MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **72** (2014) 533-550

ISSN 0340 - 6253

A Novel Visualization of DNA Sequences, Reflecting GC-Content

Zhujin Zhang[1,5,*], Jiuyong Li[2], Linqiang Pan[3], Yunming Ye[1,5], Xiangxiang Zeng[4], Tao Song[3], Xiaofeng Zhang[1,5], Eric Ke Wang[1,5,*]

[1]*Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China*
[2]*School of Computer & Information Science, University of South Australia, Adelaide, SA 5095, Australia*
[3]*School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China*
[4]*Department of Computer Science, Xiamen University, Xiamen 361005, China*
[5]*Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen 518055, China*

(Received January 16, 2014)

Abstract

Significant progresses of visualization technology of DNA sequences have been made by solving visual effect problems of degeneracy, loss of information, difficulty in multi-dimensional space and difficulty for long DNA sequences. Different from traditional models focusing on visualization effect problems, we propose a novel visualization tool — GC-Curve, which not only solves all the visual effect problems mentioned above, but also can show the GC-content of a DNA. GC-content is an important feature of DNA, which is related to the stability of DNA, the density of genes, natural selection, mutational bias, etc. So the visualization reflecting GC-content has great potential in many applications. The applications of GC-Curve on similarity analysis, GC-content analysis, stability analysis and melting temperature prediction are presented. A software of GC-Curve is available at https://www.box.com/s/g872v3pq4kuz86sj5coq.

1 Introduction

The availability of massive data of DNA sequences creates a great need of viewing and analyzing DNA sequences. Graphical representation is considered as a visualization tool of DNA

*Corresponding author. Email: zhangzhujin@gmail.com; wk_hit@hitsz.edu.cn

sequence, and provides useful insights into local and global characteristics of a sequence, which are not easily obtainable by other methods [1].

Since the first graphical representation of DNA sequences was proposed by Hamori and Ruskin [2, 3], a large number of different graphical representations have been published. In the early phase, Gates [4], who viewed a DNA sequence as a "path" in the (x, y) plane [5], designed an important graphical representation, which is very simple and easy to observe. In this representation, four vectors in orthogonal directions, representing four bases, move around in the Cartesian coordinate to draw a graph, which is called DNA walk technology. However, these three visualization methods are accompanied by high degeneracy and loss of information.

Degeneracy and loss of information became two main barriers of DNA graphical representations [6]. Many researchers made great efforts to solve these two problems. Guo *et al.* [7] improved Gates's model with a low degeneracy representation. Wu *et al.* [8] introduced a DB-Curve with non-degeneracy, but with loss of information.

Some scholars found that it was much easier to solve these two problems in multi-dimensional space than in 2D space. For example, Zhang and Zhang [9, 10], Yao *et al.* [11], Qi and Fan [12], Qi *et al.* [13], Cao *et al.* [14], Yu *et al.* [15], Xie and Mo [16], Yu *et al.* [17], Cao *et al.* [18], Huang and Wang [19] adopted 3D graphical representations. Liao *et al.* [20], Chi and Ding [21], Tang *et al.* [22] used 4D approaches.

But as stated in [23], visualization effect in multi-dimensional space is not as good as in 2D space. Scientists continued to study visualization methods in 2D space. For example, Randić *et al.* [24], Qi and Qi [25], Zhang *et al.* [26], Bielińska-Wąż, D. and Subramaniam [27,28] adopted spectral representations. Stephen *et al.* [29], Zhang *et al.* [1], Huang *et al.* [30], Zhang [23] continued to improve DNA walk technology. Randić *et al.* [31], Qi *et al.* [32] used coding methods. Randić *et al.* [33, 34] proposed representations of DNA as maps.

Besides degeneracy and loss of information, many representations need a lot of space. Jeffrey [35] first designed a compact representation, which needs limited space to represent long sequences. But the representation has degeneracy. Randić *et al.* [31], Zhang *et al.* [36] proposed compact graphical representations, avoiding degeneracy and loss of information. In addition to these three properties, Randić *et al.* [34], and Zhang *et al.* [37] made representations colorful.

Zhang [23] found that many models, such as [24, 25, 38–40], suffered difficulty of observing when showing long DNA sequences, and built a dual-vector model to conquer the problem.

However, all the problems mentioned above are visual effect problems. We ask: "beside

these visual effect problems, is a visualization tool able to reflect some biological feature of a DNA? And build more applications according to the biological feature." There are some buds of this idea. Some visualizations can reflect the length of a DNA sequence, for example [23, 24], and some can reflect the difference of numbers between two kinds of DNA bases, for example [16, 23]. But they are incidental and too simple. In this paper, we propose a novel visualization—GC-Curve. GC-Curve not only solve four visual effect problems at the same time, but also can reflect the GC-content of a DNA. GC-content is an important feature of DNA, which is related to the stability of DNA [41], the density of genes [42], natural selection [43,44], mutational bias [43], etc. So the visualization of GC-content is very helpful to analyze DNA sequences.

