Journal of Mathematical Chemistry, Vol. 40, No. 2, August 2006 (© 2006)
DOI: 10.1007/s10910-006-9063-3

TB-curve, a new 2-D graphical representation of DNA sequences

Ming Ji and Chun Li*

Department of Mathematics, Bohai University, Jinzhou 121000, China
E-mail: lchlmb@yahoo.com.cn

Recieved 27 August 2005; revised 12 December 2005; published online 12 March 2006

Based on the classifications of the four nucleic acid bases, we introduce a new 2-D method of DNA representation: TB-curve, which avoids loss of information accompanying alternative 2-D representation in which the curve standing for DNA overlaps and intersects itself. The method is illustrated on the coding sequence of the first exon of human *beta*-globin gene.

KEY WORDS: DNA, graphical representation, matrix, eigenvalue

1. Introduction

Recently, several authors outlined different 2-D graphical representations of DNA sequences [1–6]. The advantage of graphical representations of DNA sequences is that they provide a simple way of viewing, sorting and comparing various sequences. However, the 2-D graphical representation is accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. Instead of considering the four directions along the cartesian coordinate axes, as the above authors did, Randic et al. [7] proposed a novel 2-D graphical representation of DNA sequences, which avoided the limitation because the curve has a monotonic increasing characteristic. Nevertheless, by this method, there are 12 essentially different patterns of the curves representing the same DNA sequence.

In this paper we introduce a simpler 2-D graphical representation of DNA primary sequences, which also avoids loss of information due to overlapping and crossing of the curve with itself and allows numerical characterization. While there are at most three essentially different curves representing the same DNA sequence.

*Corresponding author.

2. The 2-D graphical representation of DNA

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 $Y = \{C, T\}$.

Let $X = X_1X_2 \cdots X_n$ be a DNA primary sequence with n bases. Define a homomorphic map ϕ_1 by $\phi_1(X) = \phi_1(X_1)\phi_1(X_2) \cdots \phi_1(X_n)$, here

$$\phi_1(X_i) = \begin{cases} (1, R_i) & \text{if } X_i \in R \\ (0, Y_i) & \text{if } X_i \in Y \end{cases} \quad (i = 1, 2, \ldots, n),$$

where $R_i(Y_i)$ is the *cumulative* occurrence numbers of the bases $\in R(Y)$ in the first i bases. Then we map the DNA sequence into a series of nodes P_i's. Connecting adjacent nodes, we obtain a 2-D curve.

Take the segment of DNA consisting of the first 20 bases, ATGGTGCACCTGACTCCTGA, of the first exon of human *beta*-globin gene for an example; its corresponding points are listed in table 1. By connecting these points one bye one, a 2-D curve of this sequence segment is drawn (see figure 1(a)). This curve displays two kinds of bases, the purine R and pyrimidine Y, at a time on a plane.

Table 1
2-D Coordinates for the First 20 Bases of the First Exon of Human *Beta*-globin Gene.

Base	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
nucletic	A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A
x	1	0	1	1	0	1	0	1	0	0	0	1	1	0	0	0	0	0	1	1
y	1	1	2	3	2	4	3	5	4	5	6	6	7	7	8	9	10	11	8	9

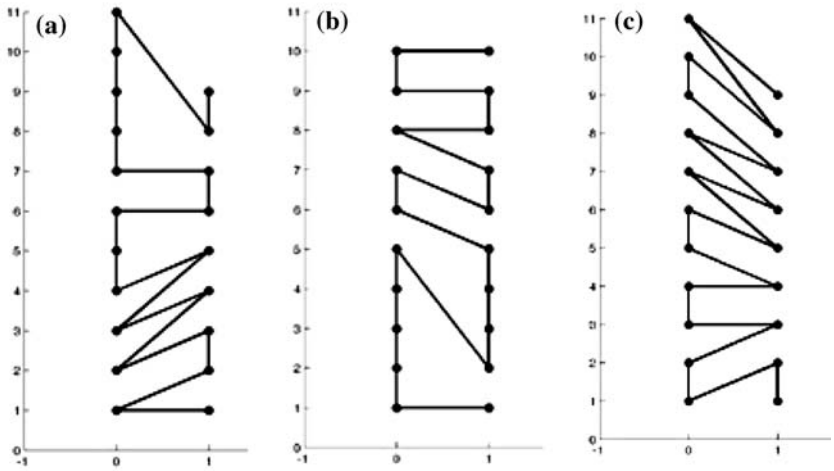

Figure 1. The three TB curves of the sequence ATGGTGCACCTGACTCCTGA. (a) RY-TB curve, (b) MK-TB curve, (c) WS-TB curve.

