MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 679-690

ISSN 0340 - 6253

A Novel Method for Visualizing and Analyzing DNA Sequences

Ronghui Wu[1], Renfa Li[1*], Benyou Liao[1],Guangxue Yue[2]

[1]School of computer and communication, Hunan University, Changsha Hunan, 410082, China

[2]College of Mathematics and Information Engineering, Jiaxing University, Jiaxing Zhejiang, 314001, China

(Received February 1, 2010)

Abstract

One important task in the study of genome sequences is to determine densities of specific nucleotides and to understand the implications for exons or coding regions. Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists. In this manuscript, we introduce a novel method for visualizing and analyzing DNA sequences, the applications on mutation analysis and similarity analysis are presented in detail based on DC-curve (Delta coding curve).

1 Introduction

With the rapid development of Human Biology information technology, the DNA database is growing rapidly, but with the alphabet representation of DNA sequences, it is difficult to obtain information or observe meaningful features from DNA sequences directly. So many bio-scientists are dedicated to the study of DNA sequence. Several researches have outlined different graphical representations of DNA sequences [1-21]. Graphical representation of DNA sequences provides a simple way of viewing, sorting, and comparing various gene structures, also helping in recognizing major differences among similar sequences. It is possible to derive numerical characterization for sequences, providing more information for analyzing the

* Corresponding author Fax: +86 731 88821715
E-mail address: jt_lrf@163.com (R. Li)

sequences [3, 4, 8].

However, most graphical approach are accompanied by two drawbacks or one of this : (1) some loss of visual information associated with crossing and overlapping of the resulting curve by itself; and (2) an arbitrary decision with respect to the choice of the directions for the four base. In order to avoid the limitations related to crossing and overlapping, Liao [6-12,14-16,20,21] and Randic [3,4] present their 2D or 3D graphical representations. However, their approaches are associated with the computations of D/D, L/L and leading eigenvalue, which need a great deal of running time and memory space. The more dimensional methods, 4D [12,13], 6D [10] are difficult to visualize, which are no longer graphical representations of DNA sequences to a certain extent.

Here, DC-curve (Delta coding curve) without degeneracy and loss of information has good visualization to represent long sequences. It is very simple and can reflect the length of DNA sequence. The applications of DC-curve on mutation analysis and similarity analysis are presented in detail.

2 DC-curve, the new 2D representation of DNA sequences

2.1 Construction of DC-curve (Delta coding curve)

In DNA sequences, the four bases A(adenine), C(cytosine), G(guanine) and T(thymine) can be divided into two classes based on their chemical structure, i.e. Purine R={A,G} and pyrimidine Y={C,T}. We present a 2D graphical representation of DNA sequences consisting of two characteristic curves according to the bases classification.

We constructed a DC-R curve and a DC-Y curve on the Cartesian coordinate system. For a DNA sequence $G = g_1 g_2 \cdots g_i \cdots g_n$, where n is the length of this sequence and g_i is the *i*-th base of this sequence, we assign the *x* and *y* axis as follows:

$$\text{DC-R curve:} \quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = A \\ y_{i-1}-1 & \text{if } g_i = G \\ y_{i-1} & \text{if } g_i = T \text{ or } C \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-1}$$

$$\text{DC-Y curve:} \quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = T \\ y_{i-1}-1 & \text{if } g_i = C \\ y_{i-1} & \text{if } g_i = A \text{ or } G \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-2}$$

From (2-1) and (2-2), we can see that the R curve is based on the increase of A and the decrease of G, while the Y curve is based on the increase of T and the decrease of C. Then merge the two curves into DC curve,

$$\text{DC (AG) curve:} \quad \begin{cases} y = y(\text{R curve}) + y(\text{Y curve}) \\ x = i \end{cases} \tag{2-3}$$

For example, the corresponding curves of the sequence AO = "ATGGTGCACCTGAC" is shown in Fig. 1.

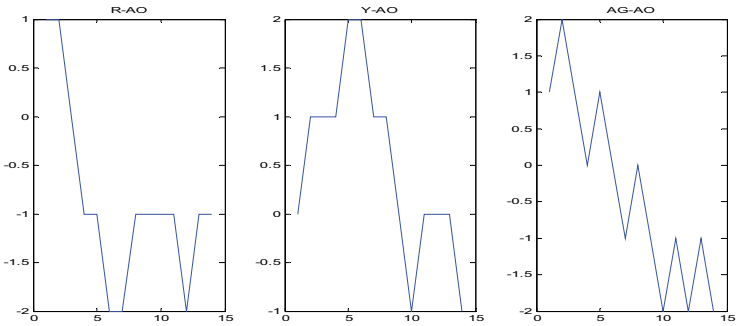

Figure 1. The DCR-curve, DCY-curve, DC-curve of sequence AO = "ATGGTGCACCTGAC".