2 GC-Curve

Zhang [23] proposed a dual-vector visualization model of DNA sequences, which shows that dual-vector technology is more powerful than one-vector technology. In this work, we still adopt the dual-vector technology. In this section, we will introduce the construction of GC-Curve. According to the construction of GC-Curve, two mathematic models have been built. Finally, we present properties of GC-Curve.

2.1 Construction of GC-Curve

In this subsection, the construction of GC-Curve is given. As shown in Fig. 1a, each DNA base of G, C, T and A is indicated by two vectors as follows:

$$(1, 1), (1, 1) \Longrightarrow \text{G}$$
$$(1, 2), (1, 0) \Longrightarrow \text{C}$$
$$(1, 0), (1, 0) \Longrightarrow \text{T}$$
$$(1,\text{-}1), (1, 1) \Longrightarrow \text{A}$$

A GC-Curve can be obtained by connecting all the vectors one by one. For example, as shown in Fig. 1b, the GC-Curve of DNA sequence 'CATG' is given. In particular, we start at the origin the point (0,0). The first base is 'C', so we move from the point (0,0) to the point (1,2), and then to the point (2,2) according to the assignments mentioned above. The second base is 'A', so we move from the point (2,2) to the point (3,1), and then to the point (4,2). Continue the moving, we can get the whole GC-Curve.

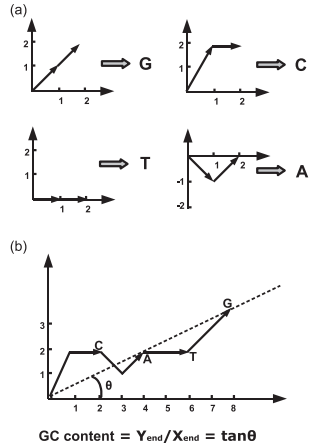

Figure 1: (a) The representations of four DNA bases in GC-Curve. (b) The GC-Curve of sequence 'CATG'. From this figure, we can know that GC-Curve extends 2U along X axes to represent each nucleotide no matter 'A', 'T', 'C' or 'G', so the value of the end point in X axes can reflect the length of the given DNA sequence. From this figure, we can also know that GC-Curve extends 2U along Y axes to represent 'G' or 'C' whereas GC-Curve extends nothing along Y axes to represent 'A' or 'T', so the value of the end point in Y axes can reflect the number of G+C of the given DNA sequence. So GC-content can be reflect by $Y_{end}/X_{end} = tan\theta$. The mathematical proofs are presented.

2.2 Mathematical Models of GC-Curve

According to the construction of GC-Curve, we present two mathematical models of GC-Curve in this subsection. Firstly, we need to define some variables:

1, we describe a DNA sequence as $SEQ = S_1S_2...S_i...S_n$ where $S_i \in \{A, T, C, G\}$, n is the length of this DNA sequence and SEQ is the whole DNA sequence. It means the DNA sequence SEQ is a sequence of alphabets.

2, (X_j, Y_j) is the point of GC-Curve. $(X_0, Y_0) = (0, 0)$ is the start point, and (X_{end}, Y_{end}) is the end point.

Model 1. Given a DNA sequence, draw the GC-Curve

$$Y_{2i-1} = \begin{cases} Y_{2i-2} + 1, & \text{if } S_i = \text{G}, \\ Y_{2i-2} + 2, & \text{if } S_i = \text{C}, \\ Y_{2i-2}, & \text{if } S_i = \text{T}, \\ Y_{2i-2} - 1, & \text{if } S_i = \text{A}. \end{cases}, i = 1, 2, ..., n. \quad (1)$$

$$Y_{2i} = \begin{cases} Y_{2i-1} + 1, & \text{if } S_i = \text{G}, \\ Y_{2i-1}, & \text{if } S_i = \text{C}, \\ Y_{2i-1}, & \text{if } S_i = \text{T}, \\ Y_{2i-1} + 1, & \text{if } S_i = \text{A}. \end{cases}, i = 1, 2, ..., n. \tag{2}$$

$$X_{2i-1} = 2i - 1, i = 1, 2, ..., n. \tag{3}$$

$$X_{2i} = 2i, i = 1, 2, ..., n. \tag{4}$$

In this model, each S_i is given. We can calculate each point (X_j, Y_j) of GC-Curve according to these 4 equations, and then connect all points with beelines. In this way, we can obtain the GC-Curve.

