MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **68** (2012) 621-637

ISSN 0340 - 6253

ColorSquare: A Colorful Square Visualization of DNA Sequences

Zhujin Zhang[a,1] , Tao Song[b], Xiangxiang Zeng[c], Yunyun Niu[b], Yun Jiang[d], Linqiang Pan[b], Yunming Ye[a]

[a] *Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China*

[b] *Key Laboratory of Image Processing and Intelligent Control Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*

[c] *Department of Computer Science, Xiamen University, Xiamen 361005, China*

[d] *School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067, China*

(Received November 23, 2011)

Abstract

Visualization tool provides a simple way of viewing and analyzing DNA sequences. Here we propose a five-color map visualization of DNA sequences — ColorSquare. ColorSquare has several advantages: (1) no degeneracy, (2) no loss of information, (3) highly compact, (4) colorful, and (5) square. Due to square, it can be converted into a matrix, from which numerical characterizations can easily be extracted. Moreover, using the feature that human eyes are more sensitive to color than to shape, we proposed a new mutation analysis based on ColorSquare. It shows that colorful visualization tool is more effective for human than visualization tool based on shape. Similarity analyses based on two kinds of numerical characterizations are also presented.

1 Introduction

The rapid growth in available DNA sequence data creates a great need of viewing and analyzing DNA sequences. Graphical representation is considered as visualization tool

[1]Corresponding author. Email: zhangzhujin@gmail.com

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

Since H curve, the first graphical representation of DNA sequences, was proposed by Hamori and Ruskin [2, 3], a number of different graphical representations have been introduced. In the early phase, Gates [4] designed an important visualization in 2D space, using four orthogonal directions to represent the four bases. The work was followed by Nandy [5], Leong and Mogenthaler [6]. However, these visualizations are accompanied by high degeneracy and loss of information.

Degeneracy and loss of information became two main barriers of DNA graphical representations [7]. Many researchers have made great efforts to solve these two problems. Guo *et al.* [8] built a low degeneracy representation. Wu *et al.* [9] introduced a representation with non-degeneracy, but with loss of information. Qi and Fan [10], Zhang and Zhang [11, 12], Xie and Mo [13], Qi *et al.* [14] used 3D graphical representations. Liao *et al.* [15], Tang *et al.* [16], Chi and Ding [17] adopted 4D approaches. Qi and Qi [18], Huang and Wang [19], Yu *et al.* [20], Cao *et al.* [21]designed dinucleotide models. Yu *et al.* [22], Liao and Wang [23] introduced trinucleotide representations. Bielińska-Wąż and Subramaniam [24, 25], Zhang *et al.* [26], Randić *et al.* [27] adopted spectral representations. Zhang [28] designed a dual-vector model. Randić *et al.* [29], Qi *et al.* [30], Cao *et al.* [31] provided coding methods.

Besides degeneracy and loss of information, most representations need a lot of space. Jeffrey [32] designed a compact representation, which needs limited space to represent long sequences. This representation avoids loss of information, but it still has degeneracy. Randić *et al.* [29], Zhang *et al.* [33] proposed compact graphical representations, avoiding degeneracy and loss of information. Besides these three advantages, Randić *et al.* [34] added a colorful advantage with a four-color map representation.

Several important applications of graphical representation are based on numerical characterizations, for example, similarity analysis, classification, phylogenetic tree study, and so on. How to fast and easily obtain numerical characterizations from a visualization is another important topic. As the literature [25] said, finding a proper balance between fast numerical identification of the sequences and good visualization became a subject of many recent studies, such as the works [14,25,28,35–42]. In this paper, we find an excellent solution for these two goals — very good visualization effect and very easily numerical identification. In particular, we improve the colorful model [34] with a new visualization

— ColorSquare. ColorSquare not only has the four advantages mentioned above, but also has an unique advantage — it is square, and can be converted into a matrix, from which numerical characterizations can easily be extracted. Moreover, we proposed a new mutation analysis based on ColorSquare, and show that colorful visualization tool is more effective for human than visualization tool based on shape. Similarity analyses based on two kinds of numerical characterizations are also presented.

2 The ColorSquare Visualization

We propose a five-color map visualization of DNA sequences — ColorSquare. In this section, we outline the construction of ColorSquare. With figures and examples, readers can understand easily about it. Then we present the algorithm for ColorSquare, and describe several advanced properties of ColorSquare.