Also, the four bases can be divided into two groups as amino(A, C)/keto(G, T) and week-H bond(A,T)/strong-H band(G, C). Similar with DC(AG) curve, we can present the following models:

$$\text{AT-W curve:} \quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = A \\ y_{i-1}-1 & \text{if } g_i = T \\ y_{i-1} & \text{if } g_i = G \text{ or } C \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-4}$$

$$\text{AT-H curve:}\quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = G \\ y_{i-1}-1 & \text{if } g_i = C \\ y_{i-1} & \text{if } g_i = A \text{ or } T \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-5}$$

$$\text{AC-A curve:}\quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = A \\ y_{i-1}-1 & \text{if } g_i = C \\ y_{i-1} & \text{if } g_i = T \text{ or } G \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-6}$$

$$\text{AC-K curve:}\quad \begin{cases} y = \begin{cases} y_{i-1}+1 & \text{if } g_i = G \\ y_{i-1}-1 & \text{if } g_i = T \\ y_{i-1} & \text{if } g_i = A \text{ or } C \end{cases} \\ x = i \qquad i = 1 \cdots\cdots n \end{cases} \tag{2-7}$$

$$\text{DC (AT) curve:}\quad \begin{cases} y = y(\text{AT-R curve}) + y(\text{AT-Y curve}) \\ x = i \end{cases} \tag{2-8}$$

$$\text{DC (AC) curve:}\quad \begin{cases} y = y(\text{AC-R curve}) + y(\text{AC-Y curve}) \\ x = i \end{cases} \tag{2-9}$$

From formula on DC (AT) curve and DC (AC) curve, we get the following formula:

$$\begin{cases} y_i - y_{i-1} = 1 & \text{while } g_i = A \text{ or } T \\ y_i - y_{i-1} = -1 & \text{while } g_i = C \text{ or } G \end{cases} \quad i = 1,2,\cdots,n \tag{2-10}$$

So DC (AT) curve and DC (AC) curve are identical. Therefore, we only consider the AT, AG curves in this paper.

For example, the figure 2 shows that the AT, AC curves are identical.

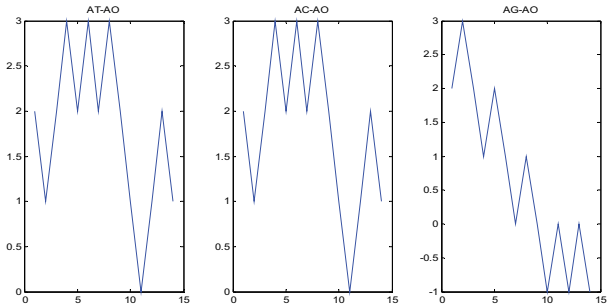

Figure 2. The DC-AT curve, DC-AO curve, DC-AG curve of sequence AO = "ATGGTGCACCTGAC".

2.2 Comparison among the coding sequences of the first exon of beta globin genes of 11 species

In Fig.3 we illustrate the two (AT, AG) curves of the coding sequences of the first exon of beta globin genes of 11 species in Table 1.

Table 1 The coding sequences of the first exon of β-globin gene of eleven different species.

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG
	CCCTGTGGGGCAAGGT
Goat	GAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCT
Opossum	GGGGCAAGGTGAAAGT
	GGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Gallus	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTA
	CCATCTGGTCTAAGGT
Lemur	GCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCG
Mouse	GCCTCTGGGGCAAGGT
	CAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Rabbit	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTC
	TCTGTGGGGCAAGGT
Rat	GGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTT
Gorilla	GCCTGTGGGCAAAGG
	TGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTG
	CCCTGTGGGGCAAGGT
Chimpanzee	GAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGG
	CCTGTGGGGAAAGGT
	GAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG
	CCCTGTGGGGCAAGGT
	GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTT
	GGGGCAAGGTGAAA
	GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG
	CCCTGTGGGGCAAGGT
	GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTG
	GTATCAAGG