Model 2. Given a GC-Curve, obtain the DNA sequence

$$S_i = \begin{cases} G, & \text{if } Y_{2i} - Y_{2i-1} = 1 \ \& \ Y_{2i-1} - Y_{2i-2} = 1, \\ C, & \text{if } Y_{2i} - Y_{2i-1} = 0 \ \& \ Y_{2i-1} - Y_{2i-2} = 2, \\ T, & \text{if } Y_{2i} - Y_{2i-1} = 0 \ \& \ Y_{2i-1} - Y_{2i-2} = 0, \\ A, & \text{if } Y_{2i} - Y_{2i-1} = 1 \ \& \ Y_{2i-1} - Y_{2i-2} = -1. \end{cases} \tag{5}$$
$$, i = 1, 2, ..., n.$$

In this model, each point (X_j, Y_j) of GC-Curve is given. We can calculate each S_i according to this equation, and then connect each S_i one by one. So we can obtain the DNA sequence as $SEQ = S_1 S_2 ... S_i ... S_n$.

2.3 Advanced Properties of GC-Curve

In this subsection, using mathematic methods, we present several properties of GC-Curve.

Definition 1 (Degeneracy)**.** *There is one or more circuits in a graphical representation.*

Definition 2 (Loss of information)**.** *A graphical representation has loss of information, if readers cannot reconstruct the original DNA sequence from a graphical representation, because the correspondence between DNA sequences and graphical representation is not one to one.*

Property 1. *There is no degeneracy in GC-Curve.*

Proof. Reduction to absurdity:

Suppose that there is a circuit in GC-Curve. So there exist two points overlapping themselves. That is to say, $i \neq j$ must exist, making$(X_i, Y_i) = (X_j, Y_j)$. So $X_i = X_j$. According to the Equations (3) and (4), $X_i = i$ and $X_j = j$. Hence $i = j$. This contradicts $i \neq j$. Therefore, there is no circuit and degeneracy in GC-Curve. □

Property 2. *The correspondence between DNA sequences and GC-Curves is one to one, and GC-Curve has no loss of information.*

Proof. First, we prove that for a given GC-Curve there is a unique DNA sequence correspondingly.

Reduction to absurdity:

Suppose that corresponding to one GC-Curve, there exist two different DNA sequences, $SEQ1 = S1_1S1_2...S1_i...$ and $SEQ2 = S2_1S2_2...S2_i....$ So there must exist an m, making $S1_m \neq S2_m$. Since $(X1_0, Y1_0) = (X2_0, Y2_0) = (0, 0)$, and according to the Equations (1) and (2), a k must exist, making $Y1_{2k-1} \neq Y2_{2k-1}$ or $Y1_{2k} \neq Y2_{2k}$ and $0 < k \leq m$. It means that there are two different GC-Curves. This contradicts one given GC-Curve.

Second, we prove that for a given DNA sequence there is a unique GC-Curve correspondingly.

Reduction to absurdity:

Suppose that corresponding to one DNA sequence, there exist two different GC-Curves, Curve1 and Curve2. So a point $P1(X1_m, Y1_m)$ in Curve1 must exist, which is different to the point $P2(X2_m, Y2_m)$ in Curve2. That is to say, an m must exist, making $(X1_m, Y1_m) \neq (X2_m, Y2_m)$. By the Equations (3) and (4), $X1_m = m$ and $X2_m = m$. It implies that $X1_m = X2_m$. Therefore, it must be $Y1_m \neq Y2_m$. Since $(X1_0, Y1_0) = (X2_0, Y2_0) = (0, 0)$, there must exist a l, making $Y1_l - Y1_{l-1} \neq Y2_l - Y2_{l-1}$ and $0 < l \leq m$. According to the Equation (5), a k must exist, making $S1_k \neq S2_k$, where $2k = l$. It means that there must be two different DNA sequences. This contradicts one given DNA sequence.

So the correspondence between DNA sequence and GC-Curve is one to one, and there is no loss of information. □

Property 3. *The X-coordinate value of the end point indicates the length of a sequence. That is*

$$Length = X_{end}/2. \tag{6}$$

Proof. According to the Equation (4), $X_{end} = 2n = 2 * Length$. So $Length = X_{end}/2$. □

We can understand this result easily by Fig.1a. From Fig.1a, we can see that GC-Curve extends 2U along X axes to represent each nucleotide no matter A, T, C or G. So $Length = X_{end}/2$.