2.1 Construction of ColorSquare

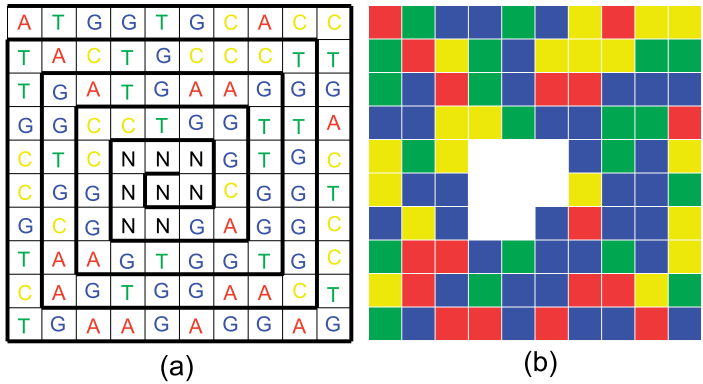

Figure 1: (a) The whirlpool construction of ColorSquare of the first exon of human β-globin gene. (b) The visualization Result of ColorSquare of the first exon of human β-globin gene.

We hope that ColorSquare can be highly compact, using limited space to view long DNA sequences. Inspired by the whirlpool, we designed a two-dimension vortex structure as Fig. 1. The details are given as follows:

Firstly, we decide to represent DNA bases by small squares with different colors. One square represents one DNA base, so we can know how many small squares we need to represent a given DNA sequence. Suppose that there is a given DNA sequence with

the length of n nucleobases, we need n small squares at least. In order to make the visualization be highly compact, we merge these n small squares into a big square, called Big Square. Big Square has a side with the length of k.

$$k = \lceil \sqrt{n} \rceil \tag{1}$$

where '$\lceil \ \rceil$' means 'round up'.

After getting Big Square, we can mark these small squares. We mark clockwise around in Big Square according to the given DNA sequence. Because $k \times k$ is generally greater than n, Big Square contains small squares more than n. It means that there are some small squares left which don't need to represent DNA bases. We mark these remaining squares with 'N'.

After marking Big Square, we fill the squares according to the assignments as follows:

$$\begin{array}{l} \text{A} \Longrightarrow \text{Red} \\ \text{G} \Longrightarrow \text{Blue} \\ \text{C} \Longrightarrow \text{Yellow} \\ \text{T} \Longrightarrow \text{Green} \\ \text{N} \Longrightarrow \text{White} \end{array}$$

Finally we get the ColorSquare visualization.

Here we take the sequence of the first exon of human β-globin gene as example to illustrate how to construct ColorSquare. The result is shown in Fig. 1. The length of the given sequence is 92. According to Equation (1), we can get $k = \lceil \sqrt{92} \rceil = 10$. So we create Big Square with 10×10 small squares. As shown in Fig. 1 (a), we start at square(1,1). The first DNA base is 'A', so we mark square(1,1) with 'A'. We mark clockwise around in Big Square, the next square is square(1,2), so we mark the square(1,2) with 'T' because the second base is 'T'. We move right and mark the small squares, according to the given sequence, till reaching the end of first row. We mark clockwise, so the 11th square is square(2,10). The 11th DNA base is 'T', so the square(2,10) is marked as 'T'. Then we move down in tenth column till we reach square(10,10). We mark clockwise around in Big Square. The next square is square(9,10), we mark the square(9,10) with 'A', because the second base is 'A'. In the same way, we can mark the whole DNA sequence. But there are some squares left. We mark the remaining squares with 'N'. Finally, we fill all the squares according to the color assignments, and get the ColorSquare visualization as Fig. 1 (b).

In short, there are four steps in the strategy of ColorSquare visualization:

Step 1: Calculate k, the length of the side of big square, according to the length of the given DNA sequence.

Step 2: Create Big Square containing $k \times k$ small squares.

Step 3: Mark clockwise around in Big Square according to the given DNA sequence.

Step 4: Fill all the squares according to the color assignments, and get ColorSquare visualization.

2.2 Algorithm of the ColorSquare

According to the construction, we build an algorithm for ColorSquare. As shown in Algorithm 1, we use 'Status' to mark the moving direction, so that we can judge when we need to turn right and when we need to turn left. After solving this problem, we can easily get the ColorSquare visualization. The time complexity of Algorithm 1 is $O(n)$, where n is the length of the DNA sequence.

