# Some Notes on 2-D Graphical Representation of DNA Sequence

Yachun Liu,*,† Xiaofeng Guo,‡ Jin Xu,§ Linqiang Pan,§ and Shiying Wang§

Department of Mathematics and Physical Science, Institute of Engineering and Technology, Nanhua University, Hengyang, Hunan 421001, P. R. China, Institute of Mathematics and Physics, Xinjiang University, Wulumuqi, Xinjiang 830046, P. R. China, and Department of Control Science and Engineering, Huazhong University of Science & Technology, Wuhan, Hubei 430074, P. R. China

Received February 28, 2001

Some 2-D and 3-D graphical representations of DNA sequences have been given by Nandy, Leong and Mogenthaler, and Randic et al., which give visual characterizations of DNA sequences. In this paper, we presented a novel graphical representation of DNA sequences by taking four special vectors in 2-D Cartesian coordinate system to represent the four nucleic acid bases in DNA sequences, so that a DNA sequence is denoted on a plane by a directed walk. It is shown that the new graphical representation of DNA sequences has lower or nondegeneracy.

## 1. INTRODUCTION

Effective representation of long DNA sequences has led to several innovative techniques to provide useful ways of viewing, sorting, analyzing, and comparing various sequences. One of the first attempts was that of Hamori and Ruskin[1] by the G- and H-curves, which is plotted in a five- or three-dimensional space, respectively. To provide rapid visual clues to sequence identity and similarities and difference in distribution of bases along the sequence, a reduced two-dimensional graphical representation of DNA sequences was proposed by M. A. Gates,[2] Nandy,[3,4] Leong, and Mogenthaler.[5] Their method is based on choosing the four cardinal directions in (*x*, *y*) coordinate system to represent the four bases in DNA sequences. The method essentially consists of plotting a point corresponding to a base by moving one unit in the positive or negative *x*- or *y*-axes depending on the defined association of a base with a cardinal direction. The cumulative plot of such points produces a graph that corresponds to the sequence. In the Gates axes system, one would move one unit in the positive *x*-direction for a cytosine (C), along the positive *y*-direction for a thymine (T), the negative *x*-direction for a guanine (G), the negative *y*-direction for an adenosine (A), implying a cumulative plot of the count of instantaneous $C-G$ against $T-A$. The Nandy axes system associates G with positive *x*-direction, C with positive *y*-direction, A with negative *x*-direction, and T with negative *y*-direction. In the Leong and Morgenthaler axes system, A is associated with positive *x*-direction, T with positive *y*-direction, C with negative *x*-direction, and G with negative *y*-direction. It was pointed out by Nandy (1995)[6] that there are three possible independent axes system to plot a two-dimensional graph of DNA sequence, and in fact the three systems mentioned above cover the three orthogonal systems (see Figure 1).

The 2-D graphical representation of DNA sequence is called the graph of the DNA sequence corresponding to ACGT-axis system (counting clockwise) of Nandy, while much work has been done using this cartesion coordinate system,[3,7,8] other pumutations have also been considered,[2,5] their studies of gene sequences of the globin family, the myosin heavy chain genes, histones, and others in such a representation has shown that evolutionary changes produce shifts in base distribution that appear to reflect evolutionary distances in the dispersion of the graphical form of the sequences. It is clear that the differences in the base composition and distribution of individual members of a homologous family will induce changes in the plots of the sequences in the graphical representation method outlined. In general, changes in base composition will result in changes in end-points represented by the coordinates ($G-A$, $C-T$), where the letters represent the total number of each base in the sequence. Changes will also arise from the differences in distribution of the bases along the sequence as measured by the instantaneous coordinates ($g-a$, $c-t$) where the lower case letters represent the number of each base up to the nucleotide number under consideration. However, because of high degeneracy of the graphical representation of DNA sequence, the essential attribute property hidden in DNA sequence could not be characterized completely. In another words, different sequences, for example, the sequences AGTC, AGTCA, AGTCAG, etc. may have the same graphical representation. If a kind of graphical representation of DNA sequences has no circuit, then each DNA sequence can be uniquely determined by the graphical representation of it. We use the minimum length of all the DNA sequences, each of which forms a circuit in a graphical representation to measure the relative degree of degeneracy. In Nandy's 2-D graphical representation of DNA sequences, the minimum length of a circuit is equal to 2, because anyone of TC, CT, AG, GA forms a circuit with the minimum length.

