Journal of Biochemistry and Molecular Biology, Vol. 39, No. 3, May 2006, pp. 292-296

JOURNAL OF Biochemistry and Molecular Biology

2-D Graphical Representation for Characteristic Sequences of DNA and its Application[#]

Chun Li[1,*] and Ji Hu[2]

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

[2]Faculty of Chemistry and Chemical engineering, Bohai University, Jinzhou 121000, P. R. China

Received 3 December 2005, Accepted 14 February 2006

DNA sequencing has resulted in an abundance of data on DNA sequences for various species. Hence, the characterization and comparison of sequences become more important but still difficult tasks. In this paper, we first give a 2-D ladder-like graphical representation for the characteristic sequences of a DNA sequence, and then construct a 3-component vector, in which the normalized ALE-indices extracted from such three 2-D graphs via *D/D* matrices are individual components, to characterize the DNA sequence. The examination of similarities/dissimilarities among sequences of the β-globin genes of different species illustrates the utility of the approach.

Keywords: ALE-index, Characteristic sequence, DNA, Graphical representation, Invariant

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, G, C 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 (Randic *et al*., 2000; Guo *et al*., 2001). The previous approaches for the comparison of DNA sequences are mainly based on the sequence alignment, which considers differences between strings due to deletion-insertion, compression-expansion, and substitution of the string elements. Such approaches, which have been hitherto widely used, are computer intensive. Recently, Randic *et al*., have proposed an alternative approach that is based on characterization of DNA by ordered sets of *invariants* derived from DNA sequence, rather than by a direct comparison of DNA sequences themselves. This is analogous to the use of graph invariants (topological indices) for characterization of molecules rather than use of information on their geometry and types of atoms involved. An important advantage of a characterization of structures (be it molecule or DNA) by invariants, as opposed to use of codes, is the simplicity of the comparison of numerical sequences based on invariants (see Randic 2000; Randic and Vracko 2000; Randic *et al*., 2000, 2001).

To obtain the invariant, we follow the strategy below:

Step 1: Represent a DNA sequence by some mathematical object of fixed geometry, such as graph, or a set of lines;

Step 2: For the selected mathematical object, construct its numerical representation in the form of a matrix or set of matrices;

Step 3: From obtained matrices extract a set of invariants.

In recent years, several graphical representations of DNA sequences based on 2-D and 3-D have been outlined (see Nandy 1994a, 1994b; Randic *et al*., 2000; Guo *et al*., 2001; Randic *et al*., 2003a, 2003b; Guo and Nandy 2003; Wu *et al*., 2003; Li and Wang 2004). The advantage of such representations is that they allow visual inspection of data, not only helping in recognizing major differences among similar DNA sequences, but also in deriving numerical characterization for DNA primary sequences. However, some of these representations are accompanied with loss of information due to overlapping and crossing of the curve representing DNA with itself. In this paper, we will give a 2-D graphical representation for the characteristic sequences of a DNA sequence instead of the DNA sequence itself, which avoids the limitation mentioned above.

For a given graph, one can transform it into another mathematical object, a matrix. Examples of the matrix

[#]Supported in part by the Science Research Project of Educational Department of Liaoning Province (20040013).

*To whom correspondence should be addressed.
Tel: 86-416-3400192
E-mail: lchlmb@yahoo.com.cn

Table 1. The β-globin genes of 10 species

Species	Database	ID	Location	Length (bp)
Human	EMBL	HSHBB	62187-63610	1424
Chimpanzee	EMBL	PTGLB1	4189-5532	1344
Gorilla	EMBL	GGBGLOBIN	4538-5881	1344
Lemur	EMBL	LMHBB	154-1595	1442
Rat	EMBL	RNGLB	310-1505	1196
Goat	EMBL	CHHBBAA	279-1749	1471
Bovine	EMBL	BTGL02	278-1741	1464
Rabbit	EMBL	OCBGLO	277-1419	1143
Opossum	EMBL	DVHBBB	467-2488	2022
Gallus	EMBL	GGGL02	465-1810	1346

