A Novel Way to Numerically Characterize DNA Sequences and Its Application

YING GUO,¹ YAN-FANG WANG,¹ SHENG-LI ZHANG²

¹College of Information Engineering, Dalian University, Dalian of Liaoning 116622, China

²School of Mathematical Sciences, Dalian University of Technology, Dalian of Liaoning, Dalian 116024, China

Received 19 January 2010; accepted 27 May 2010

Published online 4 November 2010 in Wiley Online Library (wileyonlinelibrary.com).

DOI 10.1002/qua.22872

ABSTRACT: We presented a novel way to numerically characterize DNA sequences based on the graphical representation for the sequences comparison and analysis. Instead of calculating the leading eigenvalues of the matrix for graphical representation, we computed curvature and torsion of curves as the descriptor to numerically characterize DNA sequences. The new method was tested on three data sets: the coding sequences of β -globin gene, all of their exons, and 24 coronavirus geneomes from NCBI. The similarities/dissimilarities and phylogenetic tree of these species verify the validity of our method. © 2010 Wiley Periodicals, Inc. *Int J Quantum Chem* 111: 3971–3979, 2011

Key words: graphical representation; curvature; torsion; phylogenetic tree

1. Introduction

Graphical techniques initiated in 1983 by Hamori and Ruskin [1] have emerged as a very powerful tool for the visualization and analysis of long DNA sequences. Several authors outlined different 2D graphical representations of DNA sequences based on two dimensional Cartesian coordinates. The original plot of a DNA sequence as

a random walk on a 2D grid using the four cardinal directions to represent the four bases was done by Gates [2], Leong and Morgenthaler [3], and Nandy [4]. Their method is based on the assignment of the four bases of DNA sequences to the four directions of the (x, y) coordinate system. These 2D graphical representations of DNA sequence provide useful insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence that are not easily observed from DNA sequences directly. However, these graphical representations are accompanied with some loss of information because of overlapping and crossing of the curve representing DNA with itself. To eliminate, or at least reduce the

Correspondence to: Y. Guo; e-mail: guoying197989@yahoo.cn

Contract grant sponsor: China National Natural Science Foundation.

Contract grant number: 10871219.

degeneracy of the above graphical representations, many high orders and unique graphical representations have been proposed [5–11].

In recent years, based on existing graphical representation, several authors have presented various methods to assign mathematical descriptors to DNA sequences to quantitatively compare the sequences and determine similarities and dissimilarities between them [8–10, 12–16]. In particular, the leading eigenvalues of the L/L matrices have been considered to be good descriptors of DNA sequences. However, the computation of the leading eigenvalues of the L/L matrices for long DNA sequences will be expensive. Therefore, the emergence of research into mathematical descriptors of DNA sequences is apparent and necessary.

Motivated by searching an efficient descriptor of DNA sequences, we propose a novel way to numerically characterize DNA sequences. When a DNA sequence is mapped into a 3D space, we can obtain a curve. Then, the curvature and torsion of the curve are computed to numerically characterize DNA sequences. The proposed numerical characterizations are tested by similarity analysis and phylogenetic analysis on three different data sets. Our results show that our method is preferable to numerically characterize DNA sequences. Furthermore, our method is rapid because the whole process does not involve complex algorithm.

2. Materials and Methods

2.1. 3D GRAPHICAL REPRESENTATION

Yuan et al. [7] proposed a 3D graphical representation that assigns one nucleotide base as follows:

$$\begin{aligned} (-1, 0, 0) &\rightarrow A, & (1, 0, 0) &\rightarrow G, \\ (0, -1, 0) &\rightarrow T, & (0, 1, 0) &\rightarrow C. \end{aligned}$$

That is to say, A, G, T, and C are assigned to $-x$, $+x$, $-y$, and $+y$, respectively, while the corresponding curve extend along with z -axes. In detail, for a given DNA sequence $G = g_1 g_2 \dots g_i \dots$, inspect it by stepping one base at a time. For the step i ($i = 1, 2, \dots, N$), a 3D space point $P_i(x_i, y_i, z_i)$ can be constructed by function $\phi(g_i)$ as follows:

$$\phi(g_i) = \begin{cases} (-1, 0, i) & \text{if } g_i = A, \\ (1, 0, i) & \text{if } g_i = G, \\ (0, -1, i) & \text{if } g_i = T, \\ (0, 1, i) & \text{if } g_i = C. \end{cases} \quad (1)$$