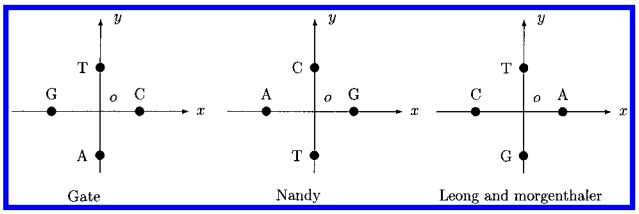


**Figure 1.** The three independent orthogonal axes systems.

* Corresponding author e-mail: liuyachun65@263.net.
† Nanhua University.
‡ Xinjiang University.
§ Huazhong University of Science & Technology.

10.1021/ci010017g CCC: $22.00 
Published on Web 03/26/2002

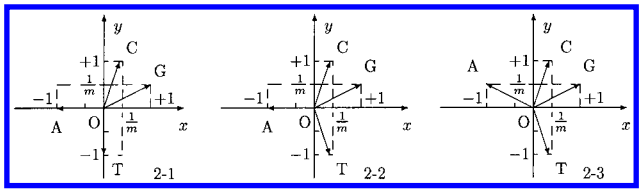


**Figure 2.** More than two of the four unit vectors that represent the corresponding bases be deviated from their original cardinal axes directions in Nandy axes system.

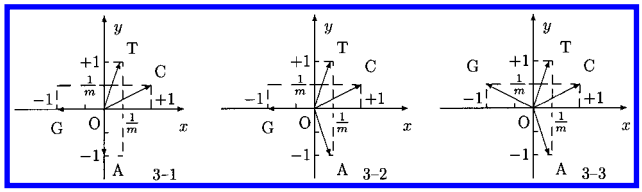


**Figure 3.** A novel axes system for two-dimensional graphical representation corresponding to Gate axes system.

On the other hand, based on the previous graphical representation, there is not an effective mathematical evaluation scheme and a very useful classification method so far. Recently, Xiaofeng Guo etc. put forward a novel 2-D graphical representation of DNA sequences of low degeneracy.[9] In this paper, we evolved Nandy's and Guo's 2-D graphical representation of DNA sequences. It will be demonstrated that the improved two-dimensional representation of DNA sequences is very useful in studies of nucleotide distribution and composition.

## 2. A NOVEL 2-D GRAPHICAL REPRESENTATION OF DNA SEQUENCES OF LOW DEGENERACY

To reduce the degeneracy of Nandy's graphical representation, more than two of the four unit vectors that represent the corresponding bases must be deviated from their original cardinal axes directions. We design four special vectors in Cartesian ($x$, $y$) coordinate system to represent the four nucleic acid bases A, C, G, T, which the initial point of the vectors is the cardinal origin. A DNA sequence of four letters A, C, G, T with length $n$ can be regarded as a successive vector sequence $V_1$, $V_2$, $\cdots$ $V_n$ of length $n$ consisting of the four vector sequence corresponding to A, C, G, T. A vector sequence $V_1$, $V_2$, $\cdots V_n$ is said to be a successive sequence if $V_1$, $V_2$, $\cdots$ $V_n$ are shifted parallel so that, for $2 \le i \le n$, the initial point of $V_i$ is identical with the terminal point of $V_{i-1}$ step by step. The graphical representation of DNA sequence may be regarded as a directed walk in digraph. There are three methods of designing four special vectors in Cartesian ($x$, $y$) coordinate system to represent the four bases A, C, G, T as following figures, where $m$ is s an integer greater than 1.

