**MATCH**
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **56** (2006) 209-216

ISSN 0340 - 6253

# 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree

Bo Liao *, Wen Zhu, Yang Liu
*School of Computer and Communication, Hunan University*
*Changsha Hunan 410082,China*

(Received January 19, 2006)

**Abstract.** A 3D graphical representation of DNA sequence using a system of three cartesian coordinates has been derived for mathematical denotation of DNA sequence. The three-dimensional graphical representation also avoids loss of information accompanying alternative 2D and 3D representation in which the curve standing for DNA sequence overlaps and intersects itself, and resolves sequences' degeneracy. The examination of similarities/dissimilarities among the DNA sequences belonging to eleven species illustrates the utility of our approach. The elements of the similarity matrix are used to construct phylogenic tree.

## 1 Introduction

Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures [1-16]. A.Nandy [12] present a graphic representation by assigning A (adenine),G (guanine),T (thymine), and C (cytosine) to the four direction,(-x),(+x),(-y),(+y), respectively. Such a representation of DNA is accompanied by (1) some loos of visual information associated with crossing and overlapping of the resulting curve by itself; (2)an arbitrary decision with respect to the choice of the direction for the four bases. In recent years several authors outlined different graphical representation of DNA sequences based on 2-D, 3-D or 4D [1-8,13-15]. But both 2-D and 3-D graphical representation are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. In mathematical terms, the overlapping and crossing of the curve forms repetitive closed loops or circuits in the DNA graph. M.Randic's 3D [7] or novel 2D [8] graphical representation and B.Liao's 3-D [2,4] or 2D [3] graphic representation avoid the limitations associated with crossing and overlapping, but the representations are not unique.

Here, we present a new three-dimensional graphical representation of DNA sequences, which has no circuit or degeneracy, so that the correspondence between DNA sequences and

*Corresponding author E-mail: dragonbw@163.com

DNA graphs is one to one. The presented graphical representation is used to construct phylogenic tree.

## 2 3-D graphical representation of DNA sequences and its properties

We construct a map between the bases of DNA sequences and plots in 3D space, then we will obtain a 3D representation of the corresponding DNA sequences. In 3D space points, vectors and directions have three components, and the unit vectors representing four nucleotides A,G,C, and T are as follows:
$(m, m, m) \longrightarrow A, (\sqrt{n}, \sqrt{n}, m) \longrightarrow G, (\sqrt{n}, m, \sqrt{n}) \longrightarrow C, (m, \sqrt{n}, \sqrt{n}) \longrightarrow T$

where m is a real number and $m \neq \sqrt{n}$, n is a positive real number but not a perfect square number. So that we will reduce a DNA sequence into a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates $x_i, y_i, z_i (i = 0, 1, 2, \ldots, N$, where $N$ is the length of the DNA sequence being studied)satisfy

$$\begin{cases} x_i = a_i m + g_i\sqrt{n} + c_i\sqrt{n} + t_i m \\ y_i = a_i m + g_i\sqrt{n} + c_i m + t_i\sqrt{n} \\ z_i = a_i m + g_i m + c_i\sqrt{n} + t_i\sqrt{n} \end{cases}$$

where $a_i, c_i, g_i$ and $t_i$ are the cumulative occurrence numbers of A, C, G and T, respectively, in the subsequence from the 1st base to the i-th base in the sequence. We define $a_0 = c_0 = g_0 = t_0 = 0$.