FIGURE 1. The 3-D graphical representation of the sequence ATGGTGCACC.

where N is the length of the given DNA sequence. When i runs from 1 to N , we have points $P_1(x_1, y_1, z_1), P_2(x_2, y_2, z_2), \dots, P_N(x_N, y_N, z_N)$. Connecting adjacent points, we obtain a 3D zigzag curve. For example, the 3D graphical representation of the sequence ATGGTGCACC is presented in Figure 1.

According to the method of the graphical representation, there are three curves corresponding to the same DNA sequence. If we assign the four nucleotide bases as follows:

$$\begin{aligned} (-1, 0, 0) &\rightarrow A & (1, 0, 0) &\rightarrow T \\ (0, -1, 0) &\rightarrow C & (0, 1, 0) &\rightarrow G, \end{aligned}$$

we will get the second 3D curve. For the same sequence, ATGGTGCACC, the graph of the second curve is shown in Figure 2.

The third curve will be gotten by assigning the four nucleotide bases as follows:

$$\begin{aligned} (-1, 0, 0) &\rightarrow A & (1, 0, 0) &\rightarrow C \\ (0, -1, 0) &\rightarrow G & (0, 1, 0) &\rightarrow T \end{aligned}$$

After having three curves corresponding to the same DNA sequence, we conveniently denote them as the curves of the patterns AGTC, ATCG, and ACGT.

2.2. THE CURVATURE AND TORSION OF THE CURVE

The most fundamental characteristics of a curve are its curvature and torsion, so we regard the curvature and torsion of curves as the descriptors to

FIGURE 2. The 3-D graphical representation of the sequence ATGGTGCACC.

numerically characterize curve of DNA sequences. The zigzag curve from the graphical representation of Yuan et al. is not smooth. In this section, we introduce a new method to calculate curvature and torsion of unsmooth curves.

Based on the reference [17], let Δ be the difference operator, which assigns to every function $f(x)$, the function $g = \Delta f$, which is defined by $g(x) = f(x+1) - f(x)$. For each integer $n \geq 2$, we define $\Delta^n f = \Delta(\Delta^{n-1}f)$, and we denote $\Delta^n f(x)$ instead of $(\Delta^n f)(x)$. Then, we have:

$$\Delta^n f(x) = \sum_{k=0}^n (-1)^{n-k} \binom{n}{k} f(x+k), \quad n = 0, 1, 2, \dots \quad (2)$$

So we can get the first to third difference form as below:

$$\Delta f(x) = f(x+1) - f(x) \quad (3)$$

$$\Delta^2 f(x) = f(x+2) - 2f(x+1) + f(x) \quad (4)$$

$$\Delta^3 f(x) = f(x+3) - 3f(x+2) + 3f(x+1) - f(x) \quad (5)$$

Then the three curves are obtained, denoted by

$$r^i(t) = (x^i(t), y^i(t), z^i(t)) \quad i = 1, 2, 3 \quad (6)$$

For the i -th curve, its curvature k and torsion τ can be calculated by the following formula [18].

$$k^i(t) = \left(\left| \begin{matrix} \Delta y^i(t) & \Delta z^i(t) \\ \Delta^2 y^i(t) & \Delta^2 z^i(t) \end{matrix} \right|^2 + \left| \begin{matrix} \Delta z^i(t) & \Delta x^i(t) \\ \Delta^2 z^i(t) & \Delta^2 x^i(t) \end{matrix} \right|^2 \right)^{1/2}$$

$$+ \left| \begin{matrix} \Delta x^i(t) & \Delta y^i(t) \\ \Delta^2 x^i(t) & \Delta^2 y^i(t) \end{matrix} \right|^2 \Big)^{1/2} / \{(\Delta x^i(t))^2 + (\Delta y^i(t))^2 + (\Delta z^i(t))^2\}^{3/2} \quad i = 1, 2, 3 \quad (7)$$

