships. Sequence alignment is a most common method used to compare DNA sequences. However, its program is only realized with the aid of dynamic programming, which will be slow due to the large number of computational steps. Therefore, it is encountered with difficulty in computational aspect with regard to large biological databases.

Graphical techniques provide us with a novel alignment-free way to compare different DNA sequences. This article devises a class of 2D graphical representations of DNA sequences based on the cells and systems. They do not require 2D Cartesian coordinates and can completely avoid loss of information. And several traditional graphical representations could essentially be ascribed to this graphical methodology. The sensitive analysis shows the high capability of the proposed representations to take into account small modifications of the DNA sequences. The example applying our method to the coding sequences of the first exon of β-globin gene of different species verifies the validity of the method. Our method offers an effective computational framework for DNA sequences to research community of bioinformatics.

Acknowledgments

We thank the referees for many valuable comments that have improved this manuscript, and appreciate the financial support of this work that was provided by Zhejiang Provincial Natural Science Foundation of China (No. Y607510).

References

1. Nandy, A.; Harle, M.; Basak, S. C. ARKIVOC 2006, ix, 211.
2. Gates, M. A. J Theor Biol 1986, 119, 319.
3. Nandy, A. Curr Sci 1994, 66, 309.
4. Leong, P. M.; Morgenthaler, S. Comput Appl Biosci 1995, 11, 503.
5. Hamori, E.; Ruskin, J. J Biol Chem 1983, 258, 1318.
6. Zhang, R.; Zhang, C. T. J Biomol Struct Dyn 1994, 11, 767.
7. Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Chem Phys Lett 2003, 368, 1.
8. Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Chem Phys Lett 2003, 371, 202.
9. Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. J Chem Inf Comput Sci 2000, 40, 1235.
10. Randić, M. Chem Phys Lett 2000, 317, 29.
11. Randić, M.; Vračko, M. J Chem Inf Comput Sci 2000, 40, 599.
12. Guo, X. F.; Randić, M.; Basak, S. C. Chem Phys Lett 2000, 350, 106.
13. Guo, X. F.; Nandy, A. Chem Phys Lett 2003, 369, 361.
14. Yao, Y. H.; Nan, X. Y.; Wang, T. M. J Comput Chem 2005, 26, 1339.
15. Yao, Y. H.; Wang, T. M. Chem Phys Lett 2004, 398, 318.
16. Yao, Y. H.; Nan, X. Y.; Wang, T. M. Chem Phys Lett 2005, 411, 248.
17. Liao, B.; Ding, K. Q. J Comput Chem 2005, 14, 1519.
18. Liao, B.; Wang, T. M. J Comput Chem 2004, 25, 1364.
19. Liao, B.; Tan, M. S.; Ding, K. Q. Chem Phys Lett 2005, 414, 296.
20. Liao, B.; Zhu, W.; Li, P. C. J Math Chem 2006, 42, 1015.
21. Song, J.; Tang, H. W. J Biochem Biophys Methods 2005, 63, 228.
22. Liao, B.; Liu, Y. S.; Li, R. F.; Zhu, W. Chem Phys Lett 2006, 421, 313.
23. Dai, Q.; Liu, X. Q.; Wang, T. M. J Mol Graph Model 2006, 25, 340.
24. Dai, Q.; Liu, X. Q.; Xiu, Z. L.; Wang, T. M. J Theor Biol 2006, 243, 555.
25. Li, C.; Wang, J. Comb Chem High Throughput Screening 2004, 7, 23.
26. He, P. A.; Wang, J. J Chem Inf Comput Sci 2002, 42, 1080.
27. Todeschini, R.; Consonni, V.; Mauri, A.; Ballabio, D. J Chem Inf Model 2006, 46, 1905.