Table 2. The three characteristic sequences of exon I of the human β-globin gene

Sequence	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
(M, K)-	10000011110011011001 00101100100110001100111000 000011100001 11000010011000000000001001110000110
(R, Y)-	10110101000110000011 11111110001001001001000010 111101111011 10101110111100110110111110000111011
(W, S)-	11001001001010100101 00101101010000111010000101 0000011001 0110010011011011001001010000001000010

representations of DNA sequences include the ***ED***, ***D/D***, ***M/M***, ***L/L*** matrices, and their 'higher order' matrices (Randic 2000; Randic and Vracko 2000; Randic *et al.*, 2000, 2003a, 2003b; Bajzer *et al.*, 2003). Once an $n \times n$ real symmetric matrix $\boldsymbol{M}$ is given, its leading eigenvalue is often used as invariant of the sequence (Randic 2000; Randic and Vracko 2000; Randic *et al.*, 2000, 2001; He and Wang 2002a, 2002b; Randic *et al.*, 2003a, 2003b; Bajzer *et al.*, 2003; Li and Wang 2003, 2004, 2005; Liao *et al.*, 2005). However, a trouble we must face is that the calculation of the eigenvalue will become more and more difficult with the order of the matrix large. Taking into account this point, here we will use the '**ALE**-index' of a matrix we proposed in 2005 (see Li and Wang 2005) as the invariant. The **ALE**-index is defined as

$$\chi = \chi(M) = \frac{1}{2}\left(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F\right) \tag{1}$$

where $\|\cdot\|_{m1}$ and $\|\cdot\|_F$ are the $m1$- and F-norms of a matrix, respectively. Clearly, the **ALE**-index is very simple for calculation so that it can be directly used to handle long DNA sequences. If desired, one can introduce weighting procedure that will normalize magnitudes of the ALE-indices to reduce variations caused by comparison of matrices of different size. For instance, one can consider instead of χ a normalized **ALE**-index $\chi' = \chi/n$, where n is the length of the sequence and the order of the corresponding matrix as well.

The distribution of this paper is as follows. In Section 2, we first represent a DNA sequence by three 2-D ladder-like graphs corresponding to three characteristic sequences of the DNA sequence. Then we transform the graphs into *D/D* matrices, and construct a 3-component vector whose components are the normalized **ALE**-indices of the *D/D* matrices to characterize the DNA sequence. In Section 3, we show the utility of our approach with an examination of similarities/dissimilarities among the full β-globin genes of 10 different species (see Table 1).

Materials and Methods

Ladder-like graphical representation of the characteristic sequences. 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} (Cornish-Bowden 1985; He and Wang 2002a, 2002b). By labeling the elements of R, M and W by 1, and that of Y, K and S by 0, respectively, He and Wang (2002a) 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 Table 2, we present the three characteristic sequences of exon I of the human β-globin gene.

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, thus a ladder-like curve is obtained. From such a 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 (cf. Fig. 1).

Obviously, the 2-D ladder-like graphical representation isn't

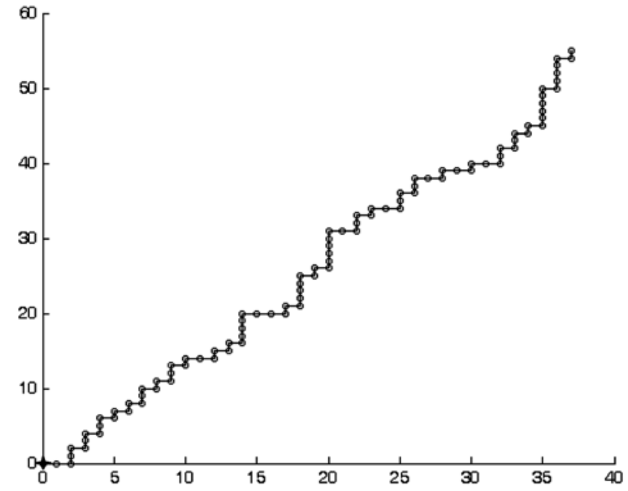

Fig. 1. The 2-D ladder-like graphical representation of the (W,S)-characteristic sequence in Table 2.

accompanied with any loss of information due to overlapping and crossing of the curve with itself. Moreover, as pointed out by He and Wang (He and Wang 2002a, 2002b), 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.

Numerical characterization of DNA sequences with 3-component vectors. In order to numerically characterize a DNA sequence denoted by three 2-D graphs above, we associate each of the three ladder-like curves with a *D/D* matrix, whose (*i*, *j*) element is defined as follows:

$$[D/D]_{ij} = \begin{cases} \dfrac{d_{ij}}{\rho_{ij}} & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i \neq j \end{cases} \tag{2}$$

Table 4. The 3-component vectors for the β-globin genes of 10 species of Table 1

Species	χ'_{MK}	χ'_{RY}	χ'_{WS}
Human	0.714336	0.716069	0.743284
Chimpanzee	0.715139	0.716589	0.745351
Gorilla	0.714511	0.716066	0.746121
Lemur	0.718384	0.715630	0.738181
Rat	0.714246	0.716774	0.722458
Goat	0.711928	0.712017	0.722996
Bovine	0.711483	0.711824	0.724913
Rabbit	0.716898	0.717475	0.727443
Opossum	0.710244	0.727289	0.725744
Gallus	0.712438	0.725014	0.716448

where d_{ij}(ρ_{ij}) is the Euclidean (the graph theoretical) distance between vertices *i* and *j* on the ladder-like curve. As an example, in Table 3 we show a part of the *D/D* matrix corresponding to the curve of Fig. 1.