$$\tau^i(t) = \frac{\left| \begin{matrix} \Delta x^i(t) & \Delta y^i(t) & \Delta z^i(t) \\ \Delta^2 x^i(t) & \Delta^2 y^i(t) & \Delta^2 z^i(t) \\ \Delta^3 x^i(t) & \Delta^3 y^i(t) & \Delta^3 z^i(t) \end{matrix} \right|}{\left(\left| \begin{matrix} \Delta y^i(t) & \Delta z^i(t) \\ \Delta^2 y^i(t) & \Delta^2 z^i(t) \end{matrix} \right|^2 + \left| \begin{matrix} \Delta z^i(t) & \Delta x^i(t) \\ \Delta^2 z^i(t) & \Delta^2 x^i(t) \end{matrix} \right|^2 + \left| \begin{matrix} \Delta x^i(t) & \Delta y^i(t) \\ \Delta^2 x^i(t) & \Delta^2 y^i(t) \end{matrix} \right|^2 \right)^{1/2}} \quad i = 1, 2, 3 \quad (8)$$

If t is equal to t_0 , the curvature and torsion values are

$$k^i(t_0) = k^i(t) |_{t=t_0}, \quad \tau^i(t_0) = \tau^i(t) |_{t=t_0} \quad i = 1, 2, 3 \quad (9)$$

Give a DNA sequence with length of N , N curvature and torsion values will be obtained. The average curvature and torsion values of the N curvature and torsion values, denoted by k_0^i and τ_0^i respectively, can be computed as:

$$k_0^i = \left\{ \sum_{t=1}^N k^i(t) \right\} / N, \quad \tau_0^i = \left\{ \sum_{t=1}^N \tau^i(t) \right\} / N \quad i = 1, 2, 3 \quad (10)$$

As the curvature and torsion are character of the curve, they, in turn, can be regarded as descriptors to numerically characterize the curve. For extracting more characters from sequence, we construct a six-component vector, which is composed of three average curvatures and three average torsions, for numerical characterization the DNA sequence.

3. Results and Discussions

3.1. SIMILARITY ANALYSIS

Comparison of different DNA sequences is main application of our method. In Table I, the coding sequences of β -globin genes of 11 species and their exons are presented. Table II shows the six-component vectors of the coding sequences of the β -globin genes of 11 species.

Having a vector representation of a DNA sequence, we can compare various sequences by

TABLE I**The accession numbers, length, and location for each β -globin genes and their exons.**

Species	Database	ID	Location	Length (bp)	Location of each exon
1 Human	NCBI	U01317	62187-63610	1424	62187...62278, 62409...62631, 63482...63610
2 Chimpanzee	NCBI	X02345	4189-5532	1344	4189...4293, 4412...4633, 5484...5532
3 Gorilla	NCBI	X61109	4538-5881	1344	4538...4630, 4761...4982, 5833...5881
4 Lemur	NCBI	M15734	154-1595	1442	154...245, 376...598, 1467...1595
5 Rat	NCBI	X06701	310-1505	1196	310...401, 517...739, 1377...1505
6 Mouse	NCBI	V00722	275-1462	1188	275...367, 484...705, 1334...1462
7 Goat	NCBI	M15387	279-1749	1471	279...364, 493...715, 1621...1749
8 Bovine	NCBI	X00376	278-1741	1464	278...363, 492...714, 1613...1741
9 Rabbit	NCBI	V00882	277-1419	1143	277...368, 495...717, 1291...1419
10 Opossum	NCBI	J03643	467-2488	2022	467...558, 672...894, 2360...2488
11 Gallus	NCBI	V00409	465-1810	1346	465...556, 649...871, 1682...1810

using any of existing distance measures for vectors. The distance between two DNA sequences can be computed as the Euclidean distance between the end points of the two vectors representing them. The Euclidean distance between u and v is defined as:

$$d(u, v) = \left\{ \sum_{i=1}^6 (u_i - v_i)^2 \right\}^{1/2} \quad (11)$$

where u and v are vectors, u_i and v_i denote the six-component of the vectors u and v , respectively. The underlying rationale is that if two vectors points in similar direction and the difference in their magnitudes is small, then the two sequences represented by these vectors are similar. In other words, the smaller the Euclidean distance between the end points of two vectors, the more similar are the two sequences represented by these vectors. Table III denotes the similarity matrix of the coding sequences of the β -globin gene of 11 species. Following the same

method, we can also get the similarity matrices of the coding sequences of the each exon of 11 species, which are represented in Tables IV–VI.

