Journal of Theoretical Biology 292 (2012) 78–85

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Sequence comparison via polar coordinates representation and curve tree

Qi Dai [a,*,1], Xiaodong Guo [b,1], Lihua Li [b,*]

[a] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[b] Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China

ARTICLE INFO

Article history:
Received 27 June 2011
Received in revised form
27 September 2011
Accepted 28 September 2011
Available online 6 October 2011

Keywords:
Graphical representation
The closed curve
Curve matching distance
Phylogenetic analysis

ABSTRACT

Sequence comparison has become one of the essential bioinformatics tools in bioinformatics research, which could serve as evidence of structural and functional conservation, as well as of evolutionary relations among the sequences. Existing graphical representation methods have achieved promising results in sequence comparison, but there are some design challenges with the graphical representations and feature-based measures. We reported here a new method for sequence comparison. It considers whole distribution of dual bases and employs polar coordinates method to map a biological sequence into a closed curve. The curve tree was then constructed to numerically characterize the closed curve of biological sequences, and further compared biological sequences by evaluating the distance of the curve tree of the query sequence matching against a corresponding curve tree of the template sequence. The proposed method was tested by phylogenetic analysis, and its performance was further compared with alignment-based methods. The results demonstrate that using polar coordinates representation and curve tree to compare sequences is more efficient.

1. Introduction

With the development of high-throughput sequencing technology, the rate of addition of new sequences to the databases increases continuously. However, such a collection of sequences does not by itself increase the scientist's understanding of the biology of organisms. Comparing a new sequence with the sequences of known functions is an effective way of assigning function to the new genes/proteins and understanding the biology of that organism from which the new sequence comes.

Many methods have been proposed for sequence comparison. They can be categorized into two classes. One is alignment-based methods, in which dynamic programming that finds an optimal alignment by assigning scores to different possible alignments and picking the alignment with the highest score. Several alignment-based algorithms have been proposed such as global alignment, local alignment, with or without overlap (Gotoh, 1982; Needleman and Wunsch, 1970; Smith and Waterman, 1981). Waterman (1995) and Durbin et al. (1998) provided comprehensive reviews about this method. However, the search for optimal solutions using sequence alignment has problems in computationally load with large biological databases and choice of the scoring schemes (Pham and Zuegg, 2004; Vinga and Almeida, 2003). Therefore, the second class, alignment-free methods, was proposed to overcome the limitations of alignment-based methods.

Graphical representation is one of widely used alignment-free methods. It provides a simple way of viewing, sorting and comparing various gene sequences with their intuitive pictures and pattern. Various graphical representations have been proposed during the past 10 years (Hamori and Ruskin, 1983; Gates, 1986; Nandy, 1994; Leong and Morgenthaler, 1995; Randic et al., 2003a, 2003b, 2006, 2001; Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009, 2011; Wu et al., 2010; Maaty et al., 2010a, 2010b; Bai et al., 2011; Xie and Mo, 2011; Yao and Wang, 2004; Liu et al., 2006; Randic, 2000; Qi and Qi, 2007; Qi et al., 2007). Randic et al. (2011) gave a comprehensive review on this method. All the graphical representations generally differ in two aspects: graphical representations and feature-based similarity measures.

Graphical representation of DNA sequences was first proposed by Hamori and Ruskin (1983) in which DNAs have been shown as 3D curves. Gates (1986), Nandy, (1994), and Leong and Morgenthaler (1995) developed 2D graphical representations of DNA sequences. These methods are straightforward but are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. Randic et al. (2003a) developed a novel 2D representation method to overcome the degeneracy of the graphical representation. Recently, several other 2D and 3D representations have been proposed (Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009, 2011; Wu et al., 2010; Maaty et al., 2010a, 2010b; Bai et al., 2011; Xie and Mo, 2011; Randic et al., 2003b, 2006, 2001; Yao and Wang, 2004; Liu et al., 2006; Randic, 2000; Qi and Qi, 2007; Qi et al., 2007). According to the handling bases of biological sequences, all the methods can be

[*] Corresponding authors.
E-mail addresses: daiailiu2004@yahoo.com.cn (Q. Dai), lilh@hdu.edu.cn (L. Li).
[1] Contributed equally to this work as co-first authors.

0022-5193/$ - see front matter Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.09.030

classified as: single nucleotide-based (Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009, 2011; Wu et al., 2010; Maaty et al., 2010a, 2010b; Bai et al., 2011; Xie and Mo, 2011; Randic et al., 2003a, 2006; Yao and Wang, 2004) and dual nucleotide-based representations (Liu et al., 2006; Randic, 2000; Randic et al., 2001; Qi and Qi, 2007; Qi et al., 2007). They often assign the n bases to corresponding points (Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009, 2011; Wu et al., 2010; Maaty et al., 2010a, 2010b; Bai et al., 2011; Xie and Mo, 2011), to the four lines (Randic et al., 2003b, 2006), or to the cell/system (Yao and Wang, 2004; Liu et al., 2006) to design the graphical representation.

Some features of graphical representations have been proposed to capture the essence of the base composition/distribution of the sequences and further facilitate biological sequence comparison. These widely used features are always associated with the central coordinate and distance matrices. The central coordinate can effectively characterize the whole changes of the geometrical curves and has been widely used for biological sequence comparison (Liao and Ding, 2006; Wen and Zhang, 2009; Abo ElMaaty et al., 2010). Another useful tool for characterization of biological sequences is distance matrix that is proposed by Randic and Vracko (2000) and further developed by Randic et al. (2001), Song and Tang (2005), and Liao and Wang (2004). They first transformed the graphical representations of biological sequences into distance matrices such as E matrix, D/D matrix, L/L matrix and their “high order” matrices. Then they extracted the invariants of matrices such as leading eigenvalue, average row element, etc. to numerically characterize the biological sequences and designed the feature-based similarity measure for sequence comparison.