2.3 Advanced Properties of ColorSquare

In this subsection, we present four advanced properties of ColorSquare. The details are given as follows:

Property 2.1 *ColorSquare avoids degeneracy.*

Degeneracy is a basic problem in graphical representation of DNA sequences. Every graphical representation needs to face and conquer. In the construction of ColorSquare, each small square represents one DNA base, and there is no DNA base overlapping each other. So there is no circuit in ColorSquare, and ColorSquare avoids degeneracy.

Property 2.2 *ColorSquare avoids loss of information.*

Loss of information is another basic problem in graphical representation of DNA sequences. If observers can not reconstruct the corresponding DNA sequence from a visualization, this visualization will lose information. In our work, as shown in Fig. 1 (b), the first square is red, so the first base in the sequence is 'A'. The second square is Green, so

Algorithm 1: ColorSquare()

Input:
SEQ — The given DNA sequence
Output:
G — ColorSquare visualization

```
N = ⌈sqrt(SEQ.Length)⌉; /* Calculate k
Square = White(N,N); /*Create Big Square
                        containing k × k small squares */
Status = 'Right';
Position = (0,1);
for i=1:SEQ.Length
    if (Status='Right')
        if (Position(2)+1 > N or Square(Position(1), Position(2)+1) ≠ White)
            Status = 'Down';
            Position(1)=Position(1)+1;
        else
            Position(2)=Position(2)+1;
        end
    else if (Status='Down')
        if (Position(1)+1 > N or Square(Position(1)+1, Position(2)) ≠ White)
            Status = 'Left';
            Position(2)=Position(2)-1;
        else
            Position(1)=Position(1)+1;
        end
    else if (Status='Left')
        if (Position(2)-1 < 1 or Square(Position(1), Position(2)-1) ≠ White)
            Status = 'Up';
            Position(1)=Position(1)-1;
        else
            Position(2)=Position(2)-1;
        end
    else if (Status='Up')
        if (Position(1)-1 < 1 or Square(Position(1)-1, Position(2)) ≠ White)
            Status = 'Right';
            Position(2)=Position(2)+1;
        else
            Position(1)=Position(1)-1;
        end
    end
    Square(Position(1),Position(2))= SEQ(i).Color;
end
G = Square;
```

the second base in the sequence is 'T'. The third square is Blue, so the third base in the sequence is 'G'. Continue the process, we can get the whole DNA sequence. Therefore, it is easy for observers to reconstruct corresponding DNA sequences from ColorSquare, and

ColorSquare avoids loss of information.

Property 2.3 *ColorSquare is highly compact.*

ColorSquare can visualize long DNA sequences in a limited space. For a DNA sequence with the length of n nucleobases, ColorSquare needs a square space with sides approximately given by $\sqrt{n}$. For example, the DNA sequence of the first β-globin exon is 92 nucleobases. According to Equation (1), we can get $k = \lceil \sqrt{92} \rceil = 10$. It only costs a 10×10 square. The complete DNA sequence of human globin with over 1400 bases will fill a square of size 38×38. And thus, the ColorSquare is highly compact.

Property 2.4 *ColorSquare is colorful, and more convenient to be observed.*

Human eyes are more sensitive to color than to shape. Humans can quickly find out a small difference in color in two pictures. The mutation analysis based on ColorSquare, using this fact, is presented in section 4. So colorful visualization tools are more effective for human.

Property 2.5 *ColorSquare is square, and can be converted into a matrix, from which numerical characterizations can easily be extracted.*

ColorSquare is a square whatever the sequence is long or short. So we can converted it into a matrix. Since that matrix is an useful tool in mathematics, we can easily extract the numerical characterizations form a matrix. Similarity analysis based on numerical characterizations can be easier. The details are given in the next section.

3 Matrix Representation and Numerical Characterizations of ColorSquare

Similarity analysis based on numerical characterizations of a visualization is an important method in sequence analysis. It was proposed by Randić et al. [27,43]. They [27] proposed E matrix, M/M matrix, L/L matrix and L^k/L^k matrix, then used the eigenvalues of these matrices as numerical characterizations of visualizations. However, these matrices need to calculate all the distance between two points, and the matrices are as large as $n \times n$, where n is the length of a DNA sequence.