Observing Table 3, we find that the entries next to the main diagonal in the *D/D* matrix are all equal to 1, while other entries are less than or equal to 1. In fact, for any two vertices v_i and v_j on the ladder-like curve, the inequality $0 < d_{ij} \leq \rho_{ij}$ always holds, and hence $0 \leq [D/D]_{ij} \leq 1$ (cf. Fig.1). Such a matrix has an interesting advantage, that is, from which one can construct a convergent sequence of matrices ${}^kD/{}^kD$, (k = 1, 2, 3, …), whose (*i*, *j*)-element is $[D/D]_{ij}^k$. Clearly, as *k* trends to infinity, the limit of the matrices sequence $\{{}^kD/{}^kD\}$ turns into a (0,1)-matrix, which is denoted by ${}^bD/{}^bD$.

For the three *D*/*D* matrices associated with the three ladder-like curves, one can calculate their normalized ALE-indices, and thus a DNA sequence can be characterized by a 3-component vector with entries being the normalized ALE-indices. In Table 4 we list the 3-component vectors for the β-globin genes of 10 species of Table 1. As can be seen, in the column corresponding to the (W,S)-characteristic sequences, gallus is different from other species by

Table 3. Part of the *D/D* matrix associated with the curve of Fig. 1, constructed from the first 12 bases

	1	1	0	0	1	0	0	1	0	0	1	0	...
1	0	1.0000	0.7071	0.7454	0.7071	0.7211	0.7454	0.7143	0.7289	0.7454	0.7211	0.7329	
1	1.0000	0	1.0000	1.0000	0.7454	0.7906	0.8246	0.7454	0.7693	0.7906	0.7454	0.7616	
0	0.7071	1.0000	0	1.0000	0.7071	0.7454	0.7906	0.7211	0.7454	0.7693	0.7289	0.7454	
0	0.7454	1.0000	1.0000	0	1.0000	0.7071	0.7454	0.7071	0.7211	0.7454	0.7143	0.7289	
1	0.7071	0.7454	0.7071	1.0000	0	1.0000	1.0000	0.7454	0.7906	0.8246	0.7454	0.7693	
0	0.7211	0.7906	0.7454	0.7071	1.0000	0	1.0000	0.7071	0.7454	0.7906	0.7211	0.7454	
0	0.7454	0.8246	0.7906	0.7454	1.0000	1.0000	0	1.0000	0.7071	0.7454	0.7071	0.7211	
1	0.7143	0.7454	0.7211	0.7071	0.7454	0.7071	1.0000	0	1.0000	1.0000	0.7454	0.7906	
0	0.7289	0.7693	0.7454	0.7211	0.7906	0.7454	0.7071	1.0000	0	1.0000	0.7071	0.7454	
0	0.7454	0.7906	0.7693	0.7454	0.8246	0.7906	0.7454	1.0000	1.0000	0	1.0000	0.7071	
1	0.7211	0.7454	0.7289	0.7143	0.7454	0.7211	0.7071	0.7454	0.7071	1.0000	0	1.0000	
0	0.7329	0.7616	0.7454	0.7289	0.7693	0.7454	0.7211	0.7906	0.7454	0.7071	1.0000	0	
⋮													

Table 5. The similarity/dissimilarity matrix for the 10 β-globin genes of Table 1 based on the quotient **D***c* of the 3-component vectors

Species	Human	Chimpanzee	Gorilla	Lemur	Rat	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0.000000	0.002278	0.002842	0.006528	0.020840	0.020830	0.019071	0.016109	0.021223	0.028357
Chimpanzee		0.000000	0.001123	0.007929	0.022914	0.023044	0.021303	0.018017	0.022870	0.030234
Gorilla			0.000000	0.008845	0.023678	0.023620	0.021840	0.018884	0.023655	0.031070
Lemur				0.000000	0.016299	0.016892	0.015432	0.010997	0.018893	0.024412
Rat					0.000000	0.005319	0.006178	0.005690	0.011721	0.010358
Goat						0.000000	0.001977	0.008618	0.015609	0.014563
Bovine							0.000000	0.008225	0.015538	0.015703
Rabbit								0.000000	0.011979	0.014058
Opossum									0.000000	0.009819
Gallus										0.000000

the smallest value 0.716448. The fact that gallus is a non-mammal while all others are mammals in the above table might be a reason for this distinct result.