Property 4. *The Y-coordinate value of the end point indicates the number of DNA bases G+C of a sequence. That is:*

$$G + C = Y_{end}/2. \tag{7}$$

Proof. Substitute Equation (1) into Equation (2), we have

$$Y_{2i} = \begin{cases} Y_{2i-2} + 2, & \text{if } S_i = \text{G}, \\ Y_{2i-2} + 2, & \text{if } S_i = \text{C}, \\ Y_{2i-2}, & \text{if } S_i = \text{T}, \\ Y_{2i-2}, & \text{if } S_i = \text{A}. \end{cases}, i = 1, 2, ..., n. \tag{8}$$

We set a new variable F_i as follows:

$$F_i = \begin{cases} 1, & \text{if } S_i = \text{G or C}, \\ 0, & \text{if } S_i = \text{T or A}. \end{cases}, i = 1, 2, ..., n. \tag{9}$$

From Equation (9) and Equation (8), we can obtain

$$\begin{aligned} Y_{2i} &= \begin{cases} Y_{2i-2} + 2, & \text{if } S_i = \text{G}, \\ Y_{2i-2} + 2, & \text{if } S_i = \text{C}, \\ Y_{2i-2}, & \text{if } S_i = \text{T}, \\ Y_{2i-2}, & \text{if } S_i = \text{A}. \end{cases} \\ &= \begin{cases} Y_{2i-2} + 2 * 1, & \text{if } S_i = \text{G}, \\ Y_{2i-2} + 2 * 1, & \text{if } S_i = \text{C}, \\ Y_{2i-2} + 2 * 0, & \text{if } S_i = \text{T}, \\ Y_{2i-2} + 2 * 0, & \text{if } S_i = \text{A}. \end{cases} \\ &= \begin{cases} Y_{2i-2} + 2F_i, & \text{if } S_i = \text{G}, \\ Y_{2i-2} + 2F_i, & \text{if } S_i = \text{C}, \\ Y_{2i-2} + 2F_i, & \text{if } S_i = \text{T}, \\ Y_{2i-2} + 2F_i, & \text{if } S_i = \text{A}. \end{cases} \\ &= Y_{2i-2} + 2F_i \\ &= Y_{2(i-1)} + 2F_i. \end{aligned} \tag{10}$$

So

$$\begin{aligned} Y_{end} = Y_{2n} &= Y_{2(n-1)} + 2F_n \\ &= Y_{2(n-2)} + 2F_{n-1} + 2F_n \\ &= Y_{2(n-2)} + 2(F_{n-1} + F_n) \\ &= ... \\ &= Y_0 + 2(1 + 2 + 3 + ... + F_{n-1} + F_n) \\ &= 0 + 2(1 + 2 + 3 + ... + F_{n-1} + F_n) \\ &= 2(G + C). \end{aligned} \tag{11}$$

Finally, we get

$$G + C = Y_{end}/2. \tag{12}$$

□

We can understand this result easily by Fig.1a. As can be seen in Fig.1a, GC-Curve extends 2U along Y axes to represent each nucleotide of C and G, while the Y-coordinate value add nothing to represent A and T. So $G + C = Y_{end}/2$.

Property 5. *The end point indicates the GC-content of a DNA sequence. That is:*

$$GC\ content = Y_{end}/X_{end} = tan\theta. \tag{13}$$

Proof. According to the Equation (6) and Equation (7), we can get

$$\begin{aligned} GC\ content &= \frac{G+C}{A+G+C+T} \\ &= \frac{G+C}{Length} \\ &= \frac{Y_{end}/2}{X_{end}/2} \\ &= Y_{end}/X_{end} \\ &= tan\theta. \end{aligned} \tag{14}$$

□

3 Applications of GC-Curve

In this section, we present several applications of GC-Curve. The first one is similarity analysis based on numerical descriptor. And the second one is sequence analysis based on a visual inspection of GC-Curves, which includes GC-content analysis, stability analysis, and melting temperature prediction. Finally, a corresponding software of GC-Curve is given.

3.1 Similarity analysis based on numerical descriptor

Numerical descriptor comparison is a kind of similarity analysis of DNA sequences, proposed by Randić *et al.* [24]. In this method, numerical descriptor is extracted from a visualization of DNA sequence, then is used to do quantitative analysis by comparing. It was proved to be effective and used by many authors. Some good numerical descriptors have been proposed, such

Table 1: The relationships of the end points of the second vector to represent four DNA bases. (From Fig.1(a))

DNA base	DNA base	relationship
G	T	different
G	C	the same
G	A	different
C	T	different
T	A	the same
C	A	different

G-C and T-A are the same, others are different. However, actually, all the relationships between two different DNA bases should be different. The end point of the second vector to represent a DNA base will save at the even points of GC-Curve according to the DNA walk technology. So “the same” information saving in the even points of GC-Curve is noise when we do similarity analysis. Therefore, we remove all even points when we characterize a GC-Curve.