**Property 1** For a given DNA sequence there is a unique 3D representation corresponding to it.

***Proof:***Let $(x_i, y_i, z_i)$ be the coordinates of the i-th base of DNA sequence, then we have

$$a_i(m, m, m) + g_i(\sqrt{n}, \sqrt{n}, m) + c_i(\sqrt{n}, m, \sqrt{n}) + t_i(m, \sqrt{n}, \sqrt{n}) = (x_i, y_i, z_i)$$

i.e.

$$\begin{cases} a_i m + g_i\sqrt{n} + c_i\sqrt{n} + t_i m = x_i \\ a_i m + g_i\sqrt{n} + c_i m + t_i\sqrt{n} = y_i \\ a_i m + g_i m + c_i\sqrt{n} + t_i\sqrt{n} = z_i \end{cases} \tag{1}$$

Obviously, $x_i, y_i$ and $z_i$ are irrational numbers of form $sm + k\sqrt{n}$, where $s$ and $k$ are integers. We suppose

$$x_i = s_x m + k_x\sqrt{n}$$

$$y_i = s_y m + k_y\sqrt{n}$$

$$y_i = s_z m + k_z\sqrt{n}$$

then we have

$$\begin{cases} a_i + t_i = s_x \\ g_i + c_i = k_x \\ a_i + c_i = s_y \\ g_i + t_i = k_y \\ a_i + g_i = s_z \\ c_i + t_i = k_z \end{cases} \tag{2}$$

So, for given x-projection, y-projection and z-projection of any point $P = (x, y, z)$ on the sequence, after uniquely determining $s_x, k_x, s_y, k_y, s_z, k_z$ from $x, y$ and $z$, the number $a_p, g_p, c_p$ and $t_p$ of A,G,C and T from the beginning of the sequence to the point P can be found by solving linear system (2). By successive x-projection, y-projection and z-projection of points on the sequence, we can recover the original DNA sequence uniquely from the DNA graph.

The vector pointing to the point $P_i$ from the origin O is denoted by $r_i$. The component of $r_i$,i.e. $x_i, y_i$ and $z_i$ are calculated by Eq.(1). Let $\Delta r_i = r_i - r_{i-1}$, then we have Property 2.

**Property 2** For any $i = 1, 2, \ldots, N$, where $N$ is the length of the studied DNA sequence, the vector $\Delta r_i$ has only four possible direction. Furthermore, the length of $\Delta r_i$,i.e.,$|\Delta r_i|$, is always equal to $\sqrt{m^2 + 2n}$ or $\sqrt{3m^2}$, for any $i = 1, 2, \ldots, N$.

***Proof:***Actually, the components of $\Delta r_i$,i.e., $\Delta x_i, \Delta y_i$ and $\Delta z_i$ can be calculated for each possible residue (A,G,C and T) at the i-th position of the DNA sequence by using Eq.(1). For example, when the i-th residue is A, we find $\Delta x_i = m, \Delta y_i = m$ and $\Delta z_i = m$. This result is independent of the conformation state of the (i-1)-th residue. The two numbers $(m, m, m)$ are called the direction of $\Delta r_i$. The direction number and the length of $\Delta r_i$ for each possible residue type at the i-th position are summarized as follows.

Table 1: Four possible direction

| | $\Delta x_i$ | $\Delta y_i$ | $\Delta z_i$ | $\lvert\Delta r_i\rvert$ |
|---|---|---|---|---|
| A | $m$ | $m$ | $m$ | $\sqrt{3m^2}$ |
| G | $\sqrt{n}$ | $\sqrt{n}$ | $m$ | $\sqrt{m^2+2n}$ |
| C | $\sqrt{n}$ | $m$ | $\sqrt{n}$ | $\sqrt{m^2+2n}$ |
| T | $m$ | $\sqrt{n}$ | $\sqrt{n}$ | $\sqrt{m^2+2n}$ |

**Property 3** There is no circuit or degeneracy in our three-dimensional graphical representation.

***Proof:*** We assume that: (1) the number of nucleotide forming a circuit is $l$; (2) the number of A,G,C and T in a circuit is $a', g', c'$ and $t'$, respectively. So $a' + g' + c' + t' = l$. Because $a'A, g'G, c'C$ and $t'T$ form a circuit, the following equation holds:

$$a_i(m, m, m) + g_i(\sqrt{n}, \sqrt{n}, m) + c_i(\sqrt{n}, m, \sqrt{n}) + t_i(m, \sqrt{n}, \sqrt{n}) = (0, 0, 0)$$

i.e.

$$\begin{cases} a'm + g'\sqrt{n} + c'\sqrt{n} + t'm = 0 \\ a'm + g'\sqrt{n} + c'm + t'\sqrt{n} = 0 \\ a'm + g'm + c'\sqrt{n} + t'\sqrt{n} = 0 \end{cases} \tag{3}$$

Clearly Eq.(3) hold if , and only if $a' = g' = c' = t' = 0$. Therefore, $l = 0$, which means no circuit exists in this graphical representation.