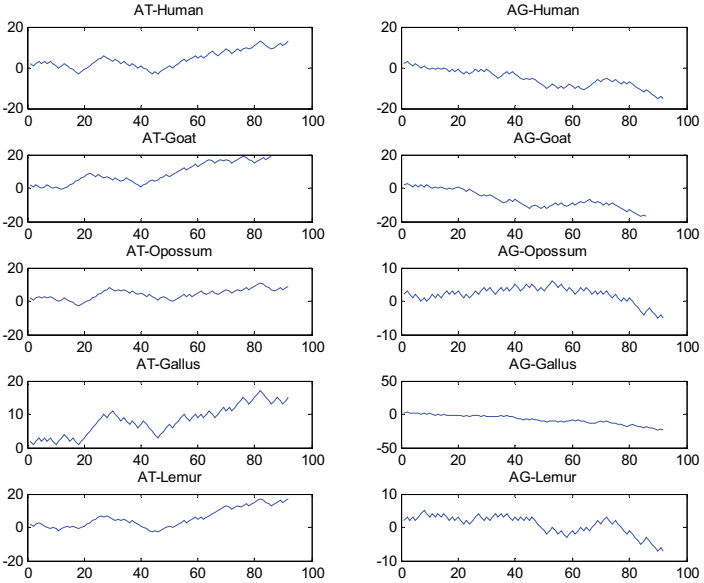

Figure 3(A)

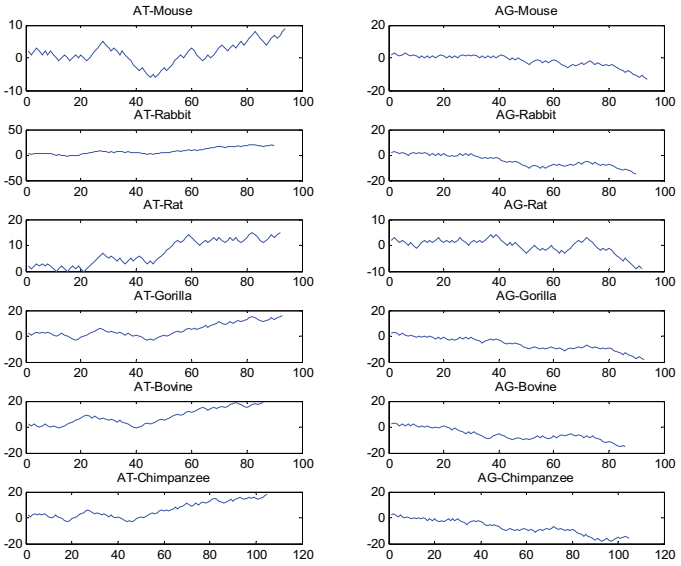

Figure 3(b)

Figure 3. The DC-AT curve, DC-AG curve of the 11 species based on Table 1.

In Figure 3, we can obtain that human and gallus, human and mouse, human and rat are different obviously, while goat-bovine, human-gorilla, human-chimpanzee are most similar, which is consistent with the history of human evolution.

2.3 Advanced properties of DC-Curve

We prove some advanced properties of DC-Curve in this subsection, using mathematical methods and experimental results as follows:

Property1. There is no circuit and degeneracy in DC-Curve.

Proof. Suppose that there is one circuits in DC-Curve at least, so there must exist two points which overlapping themselves. That means, if $i \neq j$, but $(x_i,y_i)=(x_j,y_j)$, so $x_i=x_j$. According to the Equations in subsection 2.1, $x_i=i$ and $x_j=j$. Hence $i=j$. This contradicts $i \neq j$. Therefore, there is no circuit and degeneracy in DC-Curve.

Property2. The correspondence between DNA sequences and DC-Curves(DC-R,DC-Y) is one to one and no loss of information.

First, for a given DC-R curve and DC-Y curve there is a unique DNA sequence correspondingly. From a DC-R curve, suppose it is DC(AG)-R curve, if $y_i-y_{i-1}=1$,then the base g_i=A, if $y_i-y_{i-1}=-1$, then the base g_i=G; in DC(AG)-Y curve, if $y_i-y_{i-1}=1$,we can get the base g_i=T, if $y_i-y_{i-1}=-1$,we can get the base g_i=C. So we can get the sequence uniquely by the two curves.

Second, for a given DNA sequence there is a unique DC-Curve correspondingly, this can be constructed by the formulas in subsection 2.1.