It is just in this sense we call it as the RY-TB (two kinds of bases) curve of the DNA sequence considered.

It is not difficult to see that the RY-TB curve has the following properties:

(1) The RY-TB curve has no multiple edge or circuit from the point of view of the graph theory, and thus the problem of degeneracy due to overlapping and crossing of the curve with itself is totally avoided.

(2) The RY-TB curve is condensed and thus allows visual inspection of lengthy DNA sequences without requiring excessive space.

(3) The RY-TB curve clearly shows the numbers of the bases $\in R$ and the bases $\in Y$ in the DNA sequence and its segments, and thus the relative abundance of two kinds of bases can be observed directly.

As pointed out in [8–10], the four bases can be also divided into other two 2-classes: amino group $M = \{A, C\}$ and keto group $K = \{G, T\}$; weak H-bonds $W = \{A, T\}$ and strong H-bonds $S = \{G, C\}$. Similarly, we define other two maps $\phi_j (j = 2, 3)$:

$$\phi_2(X_i) = \begin{cases} (1, M_i) & X_i \in M \\ (0, K_i) & X_i \in K \end{cases} \quad (i = 1, 2, \ldots, n),$$

$$\phi_3(X_i) = \begin{cases} (1, W_i) & X_i \in W \\ (0, S_i) & X_i \in S \end{cases} \quad (i = 1, 2, \ldots, n).$$

Then from the same DNA sequence we obtain other two curves: the MK- and WS-TB curves. Figure 1(b) and (c) show the MK- and WS-TB curves of the same sequence ATGGTGCACCTGACTCCTGA. Obviously, the three TB curves give all information of a DNA sequence because the DNA sequence is uniquely determined by its three TB curves.

3. The numerical characterization

In order to find some of the invariants sensitive to the form of the curve we transform the graphical representation of the curve into another mathematical object, a matrix. The matrices that can be constructed from a graphical representation include the Euclidean-distance matrix **ED**, The Graph Theoretical Distance matrix **GD**, the path-distance matrix **PD**, and their quotient matrices **M/M**, and **L/L** [7, 11, 12]:

3.1. The ED matrix

The (i, j)-matrix element is defined to be the Euclidean-distance between vertices i and j of the curve in the 2-D space:

$$[\mathbf{ED}]_{\mathbf{ij}} = \sqrt{(\boldsymbol{x_i} - x_j)^2 + (\boldsymbol{y_i} - \boldsymbol{y_j})^2}.$$

3.2. Path-distance matrix PD

The (i, j)-matrix element is defined to be the sum of the Euclidean-distances between the adjacent vertices along the path from vertex i to vertex j:

$$[\mathrm{PD}]_{ji} = [\mathrm{PD}]_{ij} = [\mathrm{ED}]_{i,i+1} + [\mathrm{ED}]_{i+1,i+2} + \cdots + [\mathrm{ED}]_{j-1,j}, \quad \boldsymbol{i < j}; [\mathrm{PD}]_{ii} = \mathbf{0}.$$

3.3. The M/M matrix

The (i, j)-matrix element is defined to be the quotient of the corresponding elements of the **ED** matrix and the **GD** matrix:

$$[\mathbf{M/M}]_{ij} = [\mathbf{ED}]_{ij}/[\mathbf{GD}]_{ij}, \quad i \neq j;$$
$$[\mathbf{M/M}]_{ii} = 0.$$

3.4. The L/L matrix

The (i, j)-matrix element of L/L is defined to be the quotient of the corresponding elements of the **ED** matrix and the **PD** matrix:

$$[\mathbf{L/L}]_{ij} = [\mathbf{ED}]_{ij}/[\mathbf{PD}]_{ij}, \quad i \neq j;$$
$$[\mathbf{L/L}]_{ii} = 0.$$

Clearly, all these matrices are symmetric. The upper triangle elements of the **ED, M/M** and **L/L** matrices for the same RY-TB curve of the DNA segment ATGGTGCACC are listed in tables 2–4.