**Property 4** The 3D representation possesses the reflection symmetry.

***Proof:***usually the sequence is expressed in the order from $5'$ to $3'$. Suppose that the 3D representation for DNA sequence is described by $(x_i, y_i, z_i), i = 0, 1, 2, \ldots, N$. Suppose again

that the 3D representation for the reverse sequence, i.e, the same sequence but from $3'$ to $5'$ is described by $(\hat{x_i}, \hat{y_i}, \hat{z_i})$, we find

$$\begin{cases} \hat{x_i} = x_N - x_{N-i} \\ \hat{y_i} = y_N - y_{N-i} \\ \hat{z_i} = z_N - z_{N-i} \end{cases} \tag{4}$$

# 3 Application

## 3.1 Properties of mutations

Since all morphological and physiological characters of organisms are ultimately controlled by the genetic information carried by DNA, any mutational changes in these characters are due to some change in DNA molecules. 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 this 3D graphical representation of DNA sequence. We assume the mutation appear on the i-th base. Let $(x_i, y_i, z_i), (x_i', y_i', z_i')$ be the coordinates of the primal base and mutational base, respectively. $\Delta x_i = x_i' - x_i, \Delta y_i = y_i' - y_i, \Delta z_i = z_i' - z_i$. The three numbers $(\Delta x_i, \Delta y_i, \Delta z_i)$ are called the direction of the mutation. In table 2, we list the properties of mutations.

## 3.2 Similarity analysis

For any sequence, we have a set of points $(x_i, y_i, z_i), i = 1, 2, 3, \ldots, N$, where $N$ is the length of the sequence. The coordinates of the geometrical center of the points, denoted by $x^0, y^0$ and $z^0$, may be calculated as follows:

$$x^0 = \frac{1}{N}\Sigma_{i=1}^{N} x_i, y^0 = \frac{1}{N}\Sigma_{i=1}^{N} y_i, z^0 = \frac{1}{N}\Sigma_{i=1}^{N} z_i \tag{5}$$

The element of covariance matrix CM of the points are defined:

$$\begin{cases} CM_{xx} = \frac{1}{N}\sum_1^N (x_i - x^0)(x_i - x^0) \\ CM_{xy} = \frac{1}{N}\sum_1^N (x_i - x^0)(y_i - y^0) = CM_{yx} \\ CM_{xz} = \frac{1}{N}\sum_1^N (x_i - x^0)(z_i - z^0) = CM_{zx} \\ CM_{yy} = \frac{1}{N}\sum_1^N (y_i - y^0)(y_i - y^0) \\ CM_{yz} = \frac{1}{N}\sum_1^N (y_i - y^0)(z_i - z^0) = CM_{zy} \\ CM_{zz} = \frac{1}{N}\sum_1^N (z_i - z^0)(z_i - z^0) \end{cases} \tag{6}$$

The above nine numbers give a quantitative description of a set of point $(x_i, y_i, z_i), i = 1, 2, \ldots, N$, scattering in a three-dimensional space. The eigenvalues of CM are applied to make analysis of similarity.