Table 2: The relationships of the end points of the first vector to represent four DNA bases. (From Fig.1(a))

DNA base	DNA base	relationship
G	T	different
G	C	different
G	A	different
C	T	different
T	A	different
C	A	different

They are all different. The end point of the first vector to represent a DNA base will save at the odd points of GC-Curve according to the DNA walk technology. It means the relationships at the odd points in GC-Curve among DNA bases are all different. So the odd points of GC-Curve can be used to characterize a GC-Curve.

as [23, 24, 45].Here, we adopt the numerical descriptor similar to [23], because it is reasonable, simple, and very fast to be calculated.

From the representations of four DNA bases (see Fig. 1a), we know that the end point (2,2) of the second vector to represent ‘G’ is different with the end point (2,0) of the second vector to represent ‘T’, but is the same with the end point (2,2) of the second vector to represent ‘C’. As can be seen in Table 1, there are two “the same” relationships. However, actually, all the relationships between two different DNA bases should be different. So “the same” information saving in the even points of GC-Curve is noise when we do similarity analysis. Therefore, we remove all even points when we characterize a GC-Curve.

From the representations of four DNA bases (see Fig. 1a), we know that the relationships of the end points of the first vector to represent four bases are all different (Table 2). It means the relationships at the odd points in GC-Curve among DNA bases are all different. So we use the odd points to characterize a GC-Curve. We calculate M_{xy} (like covariance) as follows:

$$(X_c, Y_c) = (\frac{1}{n}\sum_{j=1}^{n} X_{2j-1}, \frac{1}{n}\sum_{j=1}^{n} Y_{2j-1}). \tag{15}$$

$$M_{xy} = \frac{1}{n}\sum_{j=1}^{n}(X_{2j-1} - X_c)(Y_{2j-1} - Y_c). \tag{16}$$

For one sequence, we can get 24 different graphs by assigning A, T, C, G to four basic dualvectors in $4! = 24$ different ways. For each graph, we can get one M_{xy}. So we can obtain a 24-dimensional vector $\vec{D}$ (see Equation (17)) as the numerical descriptor of a sequence. We call this method, using 24-dimensional vector from all the 24 assignments as the numerical descriptor [23], as 24-enumerate technology. Using 24-enumerate technology, we can avoid the arbitrariness of DNA bases assignment.

$$\vec{D} = [M1_{xy}, M2_{xy}, ..., M24_{xy}]. \tag{17}$$

Suppose that there are two sequences i and j, and the corresponding descriptors are $\vec{D}_i$ and $\vec{D}_j$ respectively. The similarity between these two sequences can be calculated by Euclidean distance:

$$d_{ij} = \|\vec{D}_i - \vec{D}_j\|. \tag{18}$$

The smaller Euclidean distance is, the more similar the DNA sequences are. That is to say, the distances between evolutionary closely related species are smaller, while those between evolutionary disparate species are larger.

The method is illustrated on the coding sequences of the first exon of beta-globin gene of 11 species. The similarity result is shown in Table 3. From Table 3, we find that Human-Gorilla and Goat-Bovine are the most similar. It is reasonable and consistent with other publications. As shown in Table 4, we list the similarities between human and several species in current publications. It shows that Human-Gorilla is the most similar whereas Human-Gorilla, Human-Gallus, Human-Opossum, Human-Bovine are the most different. This is consistent with our results.

Table 3: The similarity result (1.0e + 3) for the coding sequences of the first exon of beta-globin gene of 10 species based on the Euclidean distances between the end points of the descriptor vectors $\vec{D}$.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine
Human	0	0.934	0.833	0.820	0.344	0.435	0.477	0.364	0.241	0.908
Goat		0	1.242	1.097	0.975	1.312	0.626	0.965	1.051	0.149
Opossum			0	0.918	0.871	0.737	1.080	0.668	0.980	1.214
Gallus				0	1.047	0.903	0.955	0.929	0.800	1.171
Lemur					0	0.626	0.445	0.213	0.435	0.919
Mouse						0	0.904	0.572	0.506	1.284
Rabbit							0	0.521	0.520	0.609
Rat								0	0.485	0.918
Gorilla									0	1.047
Bovine										0

Table 4: The similarity between human and other species.

Methods	Gorilla	Gallus	Opossum	Bovine	Goat	Lemur	Mouse	Rabbit	Rat
Our work, Table 3	0.241	0.820	0.833	0.908	0.934	0.344	0.435	0.477	0.364
Zhang [37], Table 1	1.568	5.417	4.631	4.867	4.777	2.764	3.184	3.192	2.622
Xie *et al.* [16], Table 3	0.042	1.148	0.647	0.074	0.079	0.525	1.49	0.376	1.100
Zhang [23], Table 1	0.263	1.156	1.186	0.361	0.477	0.500	0.444	0.535	0.527
Yao *et al.* [40], Table 10	0.005	0.029	0.030	0.014	0.016	0.013	0.017	0.011	0.012
Liao *et al.* [46],Table 5	0.026	0.106	0.096	0.049	0.052	0.064	0.031	0.051	0.049
Liu *et al.* [39], Table 5	0.008	0.242	0.282	0.075	0.108	0.176	0.076	0.102	0.097

It shows that Human-Gorilla is the most similar whereas Human-Gallus, Human-Opossum, Human-Bovine, Human-Goat are much more different. This is consistent with our results.