It should be mentioned here that there are also three possible independent axes systems for the novel graphical representation of DNA sequences, which are respectively corresponding to axes systems of Nandy, Gates, and Leong and Morgenthaler. The axes systems of the graphical representation given above is corresponding to Nandy's axes system. The other two independent othogonal axes systems for Figure 2 are shown in Figures 3 and 4.

It is interesting that a DNA sequence has three different graphical representations in three independent axes systems. Some, one, or all of them would be used in various applications.

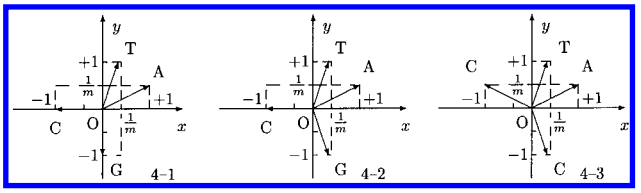


**Figure 4.** A novel axes system for two-dimensional graphical representation corresponding to Leong and Morgenthaler axes system.

## 3. THE DEGENERACY OF THE NOVEL 2-D GRAPHICAL REPRESENTATION OF DNA SEQUENCES

To determine the degeneracy of the novel 2-D graphical representation of DNA sequences, we need to calculate the minimum length of all the DNA sequences each of which forms a circuit in the graphical representation of it. Let $S = V_1V_2\cdots V_n$ be a DNA sequence which forms a circuit in its graphical representation $G(S)$, and $f_A$, $f_C$, $f_G$, $f_T$ be the frequencies of A, C, G, T in S, respectively. By the above assumption and the definition of the graphical representation of $S$, we divided the discussion of degeneracy problem to three cases as following:

**Case 1.** If $G_m(S)$ is the graphical representation of DNA sequence $S$ in Figure 2-1, then we have the following system of equations:

$$\begin{cases} -f_A + \dfrac{f_C}{2m} + f_G = 0 \\ f_c + \dfrac{f_G}{m} - f_T = 0 \end{cases}$$

$$\text{i.e.} \begin{cases} -mf_A + f_C + mf_G = 0 & (1) \\ mf_C + f_G - mf_T = 0 & (2) \end{cases}$$

Adding eqs 1 and 2, we have $(m + 1)(f_C + f_G) = m(f_A + f_T)$. Because $(m, m + 1) = 1$, implying that $f_A + f_T = c(m + 1)$, $f_C + f_G = c{\cdot}m$, where $c$ is a positive integer. So, $|S| = f_A + f_C + f_G + f_T = (2m + 1){\cdot}c \ge 2m + 1$, $min|S| = 2m + 1$ while $c = 1$, and then $f_A + f_T = m + 1$. Adding the equation into (1), and soluting the system of equations, we have that

$$f_G = \frac{m(m - f_T)}{m - 1}$$

Because $m$ is relatively prime with $m - 1$, so $m - f_T$ could be divided by $m - 1$, hence $f_T = 1$, $f_G = m$, $f_C = 0$, $f_A = m$.

**Case 2.** If $G_m(S)$ is the graphical representation of DNA sequence $S$ in Figure 2-2, then we have the following system of equations:

$$\begin{cases} -f_A + \dfrac{f_C}{m} + f_G + f_T = 0 \\ f_C + f_G m - f_T = 0 \end{cases}$$

$$\text{i.e.} \begin{cases} -mf_A + f_C + mf_G + f_T = 0 & (3) \\ mf_C + f_G - mf_T = 0 & (4) \end{cases}$$

Adding eqs 3 and 4, we have

$$\begin{cases} f_C = \dfrac{-mf_A + (m^2+1)f_T}{(m-1)(m+1)} & (5) \\ f_G = \dfrac{m(mf_A - 2f_T)}{(m-1)(m+1)} & (6) \end{cases}$$

Because $(m, m+1) = 1, (m, m-1) = 1$, we could deduce that $mf_A = 2f_T$ from eq 6, and then derive that $f_A = 2$, $f_T = m$, $f_C = m$, $f_G = 0$, $min|S| = 2(m+1)$.