In Table III, for the coding sequences of the β -globin gene of 11 species, it is obvious that the coding sequences of Gallus is the most dissimilar to the other 10 species, which is consistent with the fact that Gallus is non-mammal, whereas the others are mammal. The more similar species pairs are Human-Gorilla, Human-Chimpanzee, Rat-Mouse, and Gorilla-Chimpanzee, which are consistent with the results obtained by Randic [5, 19] and B. Liao [20]. In Tables IV–VI, for the single exon of the coding sequences of the β -globin gene of 11 species, there are some flaws. Some entries are not better than that of Table III.

To compare with other methods, we use the leading eigenvalues of E, L/L, M/M matrices [7] to perform the similarity analysis on the same data. The similarity for any pair of DNA sequences can be

TABLE II**The six-components vectors of coding sequence of the β -globin gene of 11 species.**

Pattern	AGTC (k)	AGTC (τ)	ATCG (k)	ATCG (τ)	ACGT (k)	ACGT (τ)
Human	0.63200	0.05099	0.63072	0.03132	0.62382	0.02744
Chimpanzee	0.63112	0.05498	0.62941	0.03825	0.62328	0.02229
Gorilla	0.62998	0.05751	0.62843	0.04263	0.62133	0.02117
Lemur	0.61447	0.05474	0.61306	0.01929	0.61350	0.02304
Rat	0.64599	0.03798	0.63502	0.04446	0.63577	0.04387
Mouse	0.64188	0.05029	0.63145	0.02985	0.63260	0.04137
Goat	0.64591	0.07325	0.63768	0.02751	0.63712	0.02328
Bovine	0.64807	0.06577	0.64012	0.02722	0.63735	0.01937
Rabbit	0.63439	0.04962	0.62955	0.01564	0.63016	0.02001
Opossum	0.64708	0.04546	0.63761	0.00544	0.63592	0.03010
Gallus	0.64163	0.07442	0.62966	−0.00025	0.63436	0.05336

TABLE III**The similarity matrix of the coding sequences of the β -globin gene of 11 species.**

Species	Human	Chimp-	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0	0.00965	0.01501	0.03007	0.03112	0.01929	0.03076	0.02881	0.01871	0.03360	0.04922
Chimp-		0	0.00574	0.03163	0.03467	0.02576	0.03048	0.02909	0.02455	0.04135	0.05531
Gorilla			0	0.03308	0.03752	0.03002	0.03271	0.03209	0.02984	0.04688	0.05889
Lemur				0	0.05762	0.04384	0.05063	0.05127	0.03154	0.04996	0.05601
Rat					0	0.02027	0.04431	0.04126	0.04161	0.04215	0.05888
Mouse						0	0.03057	0.02943	0.02691	0.02867	0.04048
Goat							0	0.00907	0.03094	0.03617	0.04203
Bovine								0	0.02731	0.03180	0.04631
Rabbit									0	0.02196	0.04528
Opossum										0	0.03883
Gallus											0

TABLE IV**The similarity matrix of the coding sequences of the first exon of the β -globin gene of 11 species.**

Species	Human	Chimp-	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0	0.04546	0.02183	0.15386	0.08196	0.10383	0.08294	0.06119	0.10542	0.10540	0.14157
Chimp-		0	0.04459	0.14779	0.11172	0.12163	0.11736	0.06849	0.12709	0.08631	0.13549
Gorilla			0	0.15326	0.07260	0.09546	0.08338	0.05888	0.11757	0.10732	0.15115
Lemur				0	0.17836	0.13171	0.17118	0.10132	0.11817	0.11156	0.08159
Rat					0	0.06226	0.04167	0.08762	0.13894	0.13912	0.19241
Mouse						0	0.06483	0.06798	0.12206	0.11249	0.16349
Goat							0	0.08303	0.11411	0.13033	0.17459
Bovine								0	0.08848	0.07362	0.10987
Rabbit									0	0.12287	0.09962
Opossum										0	0.10638
Gallus											0