Although the above graphical representation methods have achieved promising results, there are some problems in developing graphical representations and designing the feature-based similarity measures. First, many graphical representations were designed by assigning the single bases or dual nucleotides to corresponding direction/points/cells in Cartesian coordinates, so little attention has been paid to the whole distribution of the single nucleotide or dual nucleotides in biological sequences. Second, the choice of the direction/points/cells for the single base or dual nucleotides is arbitrary. Finally, the feature-based similarity measures are always associated with the invariants of the distance matrices that are gotten by complex repetitive computation. When the sequences are long, these kinds of feature-based similarity measures become less useful. Moreover, we believe that better representation and similarity measure will allow us to design more effective sequence comparison method.

This paper introduced a novel method to represent and compare biological sequences. Based on the whole distribution of the dual bases, we proposed a polar coordinates representation that maps a biological sequence into a closed curve. The closed curve was then transformed into a curve tree instead of the distance matrix, and a tree matching distance was proposed to estimate the similarity of two biological sequences. To assess the effectiveness of the proposed method, we took two experiments and compared its performance with the alignment-based method.

2. Method

2.1. Polar coordinates representation of DNA sequences

Given a DNA Sequence, almost all the graphical representations map it into a curve in Cartesian coordinates. The polar coordinates have not been used for graphical representation of DNA sequence until now. In addition, dual nucleotides have been introduced to design graphical representations (Liu et al., 2006; Randic, 2000; Randic et al., 2001; Qi and Qi, 2007; Qi et al., 2007), in which each dual nucleotide is assigned to a corresponding point in Cartesian coordinates, but the distribution of the dual nucleotides is not considered in graphical representation. Here, we propose a novel graphical representation of DNA sequence in polar coordinates based on the distribution of the dual nucleotides.

Given a DNA sequences, there are 16 kinds of the dual nucleotides. The distribution of the dual nucleotides consists of their frequencies in a given sequence. For a sequence s, the frequency of a dual nucleotide w_{XY}, denoted by $f(w_{XY})$, is the number of occurrence of w_{XY} in the sequence s, where $X \in \{A,C,G,T\}$, $Y \in \{A,C,G,T\}$. The standard approach for calculating the frequencies of the dual nucleotide in a sequence of length m is to use a sliding window of length 2, shifting the frame one base at a time from position 1 to $m-2+1$, in which dual nucleotides are allowed to overlap in the sequence. In this way, the distribution of the dual nucleotides is represented by a 16-dimensional vector F_2^s

$$\begin{aligned} F_2^s &= (f(w_{A,A}), f(w_{A,C}), \ldots, f(w_{T,T})) \\ &= (N(w_{A,A})/m-2+1, N(w_{A,C})/m-2+1, \ldots, N(w_{T,T})/m-2+1) \end{aligned} \tag{1}$$

where $N(w_{X,Y})$ is the count of the dual nucleotides XY in DNA sequence s, $X \in \{A,C,G,\mathrm{T}\}$, $Y \in \{A,C,G,T\}$. For convenience, we denote the vector $F_2^s(f(w_{A,A}), f(w_{A,C}), \ldots, f(w_{T,T}))$ as $F_2^s = (f_1, f_2, \ldots, f_{16})$.

When the vector F_2^s of DNA sequence is given, we are interested in its graphical representation in polar Coordinates. We first calculate the radius and angles of the distribution of the dual nucleotides as follows:

$$\begin{aligned} r(t) &= 1+\omega \times f_t, \quad t=1,2,\ldots,16, \qquad \omega=\{1,2\} \\ \theta(t) &= \theta(t-1)+2\pi \times f_t, \quad t=1,2,\ldots,16, \quad \theta(1)=f_1 \end{aligned} \tag{2}$$

where ω is a weighted value. Then we plot 16 feature points based the above radius and angles in the polar coordinates. Spline function is introduced to fit a smooth curve to a set of the radius and angles of the distribution of the dual nucleotides. Consider a cubic spline with abscissas x_i and ordinates y_i, $i=0,2,\ldots,N-1$. If the second derivatives at each point are known, the spline function has the form

$$y = S^3(x) = Ay_i + By_{i+1} + Cy_i'' + Dy_{i+1}'' \tag{3}$$

where $A = x_i - x/x_{i+1} - x_i$, $B = x - x_i/x_{i+1} - x_i$, $C = 1/6(A^3 - A)(x_{i+1} - x_i)^2$, $D = 1/6(B^3 - B)(x_{i+1} - x_i)^2$, and y'' is the second derivatives. Here, we choose x_i based on the distribution and the biological sequence

$$x(i) = x(i-1) + f(w_{s(i),s(i+1)}); \quad i=1,2,3,\ldots,n-1 \tag{4}$$

Using the spline function, we obtain the function values $y(i)$, $i=1,2,3,\ldots,n-1$. Plot $x(i)$ and $y(i)$, we will get the closed curve of a DNA sequence in the polar coordinates. For example, Fig. 1 is the polar representation of the coding sequence of the first exon of β-globin gene of Human.

As for the parameter ω in the definition of radius and angles, we have performed extensive experiments. The polar coordinates representation with different ω show a clear trends: $\omega=1$ is suitable for the short sequence. As the sequence length increases, its closed curve with $\omega=1$ is becoming more and more like a circle, which is not suitable for comparing various sequences with their intuitive pictures and pattern. For example, the polar coordinates representation of HCoV-229E coronavirus genomes is shown in Fig. 2(a). At the same time, if the value of ω is too large, the curvature difference on the small-scale will be covered that is not good for sequence representation either. Therefore, we should increase ω to a suitable value if the sequence length is large. As for the sequence, its polar representation becomes more inerratic with $\omega=2$ represented in Fig. 2(b).