**Case 3.** If $G_m(S)$ is the graphical representation of DNA sequence $S$ in Figure 2-3, then we have the following system of equations:[9]

$$\begin{cases} -f_A + \dfrac{f_C}{m} + f_G + \dfrac{f_T}{m} = 0 \\ \dfrac{f_A}{m} + f_C + \dfrac{f_G}{m} - f_T = 0 \end{cases}$$

$$\text{i.e.} \begin{cases} -mf_A + f_C + mf_G + f_T = 0 & (7) \\ f_A + mf_C + f_G - mf_T = 0 & (8) \end{cases}$$

Adding two eqs 7 and 8, we have $(m+1)(f_C + f_G) = (m-1)(f_A + f_T)$,

$$\text{i.e. } f_C + f_G = \frac{m-1}{m+1}(f_A + f_T)$$

Adding the equation into (7), and solving the system of equations, we derive

$$f_C = \frac{-2mf_A + (m^2+1)f_T}{(m-1)(m+1)} \quad (9)$$

$$f_G = \frac{(m^2+1)f_A - 2mf_T}{(m-1)(m+1)} \quad (10)$$

and then

$$min|S| = f_A + f_C + f_G + f_T = \frac{2m}{m+1}(f_A + f_T)$$

If $m$ is an odd number, $(m-1, m+1) = 2$, implying that $f_A + f_T = ((m+1)/2)x$, $f_C + f_G = ((m-1)/2)x$,$|S| = mx$, where x is a positive integer. In order to calculate the minimum length of all the DNA sequences each of which forms a circuit in the graphical representation of it in Figure 2-3, we only need to determine the minimum $x$. Substituting these equations into (9) and (10), it can be obtained that

$$\begin{cases} f_A = \dfrac{m+1}{2}x - f_T \\ f_C = \dfrac{(m+1)f_T - mx}{m-1} \\ f_G = \dfrac{(m^2+1)x - 2(m+1)f_T}{2(m-1)} \end{cases} \quad (11)$$

From (11), we have the following: $0 \le f_T \le ((m+1)/2)x$, and $f_T = (m-1)f_C + mx/m + 1$. If $x = 1$, then

$$\begin{aligned} f_T &= \frac{(m-1)f_C + m}{m+1} \\ &= \frac{(m+1)f_C - 2f_C + m}{m+1} \\ &= f_C + \frac{m - 2f_C}{m+1} \end{aligned}$$

Because $m$ is an odd number, $f_A$, $f_C$, $f_G$, $f_T$ are integers, so, the above equation is not true, hence $x \neq 1$. If $x = 2$, then

$$0 \le f_T \le m+1$$

and

$$f_T = f_C + \frac{2(m - f_C)}{m+1}$$

Since $m - f_C < m + 1$, $f_T$ has an non-negative integer solution if and only if $2(m - f_C) = 0$ or $(m+1)$. If $2(m - f_C) = 0$, then $f_C = m$, and this is in contradiction with $f_C + f_G = ((m-1)/2)x = m - 1$. So $2(m - f_C) = m + 1$, $f_C = (m-1)/2 = f_G$, and $f_A = f_T = (m+1)/2$, $min|S| = 2m$.

If $m$ is an even number, then $(m-1, m+1) = 1$, implying that $f_A + f_T = x(m+1)$, where $x$ is a positive integer. Substituting this relation into (9) and (10), we have

$$\begin{cases} f_A = x(m+1) - f_T \\ f_C = \dfrac{(m+1)f_T - 2mx}{m-1} \\ f_G = \dfrac{(m^2+1)x - (m+1)f_T}{m-1} \end{cases} \quad (12)$$

From (12), it can be obtained that

$$\begin{cases} 0 \le f_T \le x(m+1) \\ f_T = f_C + \dfrac{2(mx - f_C)}{m+1} \end{cases}$$