TABLE V**The similarity matrix of the coding sequences of the second exon of the β -globin gene of 11 species.**

Species	Human	Chimp-	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0	0.00418	0.00992	0.05362	0.05805	0.05955	0.05134	0.04771	0.03977	0.05134	0.07491
Chimp-		0	0.00838	0.05061	0.05425	0.05626	0.05315	0.05023	0.03945	0.04832	0.07303
Gorilla			0	0.05111	0.05425	0.05689	0.06028	0.05695	0.04530	0.04747	0.07298
Lemur				0	0.03670	0.02035	0.09014	0.08925	0.05604	0.01426	0.03782
Rat					0	0.02628	0.08780	0.09124	0.06804	0.03381	0.06576
Mouse						0	0.09220	0.09382	0.06639	0.01814	0.04051
Goat							0	0.01432	0.05231	0.09081	0.11111
Bovine								0	0.04682	0.08976	0.10931
Rabbit									0	0.05919	0.08091
Opossum										0	0.03770
Gallus											0

TABLE VI**The similarity matrix of the coding sequences of the third exon of the β -globin gene of 11 species.**

Species	Human	Chimp-	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0	0.04846	0.09099	0.12016	0.09176	0.12910	0.11005	0.08975	0.09191	0.04514	0.09097
Chimp-		0	0.05847	0.13822	0.13576	0.14872	0.10518	0.10569	0.09840	0.06711	0.11851
Gorilla			0	0.13308	0.17006	0.15632	0.08933	0.11427	0.10655	0.08887	0.14676
Lemur				0	0.11037	0.06751	0.06350	0.05128	0.06878	0.07917	0.10755
Rat					0	0.10577	0.13933	0.09200	0.10699	0.08413	0.07091
Mouse						0	0.09428	0.06444	0.07437	0.09237	0.11379
Goat							0	0.05293	0.04704	0.07162	0.11004
Bovine								0	0.02260	0.05066	0.07055
Rabbit									0	0.05587	0.07534
Opossum										0	0.07596
Gallus											0

gotten by calculating the Euclidean distance between their leading eigenvalues. The similarity between Human and the other species are listed in Table VII. Table VII shows that our results are better than E, L/L, and M/M matrix. For example, Human is more similar to Chimpanzee and Gorilla in our method. But in E, L/L, and M/M matrix, Human is more similar to Lemur, which does not accord with the results reported in the references [5, 19, 20].

On the other hand, it is noteworthy that the eigenvalues of E, L/L, and M/M matrixes is computationally intensive. Its running times is 6.5-times longer than that of our method. For example, in the β -globin gene, the leading eigenvalues of E, L/L, and M/M matrixes take 2.103 h, and our method just 19.4 s, using a 1.41 GHZ, AMD with 512 MB total memory. It is obvious that our method performs faster.

TABLE VII**The comparison similarity between Human and the other 10 species based on our method and Yuan's method.**