Here our matrix representation has two advantages. First, we save a lot of computation. Our matrix representation is very intuitive without complex computation. Second, we save a lot of space. Our matrix with the size of $\sqrt{n} \times \sqrt{n}$ is much smaller than others.

All these advantages are attributed to the fact that ColorSquare is square, and can be converted into be a matrix. In this section, we will introduce how to convert ColorSquare into the matrix representation, and how to get numerical characterizations of ColorSquare.

3.1 Matrix Representation of ColorSquare

ColorSquare is highly compact and square. By number assignment, ColorSquare can be converted into a matrix with the size of $\sqrt{n}\times\sqrt{n}$, where n is the length of a DNA sequence. At first we need to set the number assignment as follows:

$$\begin{array}{ll} \mathrm{N} \Longrightarrow \mathrm{White} & \Longrightarrow 0 \\ \mathrm{A} \Longrightarrow \mathrm{Red} & \Longrightarrow 1 \\ \mathrm{C} \Longrightarrow \mathrm{Yellow} & \Longrightarrow 2 \\ \mathrm{G} \Longrightarrow \mathrm{Blue} & \Longrightarrow 3 \\ \mathrm{T} \Longrightarrow \mathrm{Green} & \Longrightarrow 4 \end{array}$$

According to the number assignments above, every small square in ColorSquare can be converted into number, and the whole Big Square automatically is converted into a matrix. For example, Fig. 1 (b) can be converted into a matrix.

$$\begin{pmatrix} 1 & 4 & 3 & 3 & 4 & 3 & 2 & 1 & 2 & 2 \\ 4 & 1 & 2 & 4 & 3 & 2 & 2 & 2 & 4 & 4 \\ 4 & 3 & 1 & 4 & 4 & 1 & 1 & 3 & 3 & 3 \\ 3 & 3 & 2 & 2 & 4 & 3 & 3 & 4 & 4 & 1 \\ 2 & 4 & 2 & 0 & 0 & 0 & 3 & 4 & 3 & 2 \\ 2 & 3 & 3 & 0 & 0 & 0 & 2 & 3 & 3 & 4 \\ 3 & 2 & 3 & 0 & 0 & 3 & 1 & 3 & 3 & 2 \\ 4 & 1 & 1 & 3 & 4 & 3 & 3 & 4 & 3 & 2 \\ 2 & 1 & 3 & 4 & 3 & 3 & 1 & 1 & 2 & 4 \\ 4 & 3 & 1 & 1 & 3 & 1 & 3 & 3 & 1 & 3 \end{pmatrix}$$

Figure 2: The matrix representation of ColorSquare (Fig. 1 (b)) of the sequence of the first exon of human β-globin gene.

3.2 Numerical Characterizations of ColorSquare

Given the matrix representation of ColorSquare, we can easily get the numerical characterizations. In this subsection, we propose two numerical characterizations of ColorSquare. The details are presented.

3.2.1 24-Component Vector Based on Matrix Eigenvalue

In mathematics, leading eigenvalue is an important numerical characterization of a matrix, which has been successfully used by Randić et al. [27, 43]. We also use the leading eigenvalue, and mark it as M.

For a sequence, we can get 4! = 24 different matrix representations by assigning A, T, C, G to different numbers in 4! different ways. For example, we can set the number assignments above. We can also set the number assignments as follows:

$$\begin{array}{l} \text{N} \Longrightarrow \text{White} \Longrightarrow 0 \\ \text{A} \Longrightarrow \text{Red} \Longrightarrow 2 \\ \text{C} \Longrightarrow \text{Yellow} \Longrightarrow 1 \\ \text{G} \Longrightarrow \text{Blue} \Longrightarrow 3 \\ \text{T} \Longrightarrow \text{Green} \Longrightarrow 4 \end{array}$$

Similar to Zhang [28], we adopt 24-component vector $\vec{D}$ as the numerical characterizations of a DNA sequence:

$$\overrightarrow{DM} = [M1, M2, ..., M24] \tag{2}$$

3.2.2 96k-Component Vector Based on Parity Check Code

Parity check code was firstly used in signal transport, and now is widely applied in industry, such as bar code verification. We use parity check code in both rows and columns as the numerical characterizations of a matrix. First we need to define some common descriptions and variables which are efficient in the whole article:

(1) Matrix representation of ColorSquare: $A_{k\times k}$.