3.2 Sequence analysis based on a visual inspection of GC-Curves

The most important feature of GC-Curve is that GC-Curve can reflect the GC-content of a DNA. GC-content is an important feature of DNA, which is related to the stability of DNA [41], the density of genes [42], natural selection [43, 44], mutational bias [43], etc. So GC-Curve has great potential in many applications. In this subsection, we will use this feature to do some analyses based on a visual inspection of GC-Curves, which includes GC-content analysis, stability analysis, and melting temperature prediction.

3.2.1 By observing GC-Curve, we can know easily the GC-content of a DNA sequence, and also know the variations in GC-content within a DNA sequence.

As shown in Fig. 2, the end point of human approximates to the point (900,500). According to the Equation (13), the GC-content of the human sequence is about $500/900 = 55.5\%$. The end point of opossum approximates to the point (900,350). Analogously, the GC-content of the

opossum sequence is about $350/900 = 38.8\%$. The GC-Curve of human grows very steady and almost in a straight line, so GC-contents within the human sequence are very uniform. The GC-Curve of opossum is not as steady as human, and indicates that GC-contents within the opossum sequence change a lot. The GC-Curve of opossum grows slowly from point 'A' to point 'B', while the GC-Curve of opossum grows faster from point 'B' to point 'C'. It indicates that the GC-content of the fragment from 'A' to 'B' is higher than the GC-content of the fragment from 'B' to 'C'.

3.2.2 By a visual inspection of GC-Curve, we can do some stability analysis of a DNA sequence.

From molecular biology, we know that the GC pair is bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. So DNA with high GC-content was believed to be more stable than DNA with low GC-content. However, Yakovchuk *et al.* [41] showed that the hydrogen bonds is not the main stabilizing factor in the DNA double helix. So here we say "DNA with high GC-content is possibly more stable than DNA with low GC-content", and use it to do analysis. As can be seen in Fig. 2, the GC-Curve of human grows faster than the GC-Curve of Opossum. The GC-Curve of human is at the top of the opossum curve. It indicates that the complete coding sequence of beta-globin gene of human is possibly more stable than opossum.

3.2.3 By observing GC-Curve, we can predict the melting temperature (T_m) of a DNA.

T_m is defined as the temperature at which half of the DNA strands are in the random coil or single-stranded state. T_m is very important in nucleic acid thermodynamics, and can be predicted by following equation [47]

$$T_m = 64.9 + \frac{41 * (G + C - 16.4)}{A + T + G + C}. \tag{19}$$

The equation can be reduced as

$$\begin{aligned} T_m &= 64.9 + \frac{41 * (G + C - 16.4)}{A + T + G + C} \\ &= 64.9 + 41 * \frac{G + C}{A + T + G + C} - \frac{41 * 16.4}{A + T + G + C} \\ &= 64.9 + 41 * GC\ content - \frac{41 * 16.4}{Length}. \end{aligned} \tag{20}$$

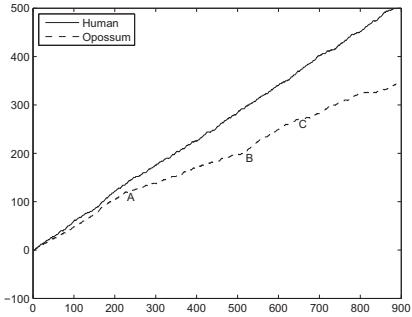

Figure 2: The GC-Curves of the complete coding sequences of beta-globin genes of human (solid) and opossum (dashed). The end point of human approximates to the point (900,500). According to the Equation (13), the GC-content of the human sequence is about $500/900 = 55.5\%$. The GC-Curve of the human sequence grows very steady and almost in a straight line, while the GC-Curve of the opossum sequence is very flexural. It indicates that the GC-contents within the human sequence are very uniform whereas the GC-contents within the opossum sequence change a lot. The GC-Curve of human is at the top of the opossum curve. It indicates that the GC-content of the human sequence is higher than the opossum sequence, and that the human sequence is possibly more stable than the opossum sequence, and that the human sequence has a higher melting temperature than the opossum sequence.