As an example, we assume $m = \frac{1}{2}, n = \frac{3}{4}$, then we compute the similarities among eleven mitochondrial sequences belonging to different species and present their phylogenic tree. In

Table 2: Properties of mutations

| | $\Delta x_i$ | $\Delta y_i$ | $\Delta z_i$ | direction |
|---|---|---|---|---|
| $A \to C$ | $\sqrt{n}-m$ | 0 | $\sqrt{n}-m$ | $(\sqrt{n}-m, 0, \sqrt{n}-m)$ |
| $C \to A$ | $m-\sqrt{n}$ | 0 | $m-\sqrt{n}$ | $(m-\sqrt{n}, 0, m-\sqrt{n})$ |
| $A \to G$ | $\sqrt{n}-m$ | $\sqrt{n}-m$ | 0 | $(\sqrt{n}-m, \sqrt{n}-m, 0)$ |
| $G \to A$ | $m-\sqrt{n}$ | $m-\sqrt{n}$ | 0 | $(m-\sqrt{n}, m-\sqrt{n}, 0)$ |
| $A \to T$ | 0 | $\sqrt{n}-m$ | $\sqrt{n}-m$ | $(0, \sqrt{n}-m, \sqrt{n}-m)$ |
| $T \to A$ | 0 | $m-\sqrt{n}$ | $m-\sqrt{n}$ | $(0, m-\sqrt{n}, m-\sqrt{n})$ |
| $C \to G$ | 0 | $\sqrt{n}-m$ | $m-\sqrt{n}$ | $(0, \sqrt{n}-m, m-\sqrt{n})$ |
| $G \to C$ | 0 | $m-\sqrt{n}$ | $\sqrt{n}-m$ | $(0, m-\sqrt{n}, \sqrt{n}-m)$ |
| $C \to T$ | $m-\sqrt{n}$ | $\sqrt{n}-m$ | 0 | $(m-\sqrt{n}, \sqrt{n}-m, 0)$ |
| $T \to C$ | $\sqrt{n}-m$ | $m-\sqrt{n}$ | 0 | $(\sqrt{n}-m, m-\sqrt{n}, 0)$ |
| $G \to T$ | $m-\sqrt{n}$ | 0 | $\sqrt{n}-m$ | $(m-\sqrt{n}, 0, \sqrt{n}-m)$ |
| $T \to G$ | $\sqrt{n}-m$ | 0 | $m-\sqrt{n}$ | $(\sqrt{n}-m, 0, m-\sqrt{n})$ |
| $A \to \Phi$ | $-m$ | $-m$ | $-m$ | $(-m, -m, -m)$ |
| $\Phi \to A$ | $m$ | $m$ | $m$ | $(m, m, m)$ |
| $C \to \Phi$ | $-\sqrt{n}$ | $-m$ | $-\sqrt{n}$ | $(-\sqrt{n}, -m, -\sqrt{n})$ |
| $\Phi \to C$ | $\sqrt{n}$ | $m$ | $\sqrt{n}$ | $(\sqrt{n}, m, \sqrt{n})$ |
| $G \to \Phi$ | $-\sqrt{n}$ | $-\sqrt{n}$ | $-m$ | $(-\sqrt{n}, -\sqrt{n}, -m)$ |
| $\Phi \to G$ | $\sqrt{n}$ | $\sqrt{n}$ | $m$ | $(\sqrt{n}, \sqrt{n}, m)$ |
| $T \to \Phi$ | $-m$ | $-\sqrt{n}$ | $-\sqrt{n}$ | $(-m, -\sqrt{n}, -\sqrt{n})$ |
| $\Phi \to T$ | $m$ | $\sqrt{n}$ | $\sqrt{n}$ | $(m, \sqrt{n}, \sqrt{n})$ |

$\Omega \to -$ corresponds a deletion, while $- \to \Omega$ corresponds a insertion, $\Omega \in \{A, C, G, T\}$