(2) $C_{(a,b)}$ is the number of A in a-th row, b-th column.

(3) α, β,γ,δ are odd check code in row, even check code in row, odd check code in column, even check code in column respectively.

Odd check code in row:

$$\alpha_i = \sum_{1\le j\le k/2} C_{(i,2j-1)}, i = 1, 2, ..., k \tag{3}$$

Even check code in row:

$$\beta_i = \sum_{1\le j\le k/2} C_{(i,2j)}, i = 1, 2, ..., k \tag{4}$$

Odd check code in column:

$$\gamma_j = \sum_{1\le i\le k/2} C_{(2i-1,j)}, j = 1, 2, ..., k \tag{5}$$

Even check code in column:

$$\delta_j = \sum_{1\le i\le k/2} C_{(2i,j)}, j = 1, 2, ..., k \tag{6}$$

	12	14	12	11	11	10	8	12	13	13	
12	1	4	3	3	4	3	2	1	2	2	13
15	4	1	2	4	3	2	2	2	4	4	13
13	4	3	1	4	4	1	1	3	3	3	14
16	3	3	2	2	4	3	3	4	4	1	13
10	2	4	2	0	0	0	3	4	3	2	10
10	2	3	3	0	0	0	2	3	3	4	10
10	3	2	3	0	0	3	1	3	3	2	10
15	4	1	1	3	4	3	3	4	3	2	13
11	2	1	3	4	3	3	1	1	2	4	13
12	4	3	1	1	3	1	3	3	1	3	11
	17	11	9	10	14	9	13	16	15	14	

Figure 3: The parity check code of the matrix representation of the sequence of the first exon of human β-globin gene. The column in left is the odd check code in rows; The column in right is the even check code in rows; The row on top is the odd check code in columns; The row at bottom is the even check code in columns.

We illustrate these parity check codes on a matrix representation of the sequence of the first exon of human β-globin gene. (Fig. 2)

As shown in Fig.2, the numbers in the first row in odd column are 1,3,4,2,2. According to the Equation (3), we can get:

$$\begin{aligned} \alpha_1 &= 1+3+4+2+2 \\ &= 12 \end{aligned} \tag{7}$$

So in Fig.3 the first element of the column in left is 12. After getting all the parity check codes, we can merge them as vector $\vec{P}$:

$$\begin{aligned} \vec{P} = [&\alpha_1, \alpha_2, ...\alpha_i, ...\alpha_k, \beta_1, \beta_2, ...\beta_i, ...\beta_k, \gamma_1, \gamma_2, ... \\ &\gamma_i, ...\gamma_k, \delta_1, \delta_2, ...\delta_i, ...\delta_k] \end{aligned} \tag{8}$$

It is not difficult to know that $\vec{P}$ has $4k$ components. Similar to previous numerical characterizations we can get 24 different matrix representations of a sequence. So we can get a vector $\overrightarrow{DP}$ with $96k$ components as numerical characterization for a DNA sequence.

4 Applications of ColorSquare

In this section, we present two applications of ColorSquare. The first one is mutation analysis based on ColorSquare. It shows that colorful visualization tool is more effective for human than visualization tool based on shape. The second application is similarity analysis of DNA sequences based on numerical characterizations comparison. The approaches are illustrated on the first exon of β-globin genes of 10 species.

4.1 Mutation Analysis Based on ColorSquare

Human eyes are more sensitive to color than to shape. This feature has been used in advertisement design, fashion design and so on. In this subsection we will use this feature to do mutation analysis. By comparing the visualizations between Randić et al. [29] and ColorSquare, we show that colorful visualization tool is more effective for human to inspect than the visualization tool based on shape.