Species	Chimp-	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
β gene										
E(10^6)	0.0770	0.0770	0.0179	0.2076	0.2142	0.0472	0.0401	0.2506	0.7159	0.0743
L/L	0.1143	0.1169	0.0140	0.3584	0.3761	0.0503	0.0409	0.4442	0.6943	0.1193
M/M(10^3)	0.0800	0.0800	0.0180	0.2279	0.2359	0.0471	0.0401	0.2810	0.5981	0.7914
My work	0.00965	0.01501	0.03007	0.03112	0.01929	0.03076	0.02881	0.01871	0.03360	0.04922
1st exon										
E(10^3)	0.8900	0.0642	0.0000	0.0001	0.1293	0.3711	0.3712	0.1266	0.0002	0.0001
L/L	0.2647	0.0165	0.0672	0.0111	0.0107	0.1455	0.1275	0.0737	0.0497	0.0177
M/M	13.0146	1.0175	0.1072	0.0032	2.0398	5.9294	6.0303	1.9919	0.1728	0.0229
My work	0.04546	0.02183	0.15386	0.08196	0.10383	0.08294	0.06119	0.10542	0.10540	0.14157
2nd exon										
E(10^4)	0.0155	0.0155	0.0000	0.0000	0.0155	0.0000	0.0000	0.0000	0.0000	0.0000
L/L	0.0087	0.0084	0.0045	0.0227	0.0165	0.0111	0.0029	0.0076	0.0025	0.0115
M/M	0.9990	0.9995	0.0394	0.0592	1.0380	0.0907	0.0524	0.0571	0.0263	0.1075
My work	0.00418	0.00992	0.05362	0.05805	0.05955	0.05134	0.04771	0.03977	0.05134	0.07491
3rd exon										
E(10^3)	4.9488	4.9488	0.0000	0.0000	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001
L/L	1.9337	1.9384	0.0150	0.0167	0.0116	0.0079	0.0204	0.0106	0.0113	0.0051
M/M	80.0582	80.0664	0.0425	0.0231	0.0701	0.0057	0.0256	0.0599	0.0985	0.1431
My work	0.04846	0.09099	0.12016	0.09176	0.12910	0.11005	0.08975	0.09191	0.04514	0.09097

TABLE VIII

The accession number, abbreviation, name and length for the 24 coronavirus geneomes.

No.	Accession	Abbreviation	Genome	Length (bp)
1	NC_002645	HCoV_229E	Human coronavirus 229E	27317
2	NC_002306	TGEV	Transmissible gastroenteritis virus	28586
3	NC_003436	PEDV	Porcine epidemic diarrhea virus	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebus	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	31028
6	AF220295	BCoVQ	Bovine coronavirus strain Quebec	31100
7	NC_003045	BCoV	Bovine coronavirus	31028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	31233
9	AF201929	MHV2	Murine hepatitis virus strain 2	31276
10	AF208066	MHVP	Murine hepatitis virus strain Penn 97-1	31112
11	NC_001846	MHV	Murine hepatitis virus strain A59	31357
12	NC_001451	IBV	Avian infectious bronchitis virus	27608
13	AY278488	BJ01	SARS coronavirus BJ01	29725
14	AY278741	Urbani	SARS coronavirus Urbani	29727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	29742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	29736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-Su10	29736
18	AY283794	SIN2500	SARS coronavirus Sin2500	29711
19	AY283795	SIN2677	SARS coronavirus Sin2677	29705
20	AY283796	SIN2679	SARS coronavirus Sin2679	29711
21	AY283797	SIN2748	SARS coronavirus Sin2748	29706
22	AY283798	SIN2774	SARS coronavirus Sin2774	29711
23	AY291451	TW1	SARS coronavirus TW1	29729
24	NC_004718	TOR2	SARS coronavirus	29751

3.2. PHYLOGENETIC ANALYSIS

Phylogenetics is the study of the evolutionary history among organisms. Moreover, it can provide information for function prediction. When sequences are grouped into families, it can provide us some clues about the general features of that family and evolutionary evidence of sequences.

Given a set of DNA sequences, their phylogenetic relationship can be gotten through the following main operation: first, we calculate the numerical characterizations of DNA sequences and the Euclidean distance between these numerical characterizations. Second, by arranging all the distance into a matrix, we obtain a distance matrix. Finally, we put the distance matrix into the UPGMA program in the PHYLIP package. We obtain the phylogenetic tree drawn by Treeview program.

To further demonstrate the utility of our method, we also analyze 24 coronavirus geneomes, which are listed in Table VIII. Recently, more attentions have been paid to atypical syndrome (SARS), which was first identified in Guangdong Province, China, and rapidly spread to several countries later. The research of the relationships between the

SARS-CoVs and the other coronaviruses can help to discover drugs and develop vaccines against the virus. The phylogenetic tree for 24 coronavirus geneomes is constructed by using our method, which is presented in Figure 3. To indicate the validity, we also constructed an evolutionary tree by the Clustal X method. Clustal X is a multiple sequence alignment program. The result is shown in Figure 4.