2.2. Curve tree

In order to facilitate characterize and compare different polar representations, we map the closed curve into a curve tree that is constructed as following two steps: (1) dividing a closed curve into two open curves and determining the direction of them; (2) constructing the curve tree.

Given a closed curve, we should divide it into two open curves and determine the direction of the curves. Take a closed curve represented in Fig. 3 as an example, we first find the two farthest points on the curve: A and B, with which the closed curve is divided into two open curves AM_1B and BM_2A, M_1 and M_2 are the intersections of the perpendicular bisector of line AB with the two open curves. Then, compare the curvilinear path of the four curves: M_1A, M_1B, M_2A and M_2B. Their comparison determines the initial points and directions of the two open curves. For example, if the curvilinear path of curve M_1A is longer than other three curves, we define the curve AM_1B as AB that is the first curve with initial point A, and the curve BM_2A is the second curve denoted by CD whose initial point is C. When the lengths of the four curves are equal, we will define the initial point and the direction as the curvilinear path of curve M1A is the longest.

As for an open curve presented in Fig. 4, we will find its initial point A and M_0, the midpoint of line AB. We draw the perpendicular bisector of line AB denoted by M_0M that intersects the line AB at M. Then, we define $h=(-1)^{a}MM_0/AM_0$ as the directed relative height of the line AB, where $a\in\{0,1\}$. If the vector $\overrightarrow{AM}$, $\overrightarrow{MM_0}$ and the vector $\boldsymbol{z}$ that is upright perpendicular to the plane AMM_0, are satisfied with the Right-Hand Rule, a is equal to 1; otherwise, a is equal to 0.

The curve tree nodes store the directed relative heights of the line. Taking the curve of Fig. 5(a) for an example, the directed relative height of line AB, denoted as $h_{0,0}$, is stored in the root node of the curve tree. The curve AB is divided into two open curves by the point $M_{0,0}$. For these two open curves, we repeat the procedure mentioned above, their results are shown in Fig. 5(b). We obtain the left child node and right child node of the root node $h_{0,0}$, which are denoted as $h_{1,0}$ and $h_{1,1}$, respectively. We then take the notes $h_{1,0}$ and $h_{1,1}$ as root nodes of the sub-trees, and repeat the operations until the obtained curve is almost a straight line. At last, we get a curve tree of the line AB, which is presented in Fig. 6.

2.3. Curve match distance

Given a DNA sequence, we can map it into a closed curve and construct a curve tree to characterize the closed curve. Here, we are not only interested in using the curve tree to characterize the closed curve, but also interested in facilitating comparison of the polar representation of DNA sequences.

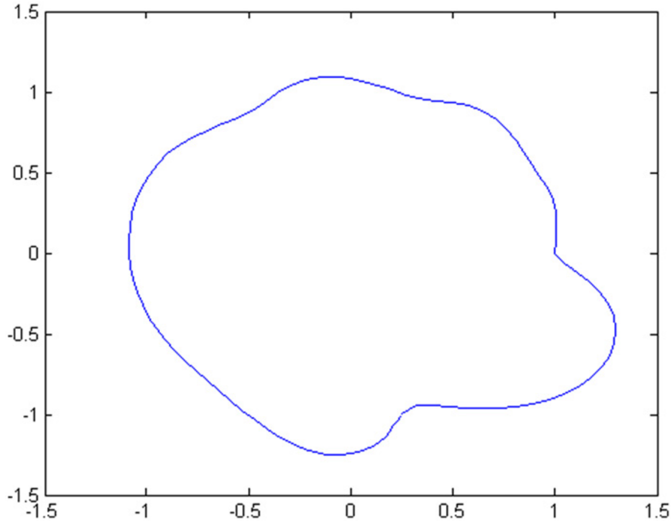

Fig. 1. Polar representation of coding sequence of the first exon of Human β-globin gene with $w=1$.

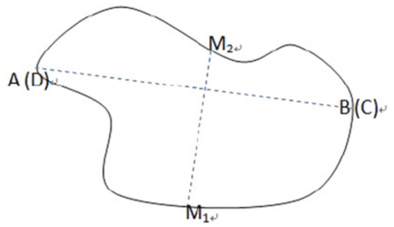

Fig. 3. Closed curve with two farthest points A and B, M_1M_2 is the perpendicular bisector of the line AB.

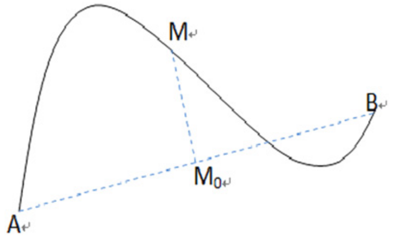

Fig. 4. Curve with its initial point A, M_0 is the midpoint of line AB.

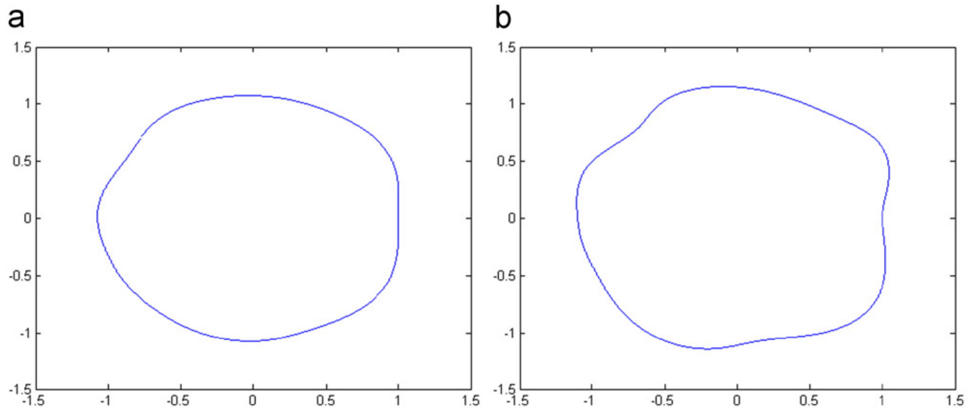

Fig. 2. Polar representation of HCoV-229E coronavirus genomes with $\omega=1$ (a) and $\omega=2$ (b).