The entries next to the main diagonal both in the **ED** and **M/M** matrix are more than or equal to 1, and their upper bound is $\sqrt{(n-1)^2 + 1}$ for $n > 2$, where n is the length of the DNA sequence considered. While in the **L/L** matrix they are always equal to 1. In fact, by the definition of **L/L**, the matrix elements for all adjacent pairs of vertices must equal 1 and for nonadjacent vertices, the matrix elements are less than or equal to 1. In other words, every entry of an

Table 2
The upper triangles of the **ED** matrix for the RY-TB curve of the sequence ATGGTGCACC.

ED	A	T	G	G	T	G	C	A	C	C
A	0.000000	1.000000	1.000000	2.000000	1.414214	3.000000	2.236068	4.000000	3.162278	4.123106
T		0.000000	1.414214	2.236068	1.000000	3.162278	2.000000	4.123106	3.000000	4.000000
G			0.000000	1.000000	1.000000	2.000000	1.414214	3.000000	2.236068	3.162278
G				0.000000	1.414214	1.000000	1.000000	2.000000	1.414214	2.236068
T					0.000000	2.236068	1.000000	3.162278	2.000000	3.000000
G						0.000000	1.414214	1.000000	1.000000	1.414214
C							0.000000	2.236068	1.000000	2.000000
A								0.000000	1.414214	1.000000
C									0.000000	1.000000
C										0.000000

Table 3
The upper triangles of the **M/M** matrix for the RY-TB curve of the sequence ATGGTGCACC.

M/M	A	T	G	G	T	G	C	A	C	C
A	0.000000	1.000000	0.500000	0.666667	0.353553	0.600000	0.372678	0.571429	0.395285	0.458123
T		0.000000	1.414214	1.118034	0.333333	0.790569	0.400000	0.687184	0.428571	0.500000
G			0.000000	1.000000	0.500000	0.666667	0.353553	0.600000	0.372678	0.451754
G				0.000000	1.414214	0.500000	0.333333	0.500000	0.282843	0.372678
T					0.000000	2.236068	0.500000	1.054093	0.500000	0.600000
G						0.000000	1.414214	0.500000	0.333333	0.353553
C							0.000000	2.236068	0.500000	0.666667
A								0.000000	1.414214	0.500000
C									0.000000	1.000000
C										0.000000

Table 4
The upper triangles of the **L/L** matrix for the RY-TB curve of the sequence ATGGTGCACC.

L/L	A	T	G	G	T	G	C	A	C	C
A	0.000000	1.000000	0.414214	0.585786	0.292893	0.424659	0.263727	0.373316	0.260721	0.314046
T		0.000000	1.000000	0.926210	0.261204	0.521441	0.267426	0.424416	0.269566	0.329788
G			0.000000	1.000000	0.414214	0.430081	0.233196	0.361421	0.230172	0.295132
G				0.000000	1.000000	0.273951	0.197453	0.273951	0.162278	0.230172
T					0.000000	1.000000	0.273951	0.537222	0.273951	0.361421
G						0.000000	1.000000	0.273951	0.197453	0.233196
C							0.000000	1.000000	0.273951	0.430081
A								0.000000	1.000000	0.414214
C									0.000000	1.000000
C										0.000000

Table 5
The coding sequence of the first exon of human *beta*-globin gene. The first ten eigenvalues λ_i, and the scaled eigenvalues, ${}^s\lambda_i$, of the **D/D** and **L/L** matrices of the sequence.

	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTG GGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG				
	D/D[a]	**L/L**	Scaled	**D/D**[a]	**L/L**
λ_1	0.27136	28.3424	${}^s\lambda_1$	100	100
λ_2	0.07010	6.9091	${}^s\lambda_2$	25.833	24.377
λ_3	0.06208	5.2627	${}^s\lambda_3$	22.877	18.568
λ_4	0.04361	4.7146	${}^s\lambda_4$	16.071	16.634
λ_5	0.03315	4.2593	${}^s\lambda_5$	12.216	15.028
λ_6	0.03236	3.7643	${}^s\lambda_6$	11.925	13.282
λ_7	0.02842	3.1005	${}^s\lambda_7$	10.473	10.939
λ_8	0.02563	2.5708	${}^s\lambda_8$	9.445	9.071
λ_9	0.02429	2.0984	${}^s\lambda_9$	8.951	7.404
λ_{10}	0.02071	1.7719	${}^s\lambda_{10}$	7.632	6.252

a: Taken from [ref 7, table 3].