3 Applications

3.1 Mutation analysis

There are four basic types of changes in DNA. They are substitution of a nucleotide for another nucleotide, deletion of nucleotides, insertion of nucleotides, and inversion of nucleotides. We shall consider the properties of mutations based on DC-Curve.

We assume that the mutation appear on the *i*-th base. Let $(x_i,y_i),(x_i',y_i')$be the coordinates of the primal base and mutational base, respectively.

$D(i,j)=\sqrt{(x_i-x_i')^2+(y_i-y_i')^2}$ is called the direction of the mutation. In Table 2, we list the properties of mutations. Obviously, from the 6 different curves we can get the mutations.

Table 2 properties of mutations

	R-curve	**Y-curve**	
DC(AG)-curve	$\sqrt{2}$	$\sqrt{2}$	**A↔T G↔T A↔C G↔C**
	$\sqrt{5}$	**0**	**A↔G**
	0	$\sqrt{5}$	**T↔C**
	W-curve	**H-curve**	
DC(AT)-curve	$\sqrt{2}$	$\sqrt{2}$	**A↔G T↔G A↔C T↔C**
	$\sqrt{5}$	**0**	**A↔T**
	0	$\sqrt{5}$	**G↔C**
	A-curve	**K-curve**	
DC(AC)-curve	$\sqrt{2}$	$\sqrt{2}$	**A↔T C↔T A↔G C↔G**
	$\sqrt{5}$	**0**	**A↔C**
	0	$\sqrt{5}$	**T↔G**

4 Similarities/dissimilarities among the complete coding sequences of the beta globin gene of different species

The goal of the studies are not only finding visual characteristic curves for the sequences, but also finding more information that is easy for sequences analysis. In order to numerically characterize a DNA sequence given by the former methods, one can associate with many matrices according with the curves, and then consider invariants that are sensitive to the form of the curves. In our paper, we use the CM matrix.

For every curve, the coordinates of the geometrical center of the points, denoted

by x^0 and y^0, may be calculated as follows:

$$x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{3-2}$$

The elements of the covariance matrix CM are defined as:

$$\begin{cases} CM_{xx} = \frac{1}{N}\sum_{1}^{N}(x_i - x^0)(x_i - x^0) \\ CM_{xy} = \frac{1}{N}\sum_{1}^{N}(x_i - x^0)(y_i - x^0) = CM_{yx} \\ CM_{yy} = \frac{1}{N}\sum_{1}^{N}(y_i - y^0)(y_i - y^0) \end{cases} \tag{3-3}$$

We get the two normalized eigenvalues of CM matrix for DC-AT curve and DC-AG curve and construct a four-component vector, which is used to compute the similarity/dissimilarity matrix in table 3.

Table 3 The Similarity/Dissimilarity Matrix for the Coding Sequences of Table 1 Based on the Euclidean Distances between the four-Component Vectors of the Normalized Eigenvalues of the CM Matrices.(the scaled value based the maximum value)

species	human	goat	opossum	gallus	lemur	mouse	rabbit	rat	gorilla	bovine	chimpanzee
human	0.0000	0.2622	0.2221	0.5363	0.4674	0.0915	0.4227	0.4532	0.0610	0.3861	0.0714
goat		0.0000	0.1232	0.7803	0.2839	0.3537	0.2330	0.2114	0.3232	0.1498	0.2463
opossum			0.0000	0.6571	0.4028	0.3135	0.3518	0.2471	0.2831	0.2686	0.2000
gallus				0.0000	1.0000	0.4773	0.9589	0.8581	0.5445	0.9224	0.5581
lemur					0.0000	0.5228	0.0728	0.2005	0.4629	0.1342	0.4660
mouse						0.0000	0.4817	0.5447	0.0834	0.4452	0.1538
rabbit							0.0000	0.1495	0.4508	0.1414	0.4250
rat								0.0000	0.5142	0.1708	0.4311
gorilla									0.0000	0.3853	0.0998
bovine										0.0000	0.3885
chimpanzee											0.0000

Observing Table 3, we can find that Human-Gorilla, Human-Chimpanzee have the smallest entry, so they are the most similar species pairs, which is consistent with the history of human evolution. Rabbit-Lemur and Gorilla-Mouse have smaller entries, so they are more similar species pairs, which show they have close evolutionary relationship. The largest entry appears at Gallus-Lemur, and the larger entries appear at Gallus-Rabbit, Gallus-Bovine. We can obtain that Gallus is most different with others.