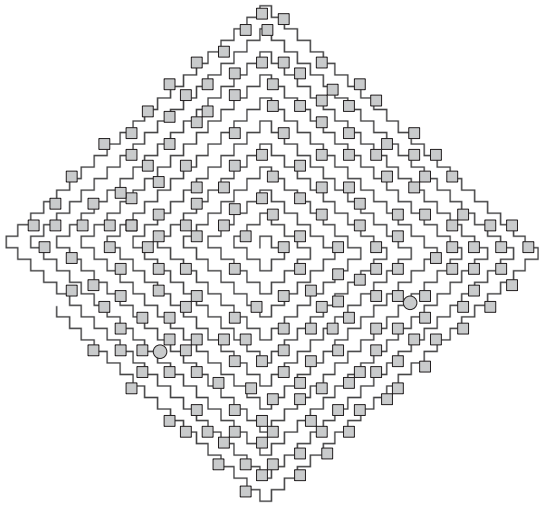

Figure 4: The visualization of a compact model [29] of two sequences with the length about 500 nucleobases. In these two sequences, there are two mutations, marked by circinal spots. However, we are not very easy to find these two mutations.

Fig.4 and Fig.5 are two visualizations of compact model [29] and ColorSquare respectively. Both of them have two mutations, marked by circinal spots in Fig.4 and black spots in Fig.5. Obviously we can find the mutations immediately in ColorSquare whereas we are difficult to find mutations in Fig.4. Because our human eyes are more sensitive to color than to shape.

4.2 Similarity Analysis Based on Numerical Characterizations

Similarity analysis based on numerical characterizations of visualization is an important method in sequence analysis. It was proposed by Randić et al. [27, 43]. It is based on a hypothesis: if the distance between two numerical characterizations is smaller, the

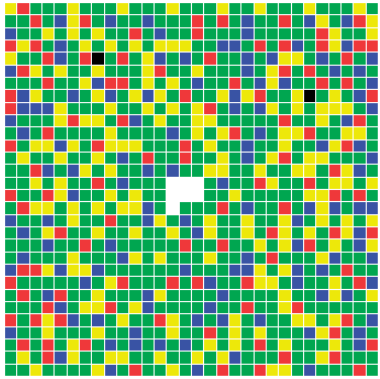

Figure 5: The visualization of ColorSquare of two sequences with the length about 1000 nucleobases. In these two sequences, there are also two mutations, marked by two black spots. Obviously we can easily find these two mutations. This advantage of ColorSquare should be attributed to the fact that human eyes are more sensitive to color than to shape.

corresponding DNA sequences are more similar, the distance between evolutionary closely related species is smaller.

In the previous section, we have proposed two kinds of numerical characterizations. In this section, we will do similarity analysis based on these two kinds of numerical characterizations. It will be illustrated on the first exon of β-globin genes of 10 species.

Suppose that there are two sequences i and j, the numerical characterization vectors are $\vec{D}_i$ and $\vec{D}_j$. Similarity between two sequences can obtained by calculating the Euclidean distance between two vectors.

$$d_{ij} = \|\vec{D}_i - \vec{D}_j\| \tag{9}$$

As shown in Table 1 and Table 2, the similarity results of the DNA sequences of the first exon of β-globin genes of 10 species based on two kinds of numerical characterizations are presented. The two methods come with the consistent results: The values of Goat–Bovine and Human–Gorilla are the smallest, so they are the most similar; the value of Bovine–Mouse is the largest, it means that they are the most different in these 10 species.

Finally we compare our results with other related works. As shown in Table 3, we list the similarities between human and several species in current publications. It shows that Human–Gorilla is the most similar. This is consistent with our results, and further illustrates the effectiveness of our approaches.

Table 1: The similarity result of the DNA sequences of the first exon of β-globin genes of 10 species by the numerical characterizations of 24-component vector based on matrix eigenvalue.

species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine
Human	0	4.777	4.631	5.417	2.764	3.184	3.192	2.622	1.568	4.867
Goat		0	7.223	6.360	5.930	7.778	3.656	5.712	5.538	0.923
Opossum			0	6.332	5.356	5.310	6.337	3.098	5.674	7.197
Gallus				0	7.590	6.407	7.242	5.775	5.257	7.045
Lemur					0	4.299	2.689	2.743	3.234	5.736
Mouse						0	6.013	4.119	2.936	7.916
Rabbit							0	3.826	3.798	3.413
Rat								0	3.376	5.719
Gorilla									0	5.769
Bovine										0

The values of Goat – Bovine and Human – Gorilla are the smallest, so they are the most similar. The value of Bovine – Mouse is the largest. It means that they are the most different in these 10 species.