Table 3: The $(x^0, y^0, z^0)$ and eigenvalues for the mitochondrial sequences belonging to 11 species

| | $(x^0, y^0, z^0)$ | eigenvalues(1.0e+004) |
|---|---|---|
| Chi | (295.9726,285.3599,322.3035) | $0.0001, 0.0003, 8.9809$ |
| Gor | (297.1336,284.6920,320.7827) | $0.0001, 0.0004, 8.9702$ |
| Hyl | (299.4339,285.1499,319.8300) | $0.0001, 0.0003, 8.9084$ |
| L. cat | (287.7326,286.8672,315.9325) | $0.0001, 0.0003, 8.7160$ |
| M. fas | (294.7454,288.2121,321.2512) | $0.0001, 0.0002, 8.8982$ |
| M. fus | (294.1449,285.6169,320.6446) | $0, 0.0002, 8.8555$ |
| M. mul | (294.1523,286.9625,320.5563) | $0, 0.0003, 8.8654$ |
| M. syl | (293.0922,286.3371,320.3946) | $0.0001, 0.0005, 8.7998$ |
| Pon | (299.7722,279.9737,321.3893) | $0, 0.0005, 8.9272$ |
| S. sci | (288.3447,290.4896,313.3262) | $0.0001, 0.0003, 8.7007$ |
| T. syr | (288.3530,288.9300,315.3849) | $0.0001, 0.0006, 8.6510$ |

Chi: Chimpanzee; Gor; Gorilla; Hyl: Hylobates; L.cat: Lemur catta; M.fas: Macaca fascicularis; M.fus: Macaca fuscata; M.mul: Macaca mulatta; M.syl: Macaca sylvanus; Pon: Pongo; S.sci: Saimiri sciureus ; T.syr: Tarsius syrichta.

table 3, we list the $(x^0, y^0, z^0)$ belonging to 11 species. In Table 4, we show the similarity matrix.

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we introduce a distance scale and an angle scale as defined below. Suppose that there are two species $i'$ and $j'$, the parameters are $\lambda_1^{i'}, \lambda_2^{i'}, \lambda_3^{i'}, \lambda_1^{j'}, \lambda_2^{j'}, \lambda_3^{j'}$, respectively, where $\lambda_1^{i'}, \lambda_2^{i'}, \lambda_3^{i'}$ are the three eigenvalues of matrix $CM_{i'}$ corresponding to species $i'$. We will illustrate the use of the 3-D quantitative characterization of DNA sequence with an examination of similarities/dissimilarties among the eleven species. We construct a three-component vector consisting of the three eigenvalues of matrix CM. The underlying assumption is that if two vectors point to a similar direction in the three-dimensional space, and then the two DNA sequences represented by the three-component vectors are similar. The similarities among such vectors can be computed by calculating the Euclidean distance between the end point of the vectors. The distance $d_{i'j'}$ between the two vectors is:

$$d_{i'j'} = \sqrt{(\lambda_1^{i'} - \lambda_1^{j'})^2 + (\lambda_2^{i'} - \lambda_2^{j'})^2 + (\lambda_3^{i'} - \lambda_3^{j'})^2} \tag{7}$$

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

Table 4: The similarity/dissimilarity matrix(1.0e+004) for the coding sequences based on the Euclidean distances between the end points of the 3-component vectors of the eigenvalues of the CM matrices

| *Species* | Chi | Gor | Hyl | L. cat | M. fas | M. fus | M. mul | M. syl | Pon | S. sci | T. syr |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chi | 0 | 0.0107 | 0.0725 | 0.2649 | 0.0827 | 0.1254 | 0.1155 | 0.1811 | 0.0537 | 0.2802 | 0.3299 |
| Gor | | 0 | 0.0618 | 0.2542 | 0.0720 | 0.1147 | 0.1048 | 0.1704 | 0.0430 | 0.2695 | 0.3192 |
| Hyl | | | 0 | 0.1924 | 0.0102 | 0.0529 | 0.0430 | 0.1086 | 0.0188 | 0.2077 | 0.2574 |
| L. cat | | | | 0 | 0.1822 | 0.1395 | 0.1494 | 0.0838 | 0.2112 | 0.0153 | 0.0650 |
| M. fas | | | | | 0 | 0.0427 | 0.0328 | 0.0984 | 0.0290 | 0.1975 | 0.2472 |
| M. fus | | | | | | 0 | 0.0099 | 0.0557 | 0.0717 | 0.1548 | 0.2045 |
| M. mul | | | | | | | 0 | 0.0656 | 0.0618 | 0.1647 | 0.2144 |
| M. syl | | | | | | | | 0 | 0.1274 | 0.0991 | 0.1488 |
| Pon | | | | | | | | | 0 | 0.2265 | 0.2762 |
| S. sci | | | | | | | | | | 0 | 0.0497 |
| T. syr | | | | | | | | | | | 0 |