Table4 Similarity/dissimilarity comparison of the first exon of beat-globin genes between human and other species

species	goat	opossum	gallus	lemur	mouse	rabbit	rat	gorilla	bovine	chimpanzee
A	0.2622	0.2221	0.5363	0.4674	0.0915	0.4227	0.4532	0.0610	0.3861	0.0714
B	0.061	0.148	0.109	0.087	0.083	0.042	0.043	0.021	0.084	0.017
C	0.311	6.322	7.170	1.188	0.735	1.372	1.966	0.339	2.489	0.863
D	0.0436	0.0799	0.0883	0.0533	0.0259	0.0427	0.0407	0.0220	0.0408	0.0206
E	0.4341	0.3805	0.4479	0.3688	0.3089	0.2968	0.4256	0.3070	0.4172	0.3101
F	24.571	23.2195	105.827	10.4924		14.778	19.1187	5.8509	25.2522	7.5617
G	0.0162	0.0601	0.0133	0.0443	0.0111	0.0081	0.0078	0.0012	0.0580	0.0094
H	0.0732	0.1464	0.0895	0.1018	0.0833	0.0725	0.0502	0.0329	0.0493	0.0249
I	0.2061	0.1978	0.2198	0.2637	0.1676	0.1292	0.1758	0.0545	0.1590	0.0467
J	0.3225	0.1666	0.3612	0.2943	0.1677	0.1602	0.2558	0.0835	0.2417	0.0741

A:this work Table 3;B:from Table 3[5];C:from Table 4[9];D:from Table 4[12];E:from Table 3[14];F:from Table 5[18];G:from Table 6[13];H:from Table 2[19];I:from Table 3[19];J:from Table 4[19].

In addition, for comparison, we list the recently published results of the examination of the degree of similarity between human and other several species in Table 4. We can see that there is an overall agreement among similarities obtained by different approaches despite some variations among them. All of them, Human-Gorilla and Human-Chimpanzee have the smallest entry.

5 Conclusion

We present a new method for visualizing and analyzing DNA sequences. The advantage of our approach is that it allows visual inspection of data no matter whether sequences are long. It also can be used to do mutation analysis and similarity analysis of DNA sequences. The examinations show that our method is useful and rationality.

Acknowledgment

This work is supported in part by the National Nature Science Foundation of China (Grant 60973082), the National Nature Science Foundation of Hunan province (Grant 07JJ5080), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195) and the National Nature Science Foundation of Zhejiang province (Grant Y1090264).

References

[1] 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.

[2] A. Nandy, Two–dimensional graphical representation of DNA sequences and intron–exon discrimination in intron–rich sequences, *Comput. Appl. Biosci.* **12** (1996) 55–62.

[3] M. Randić, Condensed representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **40** (2000) 50–56.

[4] 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.

[5] 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.

[6] B. Liao, A 2D graphical representation of DNA sequence, *Chem. Phys. Lett.* **401** (2005) 196–199.

[7] B. Liao, T. Wang, New 2D graphical representation of DNA sequences, *J. Comput. Chem.* **25** (2004) 1364–1368.

[8] B. Liao, T. Wang, 3–D Graphical representation of DNA sequences and their numerical characterization, *J. Mol. Struct. (Theochem)* **681** (2004) 209–212.

[9] Z. B. Liu, B. Liao, W. Zhu, G. H. Huang, A 2–D graphical representation of DNA sequence based on dual nucleotides and its application, *Int. J. Quant. Chem.* **109** (2009) 948–958.

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

[11] 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.

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

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

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

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

[16] Z. Cao, B. Liao, R. F. Li, A group of 3D graphical representation of DNA

sequences based on dual nucleotides, *Int. J. Quant. Chem.* **108** (2008) 1485–1490.

[17] Y. Yao, X. Nan, T. Wang, A new 2D graphical representation – Classification curve and the analysis of similarity/dissimilarity of DNA sequences, *J. Mol. Struct. (Theochem)* **764** (2006) 101–108.

[18] C. Li, X. Q. Yu, N. Helal, Similarity analysis of DNA sequences based on codon usage, *Chem. Phys. Lett.* **459** (2008) 172–174.

[19] W. Chen, Y. S. Zhang, Three distances for rapid similarity analysis of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 781–788.

[20] Y. Li, G. Huang, B. Liao, Z. Liu, H–L curve: A novel 2–D graphical representation of protein sequences, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 519–532.

[21] Z. Liu , B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 541–552.