In most cases, DNA sequence is long. So we have

$$\frac{41 * 16.4}{Length} \approx 0. \tag{21}$$

So T_m can be predicted by GC-content as

$$T_m \approx 64.9 + 41 * GC\ content. \tag{22}$$

From the Equation (22), we know that the DNA with higher GC-content has higher melting temperature. As shown in Fig. 2, the GC-Curve of human is at the top of the opossum curve. It indicates that the human sequence has higher melting temperature than the opossum sequence. The end point of human approximates to the point (900,500). According to the Equation (22), we can predict that $T_m \approx 87.7$.

3.3 The software of GC-Curve

To facilitate biologists, we have developed a software of GC-Curve. As can be seen in Fig. 3 (top), the software is very simple and very easy to use. Input a DNA sequence and its name, then click “Draw”. The corresponding GC-Curve will be drawn immediately. Fig.

3 (bottom) shows three sequences in a window. The software of GC-Curve is available at https://www.box.com/s/g872v3pq4kuz86sj5coq.

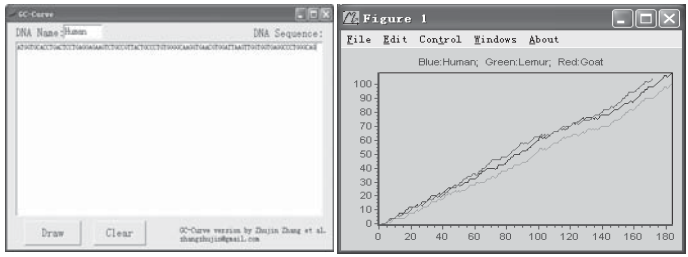

Figure 3: The software of GC-Curve (left). Three DNA sequences in the same window (right).

4 Discussion

Visualization of DNA sequences is an important means for sequence analysis, and provides useful insights into local and global characteristics of a sequence, which are not easily obtainable by other methods [1]. Significant progresses have been made in the past three decades for visualizing DNA sequences by solving the problems of degeneracy, loss of information, difficulty in multi-dimensional space and difficulty of observing when showing long DNA sequences. However, a question is whether a visualization tool is able to visualize some biological features in addition visual effect.

In this paper, different from traditional methods focusing on visual effect, we propose a novel visualization tool—GC-Curve, which not only solves all the visual effect problems mentioned above, but also can reflect the GC-content of a DNA. We also build two mathematics models for GC-Curve, and present properties of GC-Curve.

GC-content is an important feature of DNA, and it is related to the stability of DNA [41], the density of genes [42], natural selection [43,44], mutational bias [43], etc. So, GC-Curve has great potential in many applications. In this paper, applications of GC-Curve on similarity analysis, GC-content analysis, stability analysis, and melting temperature prediction are presented.

GC-Curve is the first visualization model which can reflect GC-content. GC-content is related to many important features of DNA, and GC-Curve has great potential in many applications.

Acknowledgment: The work was supported by National Natural Science Foundation of China (61272385,

61202011, 61033003), China Postdoctoral Science Foundation (2012M511485, 2013T60374), and Shenzhen Strategic Emerging Industries Program (ZDSY20120613125016389).

References

[1] Y. Zhang, B. Liao, K. Ding, On 2D graphical representation of DNA sequence of nondegeneracy, *Chem. Phys. Lett.* **411** (2005) 28–32.

[2] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–27.

[3] E. Hamori, Novel DNA sequence representations, *Nature* **314** (1985) 585–586.

[4] M. A. Gates, Simpler DNA sequence representations, *Nature* **316** (1985) 219–219.

[5] M. Randić, J. Zupan, A. T. Balaban, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins, *Chem. Rev.* **111** (2011) 790–862.

[6] Z. Zhang, L. Liu, J. Li, Z. Zhang, Spectral Representation of Protein Sequences, *J. Comput. Theor. Nanosci.* **8** (2011) 1335–1339.

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

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

[9] R. Zhang, C. T. Zhang, Z curves, an intutive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.* **11** (1994) 767–782.

[10] C. T. Zhang, R. Zhang, H. Y. Ou, The Z curve database: a graphic representation of genome sequences, *Bioinformatics* **19** (2003) 593–599.

[11] Y. Yao, X. Nan, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, *Chem. Phys. Lett.* **411** (2005) 248–255.

[12] Z. H. Qi, T. R. Fan, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434–440.

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

[14] Z. Cao, B. Liao, R. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, *Int. J. Quantum Chem.* **108** (2008) 1485–1490.

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

[16] G. Xie, Z. Mo, Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications, *J. Theor. Biol.* **269** (2011) 123–130.

[17] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 493–512.

[18] Z. Cao, R. Li, W. Chen, A 3D graphical representation of DNA sequence based on numerical coding method, *Int. J. Quantum Chem.* **110** (2010) 975–980.