The topology of the tree obtained by our method (Fig. 3) is on the whole consistent with the established taxonomic groups, except for BCoV and IBV. Coronaviruses can be divided into four groups according to serotypes. Group I (HCoV_229E, TGEV, and PEDV) and group II (BCoVL, BCoVQ, BCoV, MHVM, MHV2, MHVP, and MHV) contain mammalian viruses, while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C virus [21]. Group III (IBV) contains only avian viruses, and Group IV [22, 23] are SARS-CoVs. Compared with the results in Figures 3 and 4, we can find some difference. In Figure 4, the result of Group IV is not clear. All in all, our method gives a more intuitively acceptable arrangement compared with the result of Clustal X.

10. Liu, X. Q.; Dai, Q.; Xin, Z. L.; Wang, T. M. *J Theor Biol* 2006, 243, 555.
11. Liao, B.; Li, R. F.; Zhu, W.; Xiang, X. Y. *J Math Chem* 2007, 42, 47.
12. Randic, M.; Zupan, J.; Novic, M. *J Chem Inf Comput Sci* 2001, 41, 1339.
13. Guo, Y.; Wang, T. M. *J Biomol Struct Dyn* 2008, 26, 367.
14. Feng, J.; Wang, T. M. *J Biomol Struct Dyn* 2008, 25, 621.
15. Li, C.; Wang, J. *J Chem Inf Model* 2005, 45, 115.
16. Yang, L. P.; Zhang, X.; Wang, T. M. *J Theor Biol* 2010, 262, 742.
17. Comtet, L. *Advanced Combinatorics*; D. Reidel Publishing Co.: Dordrecht, 1974.
18. Mei, X. M.; Huang, J. Z. *Differential Geometry*; Higher Education Press: Beijing, 2003.
19. Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. *Chem Phys Lett* 2003, 371, 202.
20. Liao, B.; Wang, T. *J Comput Chem* 2004, 25, 1364.
21. Lai, M. M. C.; Holmes, K. V. In: *Fields Virology*, 4th ed.; Knipe, D. M., Howley, P. M., Eds.; Lippincott Williams and Wilkins: New York, 2001; p 1163.
22. Marra, M. A.; Jones, S. J. M.; Astell, C. R.; Holt, R. A.; Angela, B. W.; Butterfield, Y. S. N.; Khattri, J.; Asano, J. K.; Barber, S. A.; Chan, S. Y.; Cloutier, A.; Coughlin, S. M.; Freeman, D.; Girn, N.; Griffith, O. L.; Leach, S. R.; Mayo, M.; McDonald, H.; Montgomery, S. B.; Pandoh, P. K.; Petrescu, A. S.; Robertson, A. G.; Schein, J. E.; Siddiqui, A.; Smailus, D. E.; Stott, J. M.; Yang, G. S.; Plummer, F.; Andonov, A.; Artsob, H.; Bastien, N.; Bernard, K.; Booth, T. F.; Bowness, D.; Czub, M.; Drebot, M.; Fernando, L.; Flick, R.; Garbutt, M.; Gray, M.; Grolla, A.; Jones, S.; Feldmann, H.; Meyers, A.; Kabani, A.; Li, Y.; Normand, S.; Stroher, U.; Tipples, G. A.; Tyler, S.; Vogrig, R.; Ward, D.; Watson, B.; Brunham, R. C.; Krajden, M.; Petric, M.; Skowronski, D. M.; Upton, C.; Roper, R. L. *Science* 2003, 300, 1399.
23. Rota, P. A.; Oberste, M. S.; Monroe, S. S.; Nix, W. A.; Campagnoli, R.; Icenogle, J. P.; Penaranda, S.; Bankamp, B.; Maher, K.; Chen, M. H.; Tong, S.; Tamin, A.; Lowe, L.; Frace, M.; DeRisi, L. L.; Chen, Q.; Wang, D.; Erdman, D. D.; Peret, T. C. T.; Burns, C.; Ksiazek, T. G.; Rollin, P. E.; Sanchez, A.; Liffick, S.; Holloway, B.; Limor, J.; McCaustland, K.; Melissa, O. R.; Fouchier, R.; Gnther, S.; Osterhaus, A. D. M. E.; Drosten, C.; Pallansch, M. A.; Anderson, L. J.; Bellini, W. J. *Science* 2003, 300, 1394.