If $x = 1$, $f_T$ is an integer unless $2(m - f_C) = 0$ or $(m+1)$, i.e. $f_C = m$ or $(m-1)/2$. Since $m$ is even number, $f_C$ is an integer, so $f_C \neq (m-1)/2$. If $f_C = m$, then $f_T = m$, $f_A = 1$, $f_G = -1$ could be derived, in contradiction with integral $f_G$. Hence $x \neq 1$. If $x = 2$, then $f_T = f_C + 2(2m - f_C)/(m+1)$, and $0 \le f_T \le 2(m+1)$. $f_T$ is a non-negative integer if and only if $2(2m - f_C) = k(m+1)$, where $k = 0, 1, 2, 3$. If $k = 0$, then $f_C = f_T = 2m$, $f_A = 2$, $f_G = -2$, a contradict. If $k = 1$, then $f_c = (3m-1)/2$ is not an integer, a contradiction too. If $k = 2$, it could be derived that $f_A = f_T = m + 1$, $f_C = f_G = m - 1$. If $k = 3$, then $f_C = (m-3)/2$, $f_C$ is not an integer. So, when $m$ is an even number, $min|S| = 4m$.

Overviewing above demonstration, we have the following theorem:

**Theorem 1.** Let $S$ be a DNA sequence whose graphical representation $G(S, m, i)$ in Figure 2—i forms a circuit with the minimum length $\mathcal{L}(S, m, i)$, where $i = 1, 2, 3$. $m$ is a natural number greater 1. Then

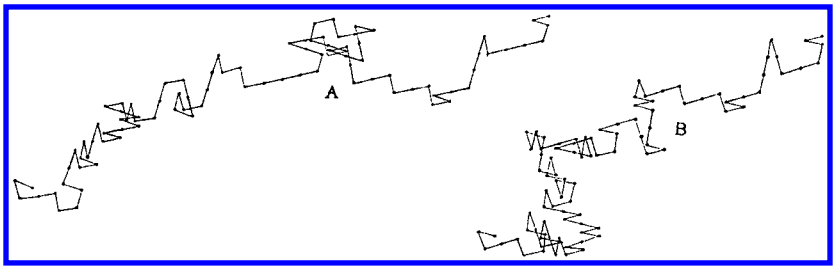


**Figure 5.** The novel graphical representations of human beta-globin exon-1 and the opossum beta-hemoglobin exon-1 based on the ACGT-axes system in Figure 2-3, where $m = 4$.

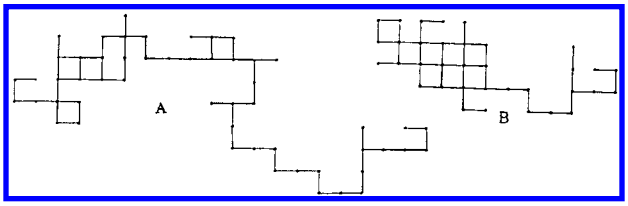


**Figure 6.** Nandy's graphical representation of human beta-globin exon-1 and the opossum beta-hemoglobin exon-1 based on Nandy's ACGT-axes system.

1. $\mathcal{N}(S, m, 1) = 2m + 1$, $f_A = m$, $f_C = 0$, $f_G = m$, $f_T = 1$;
2. $\mathcal{N}(S, m, 2) = 2(m + 1)$, $f_A = 2$, $f_C = m$, $f_G = 0$, $f_T = m$;
3. $\mathcal{N}(S, m, 3) = 2m$, $f_A = f_T = (m + 1)/2$, $f_C = f_G = (m - 1)/2$, if and only if $m$ is odd; $\mathcal{N}(S, m, 3) = 4m$, $f_A = f_T = m + 1$, $f_C = f_G = m - 1$, if and only if $m$ is even.