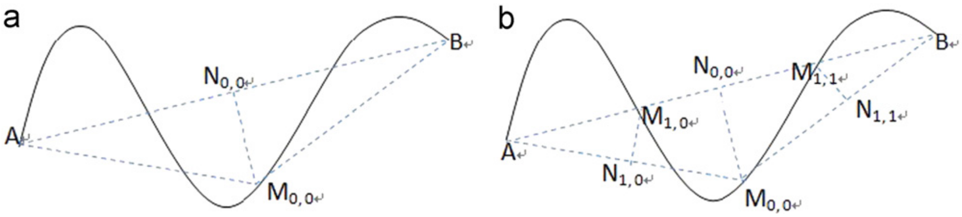

Fig. 5. Construction of curve tree, (a) the curve AB is divided into two open curves by the point $M_{0,0}$, (b) the curves $AM_{0,0}$ and $M_{0,0}B$ is further divided into two open curves by the point $M_{1,0}$ and $M_{1,1}$.

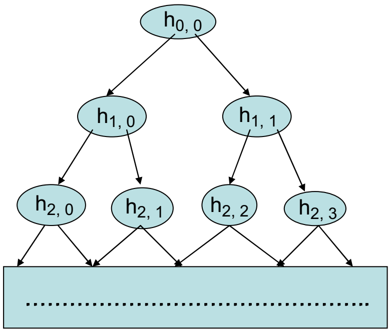

Fig. 6. Curve tree of the curve presented in Fig. 5.

As we all know, the more similar two sequences, the closer the two sequences relate. It is also true for the curve trees. That is to say, if the relations of two curve trees are closer, they are more similar. On the basis of this assumption, we define a curve match distance of the two curves C_1 and C_2 as follows:

$$Dist(C_1,C_2)=\sum_{i=0}^{n-1}\sum_{j=0}^{2^i}\zeta_i\psi(|h_{i,j}^{T_1}-h_{i,j}^{T_2}|)/\sum_{i=0}^{n-1}\sum_{j=0}^{2^i}\zeta_i \tag{5}$$

where $h_{i,j}^{T_1}(h_{i,j}^{T_2})$ is the values of the j-th node on the i-th layer in the complete binary tree corresponding to the curve tree T_1 (T_2) of the curve C_1 (C_2), ζ_i is a weight, n is the layers that is determined by the actual precision required and curvature curve, $\psi(x)$ is monotonically increasing function that is defined as following:

$$\psi(x)=\begin{cases} x^2+(1-a)x & x\in[0,a) \\ x & x\in[a,b] \\ x^2+(1-2b)x+b^2 & x\in(b,\infty) \end{cases} \tag{6}$$

The function $\psi(x)$ enlarges the distance between the larger local differences to improve comparison accuracy, and reduces the distance between the smaller local differences to increase the anti-jamming ability of the curve distance. $\{\zeta_i\}_{i=0}^{n-1}$ is a weight series that influences the distance by the element differences of the curve tree.

The value of $\{\zeta_i\}_{i=0}^{n-1}$ should be set as the appropriate value on the basis of actual needs. If the differences of large-scale curvatures are as the same as that of small-scale curvatures of the curve, it is better to choose $\{\zeta_i\}_{i=0}^{n-1}$ as a constant series. If we pay more attention on the curvature difference on the large-scale when comparing the different curves, it is better to choose $\{\zeta_i\}_{i=0}^{n-1}$ as a descending series; otherwise, it is better to choose $\{\zeta_i\}_{i=0}^{n-1}$ as an increasing series.

3. Results and discussion

Biological sequence comparison is the essential motivation of polar representation of DNA sequences. Here, we propose intuitive and quantitative methods to compare biological sequences with help of the proposed polar representation.

3.1. Sequence comparison with polar representation

The alphabet representation of biological sequences is easily handled with computer but difficult for us to observe their differences. Graphical representation provides us with a simple way to view various biological sequences and facilitate sequence comparison with the intuitive pictures and pattern.

In Fig. 7, we present the polar representations of the first coding sequences of β-globin gene of Human, Gorilla, Gallus, and Rabbit. Comparing the closed curves, it is easily to find that the most similar pair is Human–Gorilla because they are Primates. The more similar pairs are Human–Rabbit and Gorilla–Rabbit, which is consist with the (Ferungulates, (Primates, Rodents)) grouping (Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009; Liao and Ding, 2006; Wen and Zhang, 2009). The closed curve of Gallus is dissimilar to the other because it is the only non-mammal animals among them. Therefore, polar representation provides us with a simple way to compare different biological sequences.

For comparison, we list the recently published results of the examination of the degree of similarity between Human and other several species in Fig. 8 (Zhang, 2009; Yu et al., 2009; Wang et al., 2010; Xie and Mo, 2011). As one can see there is an overall agreement among similarities obtained by different approaches despite some variation among them. But it is also noted that the degree of dissimilarities of Human–Goat and Human–Bovine are larger than that of Human–Opposum and Human–Gallus, which is an undesirable result because Gallus is the only non-mammal among them, and Opossum is the most remote species from the remaining mammals.