L/*L* matrix is in the interval [0,1]. This allows one to construct a convergent sequence of matrices ${}^k\mathbf{L}/{}^k\mathbf{L}$, the product of Hadammard multiplication of the **L/L** matrix by itself k-times, ($k = 1, 2, 3, \ldots$). From these 'higher order' matrices, one can derive additional structurally related descriptors [7,11].

4. Application: the coding sequence of the first exon of human *beta*-globin gene

In this section, we construct the DNA descriptors for the coding sequence of the first exon of human *beta*-globin gene and compare them with the descriptors based on Nandy's graphical representation (eigenvalues of the **D/D** matrix). The sequence consists of 92 bases and is shown in table 5, in which we list 10 eigenvalues of **D/D** and **L/L** matrices. In order to make the comparison of the eigenvalues easier we show at the right-hand side of table 5 the eigenvalues scaled so that the leading eigenvalue for each of the two matrices equals 100. We find that they have similar overall appearance: a single major leading eigenvalue followed by positive and negative eigenvalues which are of relatively small magnitude. In figure 2 we show the plot of the eigenvalues of the **L/L** matrix of table 5 against the corresponding eigenvalues of the **D/D** matrix. Because of great disparities between the leading eigenvalues of **L/L** and **D/D** in comparison with other eigenvalues, respectively, in figure 2 we exclude the leading eigenvalue. The main conclusion one can draw from figure 2 is that: (1) the numerical characterization of the new 2-D graphical representation, being easy to constructed, and the older 2-D representation based on graphical approach of Nandy have

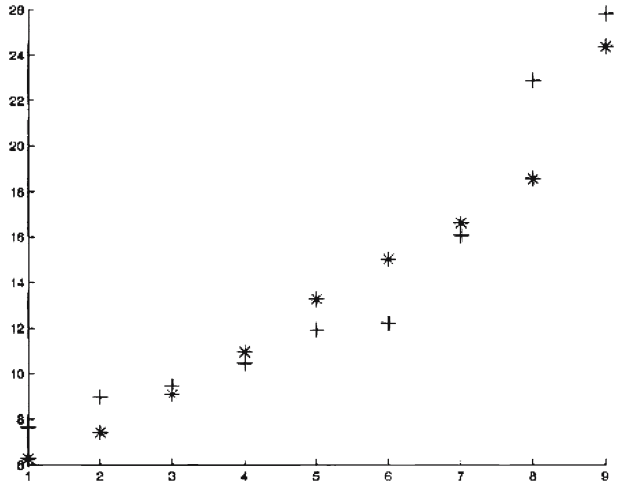

Figure 2. The plot of some eigenvalues of the **L/L** matrix of the sequence of Table 5 **VS** the corresponding eigenvalues of the **D/D** matrix.

similarities and (2) a few leading eigenvalues may suffice to characterize DNA sequences.

Acknowledgment

This work was supported in part by the Science Research Project of Educational Department of Liaoning Province of China (20040013).

References

[1] M.A. Gates, J. Theor. Biol. 119 (1986) 319.
[2] A. Nandy, Curr. Sci. 66 (1994) 309.
[3] A. Nandy, Curr. Sci. 66 (1994) 821.
[4] P.M. Leong and S. Mogenthaler, Comput. Appl. Biosci. 12 (1995) 503.
[5] X.F. Guo, M. Randic and S.C. Basak, Chem. Phys. Lett. 350 (2001) 106.
[6] Y.H. Wu, A.W. Liew, H. Yan and M. Yang, Chem. Phys. Lett. 367 (2003) 170.
[7] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 368 (2003) 1.
[8] P-an He and J. Wang, J. Chem. Inf. Comput. Sci. 42 (2002) 1080.
[9] C.T. Zhang, J. Theor. Biol. 187 (1997) 297.
[10] A. Cornish-Bowden, Nucleic Acids Res. 13 (1985) 3021.
[11] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 371 (2003) 202.
[12] Z. Bajzer, M. Randic, D. Plasic and S.C. Basak, J. Mol. Graph. Model. 22 (2003) 1.