Observing Table 4, we find that the more similar species pairs are *Chimpanzee* $\sim$ *Gorilla* and *Macaca fascicularis* $\sim$ *Macaca mulatta*, while Lemur catta and Tarsius syrichta are dissimilar to others. Therefore, the classification of species provided that the numbers of their coding sequences are sufficiently large, can be generally performed in terms of the two matrices as listed in Table 3. In other words, with the continuous increase in the number of coding sequences for various species, it is possible to perform the cluster analysis by the distance matrices.

The elements of the similarity matrix are used to construct phylogenic tree using the maximum parsimony method. In figure 1, we show the phylogenic tree belonging to eleven species.

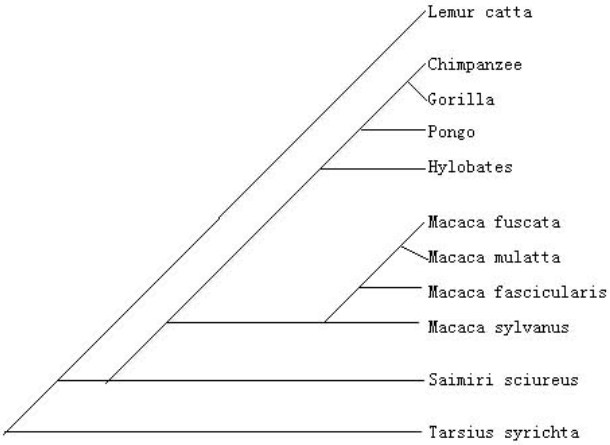


Figure 1: phylogenic tree

## 4 Conclusion

High complexity and degeneracy are major problems in previous DNA sequence representations. Our representation provides a direct plotting method to denote DNA sequences without degeneracy. From the DNA graph, the A,T,G and C usage as well as the original DNA sequence can be recaptured mathematically without loss of textual information. The current three-dimensional graphical representation of DNA sequences provides different approaches for both computational scientists and molecular biologists to analysis DNA sequences efficiently with different parameter $n$ and $m$.

## 5 Acknowledgment

This work is supported in part by the China Postdoctoral Science Foundation and the the National Natural Science Foundation of Hunan University.

## References

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

[2] C. X. Yuan, B. Liao, T. M. Wang, New 3-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.*, 379 (2003), 412-417.

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

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

[5] B. Liao, T. M. Wang, Analysis of similarity of DNA sequences based on 3D graphical representation, *Chem. Phys. Lett.*, 388 (2004), 195-200.

[6] B. Liao, K. Q. Ding, Graphical approach to analyzing DNA sequences, *J. Comput. Chem.* , 14 (2005), 1519-1523.

[7] M. Randic, M. Vracko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequence and their numerical characterization, *J. Chem. Inf. Comput. Sci.*, 40 (2000), 1235-1244.

[8] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numberical characterization, *Chem. Phys. Lett.*, 368 (2003), 1-6.

[9] E. Hamori, J. Ruskin, H curves, a novel method of representaion of nucleotides series especially suited for long DNA sequences. *J. Biol. Chem.* , 258 (1983), 1318-1327.

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

[11] M. A. Gates, Simple DNA sequence representations, *Nature* , 316 (1985), 219.

[12] 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.

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

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

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

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