3.2. Phylogenetic analysis

Since the proposed curve matching distance $Dist(C_1,C_2)$ is a distance measure, we can further evaluate the proposed method with phylogenetic analysis. Here, we choose two date sets that have been studied by many researchers (Liao and Wang, 2004; Chi and Ding, 2005; Yao et al., 2005; Zhang and Liao, 2007; Zhang and Chen, 2006; Huang et al., 2009; Liao and Ding, 2006; Wen and Zhang, 2009; Gu et al., 2004; Zhang, 2009). The first data set consists of the first exon of β-globin gene of 11 species presented in Table 1. It is a small data set with average sequence length 92. The second data set are 24 coronavirus genomes with average length about 30,000. They are downloaded from GenBank, of which 12 are SARS-CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 2.

Given a set of biological sequences, their phylogenetic relationship can be obtained through the following main operations: firstly, we construct the polar representation of biological

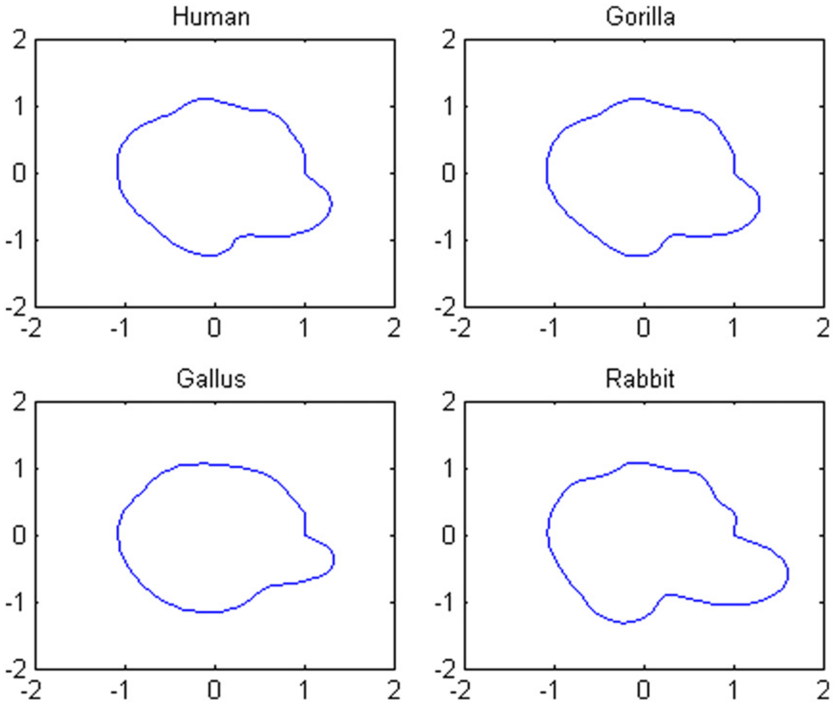

Fig. 7. Polar representations for the first coding sequences of β-globin gene of Human, Gorilla, Gallus, and Rabbit.

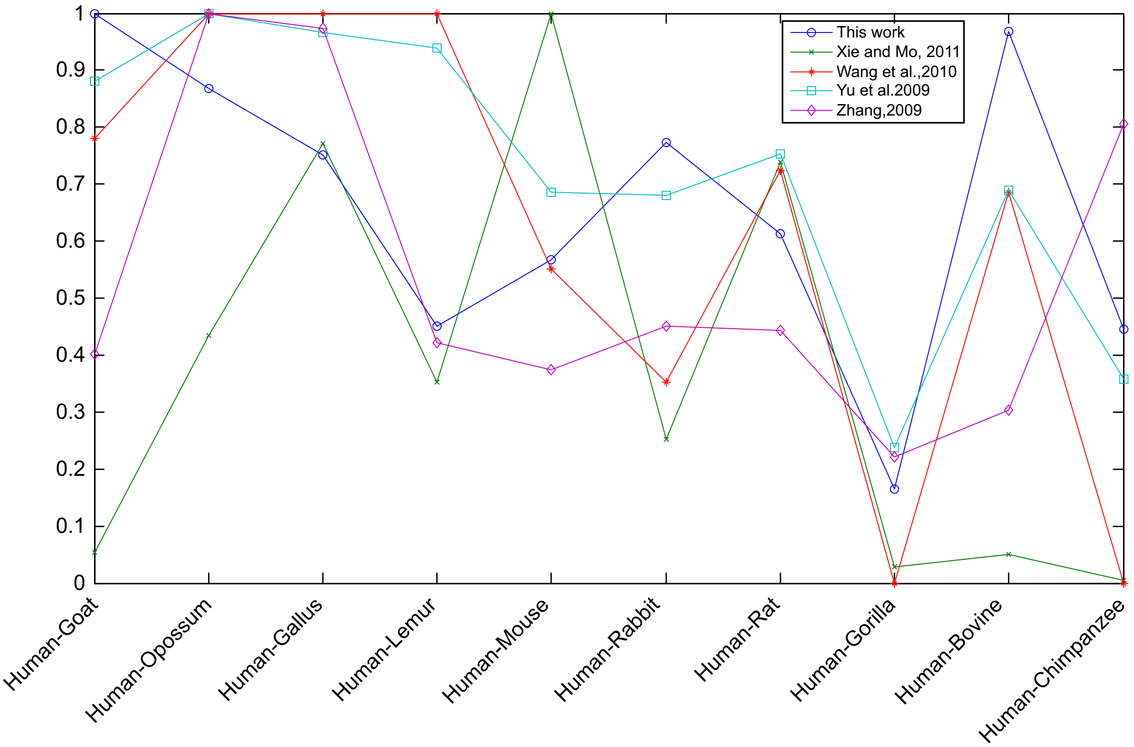

Fig. 8. Comparison degree of similarity/dissimilarity between Human and other several species in Table 1 with the proposed method and methods in Zhang (2009), Yu et al. (2009), Wang et al. (2010), and Xie and Mo (2011).