Results and discussion

In this section, we illustrate the use of the quantitative characterization of DNA sequences with an examination of similarities/dissimilarities among the 10 full β-globin genes of Table 1. The underlying assumption is that if two vectors point to a similar direction in the 3-D space and have similar magnitudes, then the two DNA sequences represented by the two 3-component vectors are similar. The similarity between any such two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ can be examined by the formula below:

$$\mathbf{D}c = d(\boldsymbol{a},\boldsymbol{b})/cos(\boldsymbol{a},\boldsymbol{b}), \quad (3)$$

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

In Table 5, the similarities/dissimilarities for the 10 sequences based on the quotient **D***c* of the 3-component vectors are listed.

Observing Table 5, we find that the most similar species pairs are (gorilla, chimpanzee), (human, gorilla), (human, chimpanzee), and (goat, bovine), which is expected because of their evolutionary relationship. It is also interesting to see that 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). Similar results have been obtained by other authors (Randic *et al.*, 2001, 2003b; He and Wang 2002a; Li and Wang 2003; Liao *et al.*, 2005). It should be mentioned that in another result reported by Liao and Wang (2004), the degree of similarity of human-chimpanzee is obviously lower than that of human and other some species including rat, rabbit and especially gallus. This seems to be a disappointing phenomenon in the evolutionary sense. While from our Table 5, one can easily see that the relationships among human, gorilla, and chimpanzee are closer than that between them and other seven species, and we believe that this is not accidental.

Conclusion

Based on the characteristic sequences of a DNA sequence, we present a similarity measure between DNA sequences. We first give a 2-D ladder-like graphical representation for the characteristic sequences of a DNA sequence, and then construct a 3-component vector, in which the normalized ALE-indices extracted from such three 2-D ladder-like graphs via *D/D* matrices are individual components, to characterize the DNA sequence. It is well known that the alignment of DNA sequences is computer intensive that is a direct comparison of DNA sequences. The structure considered in alignment of DNA sequences is only string's structure. Here, we use an approach that considers not only sequences' structure but also chemical structure for DNA sequences. On the other hand, in comparing with the existing invariant-based approaches for the comparison of DNA sequences, the calculation methodology proposed in this paper is comparatively easy so that it can be directly used to handle long DNA sequences. The examination of the similarity among sequences of the full β-globin genes of 10 species shows the utility of our approach.

References

Bajzer, Z., Randic, M., Plavsic, D. and Basak, S. C. (2003) Novel map descriptors for characterization of toxic effects in proteomics maps. *J. Mol. Graph. Model.* **22**, 1-9.

Cornish-Bowden, A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* **13**, 3021-3030.

Guo, X. F., Randic, M. and Basak S. C. (2001) A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.* **350**, 106-112.

Guo, X. F. and Nandy, A. (2003) Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy. *Chem. Phys. Lett*. **369**, 366.

He, P-an and Wang, J. (2002a) Characteristic sequences for DNA primary sequence. *J. Chem. Inf. Comput. Sci.* **42**, 1080-1085.

He, P-an and Wang, J. (2002b) Numerical characterization of DNA primary sequence. *Internet Electron. J. Mol. Des.* **1**, 668-674.

Li, C. and Wang, J. (2003) Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences. *Comb. Chem. High T. Scr.* **6**, 795.

Li, C. and Wang, J. (2004) On a 3-D representation of DNA primary sequences. *Comb. Chem. High T. Scr.* **7**, 23.

Li, C. and Wang, J. (2005) New Invariant of DNA Sequences. *J. Chem. Inf. Model.* **45**, 115-120.

Liao, B. and Wang, T. (2004) Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett*. **388**, 195-200.

Liao, B., Zhang, Y., Ding, K. and Wang, T. (2005) Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation. *Journal of Molecular Structure: THEOCHEM* **717**, 199-203.

Nandy, A. (1994a) A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Curr. Sci.* **66**, 309-314.

Nandy, A. (1994b) Graphical representation of long DNA sequences. *Curr. Sci*. **66**, 821.

Randic, M., Vracko, M., Nandy, A., Basak, S. C. (2000) On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci*. **40**, 1235-1244.

Randic, M. and Vracko, M. (2000) On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci*. **40**, 599-606.

Randic, M. (2000) On characterization of DNA primary sequences by a condensed matrix. *Chem. Phys. Lett*. **317**, 29-34.

Randic, M., Guo, X. F. and Basak, S. C. (2001) On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inf. Comput. Sci.* **41**, 619-626.

Randic, M., Vracko, M., Lers, N. and Plavsic, D. (2003a) Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **368**, 1-6.

Randic, M., Vracko, M., Lers, N. and Plavsic, D. (2003b) Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **371**, 202-207.

Wu, Y., Liew, A. W., Yan, H. and Yang, M. (2003) DB-Curve: a novel 2D method of DNA sequence visualization and representation. *Chem. Phys. Lett*. **367**, 170-176.