Table 2: The similarity result of the DNA sequences of the first exon of β-globin genes of 10 species by the numerical characterizations of 96k-component vector based on parity check code.

species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine
Human	0	95.08	66.93	63.25	60.00	90.33	47.33	48.99	24.50	96.33
Goat		0	97.57	94.66	95.08	106.21	90.11	94.23	98.39	21.91
Opossum			0	66.33	70.99	92.95	70.99	68.70	71.27	98.79
Gallus				0	72.66	98.79	73.21	73.76	66.63	95.08
Lemur					0	91.65	66.33	64.50	63.56	96.75
Mouse						0	92.09	86.72	89.67	110.27
Rabbit							0	65.73	52.15	92.74
Rat								0	54.772	96.33
Gorilla									0	100.00
Bovine										0

The similarity result is consistent with Table 1.

Table 3: The similarity between human and other species.

Methods	Gorilla	Gallus	Opossum	Bovine	Goat	Lemur	Mouse	Rabbit	Rat
Our work (Eigenvalue)	1.568	5.417	4.631	4.867	4.777	2.764	3.184	3.192	2.622
Our work (Check Code)	24.50	63.25	66.93	96.33	95.08	60.00	90.33	47.33	48.99
Xie and Mo (2011) [13]	0.042	1.148	0.647	0.074	0.079	0.525	1.49	0.376	1.100
Zhang (2009) [28]	0.263	1.156	1.186	0.361	0.477	0.500	0.444	0.535	0.527
Yao et al. (2008) [44]	0.005	0.029	0.030	0.014	0.016	0.013	0.017	0.011	0.012
Liao et al. (2006) [45]	0.026	0.106	0.096	0.049	0.052	0.064	0.031	0.051	0.049
Liu et al. (2006) [46]	0.008	0.242	0.282	0.075	0.108	0.176	0.076	0.102	0.097

5 Conclusion

We proposed a new visualization of DNA sequences — ColorSquare. ColorSquare is no degeneracy, no loss of information, highly compact, colorful, and square. Because ColorSquare is square, it can be converted into a matrix, from which numerical characterizations can easily be extracted. Mutation analysis and similarity analysis based on ColorSquare are also presented, and demonstrate the usability. Therefore, it can be a convenient tool for researchers in sequence analysis.

Acknowledgment. The work was supported by National Natural Science Foundation of China (61003038, 61033003) and Natural Scientific Research Innovation Foundation in HIT(HIT.NSFIR.2010128).

References

[1] A. Nandy, M. Harle, S. Basak, Mathematical descriptors of DNA sequences: Development and applications, *ARKIVOC* **9** (2006) 211–238.

[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) 1318–1327.

[3] E. Hamori, Novel DNA sequence representations, *Nature* **314** (6012) (1985) 585–586.

[4] M. A. Gates, Simpler DNA sequence representations, *Nature* **316** (6025) (1985) 219–219.

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

[6] P. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, Computer applications in the biosciences: CABIOS 11 (5) (1995) 503.

[7] Z. Zhang, L. Liu, J. Li, Z. Zhang, Spectral representation of protein sequences, *J. Comput. Theor. Nanosci.* **8** (2011) 1335–1339.

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

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

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

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

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

[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.*, in press.

[20] J. Yu, J. 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.

[21] 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.

[22] 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.

[23] B. Liao, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1666–1670.

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

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

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

[27] 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.

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

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

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

[31] 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.

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

[33] Z. Zhang, X. Zeng, T. Song, Z. Chen, X. Wang, Y. Ye, WormStep: An improved compact graphical representation of DNA sequences based on worm curve, *J. Comput. Theor. Nanosci.*, in press.

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

[35] B. Liao, Y. Zhang, K. Ding, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *J. Mol. Struct. (Theochem)* **717** (2005) 199–203.

[36] 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.

[37] Y. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol.* **242** (2006) 382–388.

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

[39] M. Randić, Another look at the chaos-game representation of DNA, *Chem. Phys. Lett.* **456** (2008) 84–88.

[40] G. Huang, B. Liao, Y. Li, Y. Yu, Similarity studies of DNA sequences based on a new 2D graphical representation, *Biophys. Chem.* **143** (2009) 55–59.

[41] 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.

[42] P. He, Y. Zhang, Y. Yao, Y. Tang, X. Nan, The graphical representation of protein sequences based on the physicochemical properties and its applications, *J. Comput. Chem.* **31** (2010) 2136–2142.

[43] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **371** (2003) 202–207.

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

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

[46] X. 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.