sequences and calculate their curve matching distance based on the curve tree; secondly, by arranging all the curve matching distance into a matrix, we obtain a pair-wise distance matrix; finally, we put the pair-wise distance matrix into the neighbor-joining program in the PHYLIP package (Felsenstein, 1989). Fig. 9(a) is phylogenetic tree of the first exon of β-globin gene of 11 species using the proposed method with $\omega=1$ and $\{\xi_i\}_{i=0}^{n-1}=\{(1.2)^i\}_{i=0}^{n-1}$. Fig. 10(a) is phylogenetic tree of the 24 coronavirus genomes obtained using the proposed method with $\omega=2$ and $\{\xi_i\}_{i=0}^{n-1}=\{(1.2)^i\}_{i=0}^{n-1}$.

Generally, an independent method can be developed to evaluate the accuracy of a phylogenetic tree, or the validity of a phylogenetic tree can be tested by comparing it with authoritative ones. Here, we adopt the form one to test the validity of our

Table 1
Sequences of the first exon of β-globin gene of different species.

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

Table 2
Accession number, abbreviation, name and length for each of the 24 coronavirus genomes.

No	Accession	Group	Abbreviation	Genome	Length (nt)
1	NC_002645	I	HCoV-229E	Human coronavirus 229E	27,317
2	NC_002306	I	TGEV	Transmissible gastroenteritis virus	28,586
3	NC_002436	I	PEDV	Porcine epidemic diarrhea virus	28,033
4	U00735	II	BCoVM	Bovine coronavirus strain Mebus	31,032
5	AF391542	II	BCoVL	Bovine coronavirus isolate BCoV-LUN	31,028
6	AF220295	II	BCoVQ	Bovine coronavirus strain Quebec	31,100
7	NC_003045	II	BCoV	Bovine coronavirus	31,028
8	AF208067	II	MHVM	Murine hepatitis virus strain ML-10	31,100
9	AF201929	II	MHV2	Murine hepatitis virus stain 2	31,028
10	AF208066	II	MHVP	Murine hepatitis virus strain Penn 97-1	31,233
11	NC_001846	II	MHV	Murine hepatitis virus	31,276
12	NC_001451	III	IBV	Avian infectious bronchitis virus	27,608
13	AY278488	IV	BJ01	SARS coronavirus BJ01	29,725
14	AY278741	IV	Urbani	SARS coronavirus Urbani	29,727
15	AY278491	IV	HKU-39849	SARS coronavirus HKU-39849	29,742
16	AY278554	IV	CUHK-W1	SARS coronavirus CUHK-W1	29,736
17	AY282752	IV	CUHK-Su10	SARS coronavirus CUHK-Su10	29,736
18	AY283794	IV	SIN2500	SARS coronavirus Sin2500	29,711
19	AY283795	IV	SIN2677	SARS coronavirus Sin2677	29,705
20	AY283796	IV	SIN2679	SARS coronavirus Sin2679	29,711
21	AY283797	IV	SIN2748	SARS coronavirus Sin2748	29,706
22	AY283798	IV	SIN2774	SARS coronavirus Sin2774	29,711
23	AY291451	IV	TW1	SARS coronavirus TW1	29,729
24	NC_004718	IV	TOR2	SARS coronavirus	29,751

phylogenetic tree. Both two data sets are aligned with the multiple alignmen CLUSTAL X and use the neighbor-joining to construct the phylogenetic tree presented in Figs. 9(b) and 10(b).

From Fig. 9, we find that the eleven species are separated clearly in our results: (1) three Primates (Human, Gorilla and Chimpanzee) are clustered closely; (2) two Rodents (Mouse and Rat) are grouped closely; (3) Rabbit is clustered closely with Human, Gorilla and Chimpanzee. (4) Opossum and Gallus are less closely with other species, which is consistent with the fact that Gallus is the only non-mammal among them, and Opossum is the most remote species from the remaining mammals. Our results are consistent with the results of the multiple alignment CLUSTAL X (Fig. 9(b)).

Fig. 10(a) shows that our results are quite consistent with the authoritative results (Gu et al., 2004; Zhang, 2009) and that of the multiple alignment in the following aspects. First of all, all SARS-CoVs are grouped in a separate branch, which appear different from the other three groups of coronaviruses. Secondly, BCOV, BCOVL, BCOVM, BCOVQ, MHV, MHV2, MHVM, and MHVP are grouped into a branch, which is consonant with that they belong to group II. Thirdly, HCoV-229E, TGEV, and PEDV are closely related to each other, which is consistent with the fact that they belong to group I. Finally, IBV forms a distinct branch within the genus Coronavirus, because it belongs to group III. Rota et al. (2003) found out that the overall level of similarity between SARS-CoVs and the other coronaviruses is low. Our tree also reconfirms that SARS-CoVs are not closely related to any previously isolated coronaviruses and form a new group, which indicates that the SARS-CoVs have undergone an independent evolution path after the divergence from the other coronaviruses.

4. Conclusion

Sequence comparison is one of the major goals of sequence analysis, which could serve as evidence of structural and functional conservation, as well as of evolutionary relations among the sequences. Despite the prevalence of the alignment-based methods, it is also noteworthy that it is computationally intensive and consequently unpractical for querying large data sets. Therefore, considerable efforts have been made to seek for alternative methods for sequence comparison.