[19] Y. Huang, T. Wang, New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis, *Int. J. Quantum Chem.* **112** (2012) 1746–1757.

[20] B. Liao, M. Tan, K. Ding, A 4D representation of DNA sequences and its application, *Chem. Phys. Lett.* **402** (2005) 380–383.

[21] R. Chi, K. Ding, Novel 4D numerical representation of DNA sequences, *Chem. Phys. Lett.* **407** (2005) 63–67.

[22] X. C. Tang, P. P. Zhou, W. Y. Qiu, On the similarity/dissimilarity of DNA sequences based on 4D graphical representation, *Chinese Sci. Bull.* **55** (2010) 701–704.

[23] Z. Zhang, DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences, *Bioinformatics* **25** (2009) 1112–1117.

[24] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1–6.

[25] Z. Qi, X. Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides, *Chem. Phys. Lett.* **440** (2007) 139–144.

[26] Z. Zhang, L. Liu, J. Li, Z. Zhang, Spectral representation of DNA sequences and its application, in: *Bio-Inspired Computing: Theories and Applications (BIC-TA), The IEEE Fifth International Conference*, IEEE, 2010, pp. 1023–1027.

[27] D. Bielińska-Wąż, Four–component spectral representation of DNA sequences, *J. Math. Chem.* **47** (2010) 41–51.

[28] D. Bielińska-Wąż, S. Subramaniam, Classification studies based on a spectral representation of DNA, *J. Theor. Biol.* **266** (2010) 667–674.

[29] S. T. Y. Stephen, J. Wang, A. Niknejad, C. Lu, N. Jin, Y. K. Ho, DNA sequence representation without degeneracy, *Nucleic Acids Res.* **31** (2003) 3078–3080.

[30] G. Huang, B. Liao, Y. Li, Z. Liu, H-L curve: A novel 2D graphical representation for DNA sequences, *Chem. Phys. Lett.* **462** (2008) 129–132.

[31] M. Randić, M. Vračko, J. Zupan, M. Novič, Compact 2-D graphical representation of DNA, *Chem. Phys. Lett.* **373** (2003) 558–562.

[32] Z. H. Qi, L. Li, X. Q. Qi, Using Huffman coding method to visualize and analyze DNA sequences, *J. Comput. Chem.* **32** (2011) 3233–3240.

[33] M. Randić, Graphical representations of DNA as 2-D map, *Chem. Phys. Lett.* **386** (2004) 468–471.

[34] M. Randić, N. Lerš, D. Plavšić, S. C. Basak, A. T. Balaban, Four–color map representation of DNA or RNA sequences and their numerical characterization, *Chem. Phys. Lett.* **407** (2005) 205–208.

[35] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.* **18** (1990) 2163–2170.

[36] Z. Zhang, X. Zeng, T. Song, Z. Chen, WormStep: an improved compact graphical representation of DNA sequences based on worm curve, *J. Comput. Theor. Nanosci.* **10** (2013) 189–193.

[37] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L. Pan, ColorSquare: a colorful square visualization of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **68** (2012) 621–637.

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

[39] X. Q. Liu, Q. Dai, Z. Xiu, T. Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its application, *J. Theor. Biol.* **243** (2006) 555–561.

[40] Y. Yao, Q. Dai, X. Y. Nan, P. A. He, Z. M. Nie, S. P. Zhou, Y. Z. Zhang, Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation, *J. Comput. Chem.* **29** (2008) 1632–1639.

[41] P. Yakovchuk, E. Protozanova, M. D. Frank–Kamenetskii, Base–stacking and base–pairing contributions into thermal stability of the DNA double helix, *Nucleic Acids Res.* **34** (2006) 564–574.

[42] G. Bernardi, The vertebrate genome: isochores and evolution, *Mol. Biol. Evol.* **10** (1993) 186–204.

[43] J. A. Birdsell, Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution, *Mol. Biol. Evol.* **19** (2002) 1181–1197.

[44] U. Pozzoli, G. Menozzi, M. Fumagalli, M. Cereda, G. Comi, R. Cagliani, N. Bresolin, M. Sironi, Both selective and neutral processes drive GC content evolution in the human genome, *BMC Evol. Biol.* **8** (2008) #99: 1–12.

[45] I. Pesek, J. Žerovnik, A numerical characterization of modified Hamori curve representation of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 301–312.

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

[47] R. Bruce Wallace, J. Shaffer, R. F. Murphy, J. Bonner, T. Hirose, K. Itakura, Hybridization of synthetic oligodeoxyribonucleotides to ΦX 174 DNA: the effect of single base pair mismatch, *Nucleic Acids Res.* **6** (1979) 3543–3558.