Graphical representation is one of widely used alignment-free methods to view, sort, and compare biological sequences. This work presented a novel method to represent and compare biological sequence. In contrast to the existing graphical representations, we used the whole distribution of the dual bases to map a biological sequence into a closed curve in polar coordinates. Then we transformed the closed curve into a curve tree instead of the distance matrix, and proposed a tree matching

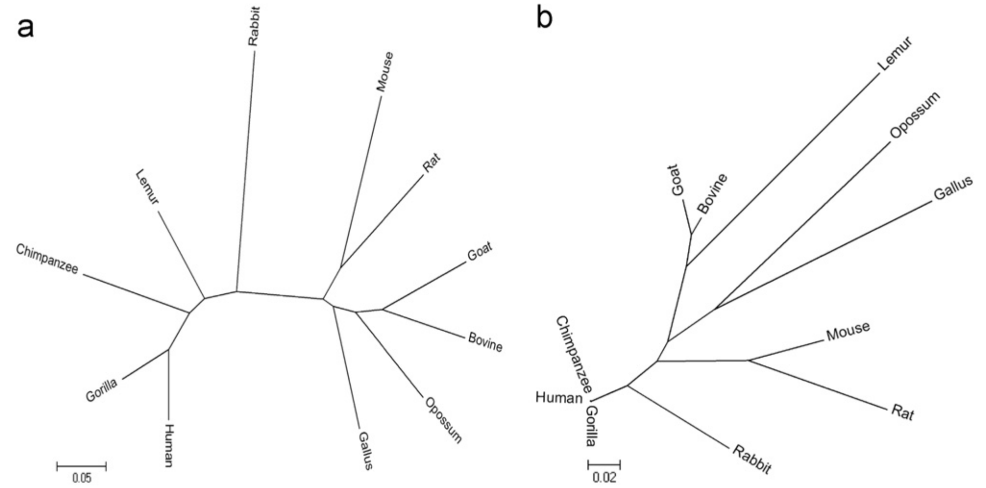

Fig. 9. Phylogenetic tree of 11 species based on (a) the proposed method, (b) the multiple alignment CLUSTAL X.

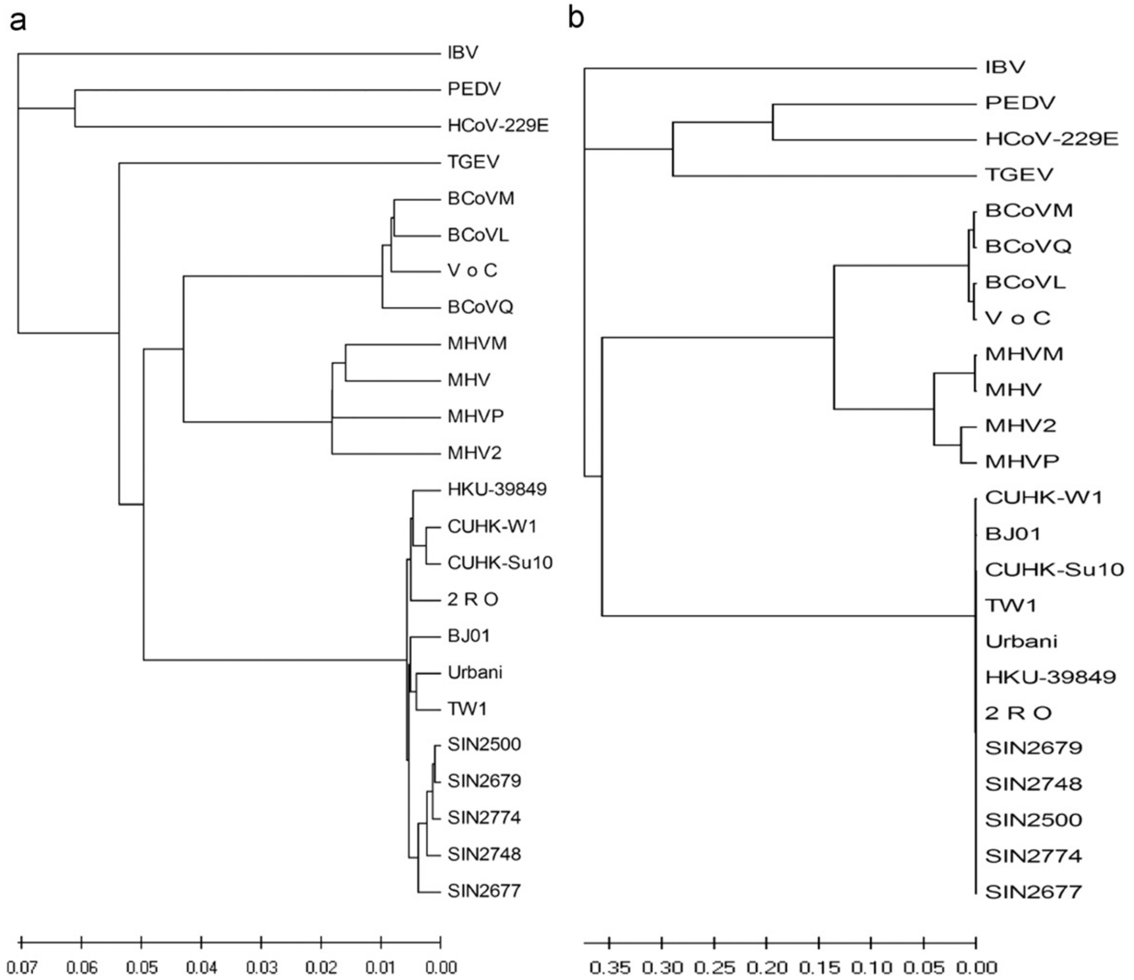

Fig. 10. Phylogenetic tree of 24 coronavirus genomes based on (a) the proposed method and (b) the multiple alignment CLUSTAL X.

distance to estimate the similarity of two biological sequences. To compare the effectiveness of the proposed method, we performed extensive tests including similarity of biological sequences and phylogenetic analysis, and compared its performance with alignment-based method. The results demonstrate that the proposed method is efficient, which highlight the necessity for graphical representation method to consider the whole distribution of the dual bases. Thus, this understanding can then be used to guide development of more powerful graphical representation for biological sequence comparison.

Acknowledgments

We thank the referees for many valuable comments that have improved this manuscript. This work is supported by the National

Natural Science Foundation of China (61170316, 61001214), National Distinguished Young Research Scientist Award (60788101), and a research Grant (Y2100930) from Zhejiang Provincial Natural Science Foundation of China.

References

Abo ElMaaty, M.I., Abo-Elkhier, M.M., AbdElwahaab, M.A., 2010. 3D graphical representation of protein sequences and their statistical characterization. Physica A 389, 4668–4676.

Bai, F.L., Zhang, J.H., Zheng, J.S., 2011. Similarity analysis of DNA sequences based on the EMD method. Appl. Math. Lett. 24, 232–237.

Chi, R., Ding, K.Q., 2005. Novel 4D numerical representation of DNA sequences. Chem. Phys. Lett. 407, 63–67.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis. Cambridge University Press.

Felsenstein, J., 1989. PHYLIP-Phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Gates, M., 1986. A simple way to look at DNA. J. Theor. Biol. 119, 319–328.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155–161.

Hamori, E., Ruskin, J., 1983. H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J. Biol. Chem. 25, 1318–1327.

Huang, G.H., Liao, B., Li, Y.F., Yu, Y.G., 2009. Similarity studies of DNA sequences based on a new 2D graphical representation. Biophys. Chem. 14, 355–359.

Huang, G.H., Zhou, H.Q., Li, Y.F., Xu, L.X., 2011. Alignment free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. 281, 107–112.

Leong, P.M., Morgenthaler, S., 1995. Random walk and gap plots of DNA sequences. Comput. Appl. Biosci. 11, 503–507.

Liao, B., Ding, K., 2006. A 3D graphical representation of DNA sequences and its application. Theor. Comput. Sci. 358, 56–64.

Liu, X.Q., Dai, Q., Xiu, Z.L., Wang, T.M., 2006. PNN-curve: a new 2D graphical representation of DNA sequences and its application. J. Theor. Biol. 243, 555–561.

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

Maaty, M.A., Abo-Elkhier, M.M., Elwahaab, M.A.A., 2010a. 3D graphical representation of protein sequences and their statistical characterization. Physica A 389, 4668–4676.

Maaty, M.A., Abo-Elkhier, M.M., Elwahaab, M.A.A., 2010b. Representation of protein sequences on latitude-like circles and longitude-like semi-circles. Chem. Phys. Lett. 493, 386–391.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure: methodology and application to globin genes. Curr. Sci. 66, 309–314.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Pham, T.D., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20, 3455–3461.

Qi, Z.H., Qi, X.Q., 2007. Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chem. Phys. Lett. 440, 139–144.

Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theor. Biol. 249, 681–690.

Randic, M., 2000. Condensed representation of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40, 50–56.

Randic, M., Vracko, M., 2000. On the similarity of DNA primary sequences. J. Chem. Inf. Comput.Sci. 40, 599–606.

Randic, M., Guo, X.F., Basak, S.C., 2001. On the characterization of DNA primary sequence by triplet of nucleic acid bases. J. Chem. Inf. Comput. Sci. 41, 619–626.

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

Randic, M., Zupan, J., Vikic-Topic, D., Plavsic, D., 2006. A novel unexpected use of a graphical representation of DNA: graphical alignment of DNA sequences. Chem. Phys. Lett. 431, 375–379.

Randic, M., Zupan, J., Balaban, A., Vikic-Topic, D., Plavsic, D., 2011. Graphical representation of proteins. Chem. Rev. 111, 790–862.

Rota, P.A., Oberste, M.S., Monroe, S.S., et al., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Song, J., Tang, H., 2005. A new 2-D graphical representation of DNA sequences and their numerical characterization. J. Biochem. Biophys. Methods 63, 228–239.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19, 513–523.

Wang, S.Y., Tian, F.C., Qiu, Y., Liu, X., 2010. Bilateral similarity function: a novel and universal method for similarity analysis of biological sequences. J. Theor. Biol. 265, 194–201.

Waterman, M.S., 1995. Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics. Chapman and Hall/CRC, Boca Raton, FL.

Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267, 29–34.

Wen, J., Zhang, Y., 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chem. Phys. Lett. 476, 281–286.

Xie, G.S., Mo, Z.X., 2011. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. J. Theor. Biol. 269, 123–130.

Yao, Y.H., Wang, T.M., 2004. A class of new 2-D graphical representation of DNA sequences and their application. Chem. Phys. Lett. 398, 318–323.

Yao, Y.H., Nan, X.Y., Wang, T.M., 2005. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. Chem. Phys. Lett. 411, 248–255.

Yu, J.F., Sun, X., Wang, J.H., 2009. TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J. Theor. Biol. 261, 459–468.

Zhang, Z.J., 2009. DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics 25, 1112–1117.

Zhang, Y.S., Chen, W., 2006. Invariants of DNA sequences based on 2DD-curves. J. Theor. Biol. 242, 382–388.

Zhang, Y.S., Liao, B., 2007. On the similarity of DNA sequences based on 3-D graphical representation. J. Biomath. 22, 583–590.