The above results show that, if $m$ is a greater natural number, then the graphical representation of a DNA sequence has a lower degeneracy. For an individual DNA sequence, we can also measure the degree of degeneracy of a graphical representation of it by the quotient of the number of the bases in the sequence and the number of edges in the graphical representation. For the human beta-globin exon-1 and the opossum beta-hemoglobin exon-1, under the ACGT-axes system in Figure 2-3, the novel graphical representation of the two DNA sequences has no circuit, as shown as Figure 5 where the graphical representations of DNA sequences take $m$ equal to 4, and so the corresponding quotient is equal to 1, implying that the graphical representations can uniquely determine the corresponding DNA primary sequences. In contrast, in Nandy's graphical representation of the same two DNA sequences, there are closed walks and repeated overlapping at numerous points and edges (as shown as Figure 6). The quotients for the two Nandy's graphical representations are respectively equal to 1.64 and 2.190, hence, the degeneracy is very high, and then there is a considerable loss of information in the case of Nandy's graphical representation. Moreover, from Figure 5 one can see an important advantage of the novel graphical representations by comparing the "tail" parts of graphical representations in Figures 5 and 6: One can see that the last 20 bases are identical except for a single insertion of G (which makes a longer segment near the end). This cannot be seen from Figure 6 in which the "tail" parts of human beta globin and opossum look quit different.

Although the degeneracy in Theorem 1 tends to zero when $m$ is prone to infinite, the degeneracy always exists in actual application. Therefore, there is continuing interest in finding a novel 2-D graphical representation of DNA sequences which has no degeneracy.

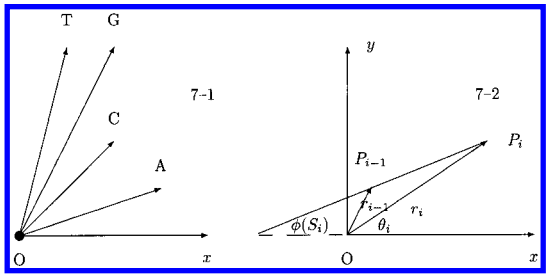


**Figure 7.** Polar coordinate system and the 2-D graphical representation of DNA sequences of no degeneracy.

## 4. A NEW 2-D GRAPHICAL REPRESENTATION OF DNA SEQUENCES OF NO DEGENERACY IN POLAR COORDINATE SYSTEM

Let $O$ be the pole and $\overrightarrow{OX}$ be the polar axes. $r_1, r_2, \cdots r_n, \cdots$ is a given increasing sequence of numbers, and $\phi_1$, $\phi_2$, $\phi_3$, $\phi_4$ are four given different angles, where $0 \le \phi_1, \phi_2, \phi_3, \phi_4 < 2\pi$, making four rays $\overrightarrow{OA}$, $\overrightarrow{OC}$, $\overrightarrow{OG}$, $\overrightarrow{OT}$, where their polar angles are respectively $\phi_1$, $\phi_2$, $\phi_3$, $\phi_4$ which match to nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T), respectively (see Figure 7-1). If $S_1S_2\cdots S_n$ is a DNA sequence of length $n$, where $S_i \in \{A, C, G, T\}$, the corresponding sequence of points $P_0P_1P_2\cdots P_n$ will be constructed such that the vector $\overrightarrow{P_{i-1}P_i}$ corresponds to $S_i$, where $P_0$ is the pole; the distance $|OP_i|$ is equal to $r_i$, the angle that $\overrightarrow{OX}$ rotates to $\overrightarrow{OP_i}$ counterclockwise is $\theta_i$, $i = 1, 2, \cdots n$. If $S_1 = A$, then $\theta_1 = \phi$; if $S_1 = C$, then $\theta_1 = \phi_2$; if $S_1 = G$, then $\theta_1 = \phi_3$; if $S_1 = T$, then $\theta_1 = \phi_4$. Let the angle from $\overrightarrow{OX}$ to $\overrightarrow{P_{i-1}P_i}$ in counterclockwise be denoted by $\phi(S_i)$, if $S_i = A$, then $\phi(S_i) = \phi_1$; if $S_i = C$, then $\phi(S_i) = \phi_2$; if $S_i = G$, then $\phi(S_i) = \phi_3$; if $S_i = T$, then $\phi(S_i) = \phi_4$. Obviously, $\theta_1 = \phi(S_1)$. By Figure 7-2 and the sine theorem, we have

$$\frac{r_{i-1}}{sin[\theta_i - \phi(S_i)]} = \frac{r_i}{sin[\theta_{i-1} - \phi(S_i)]}$$

$$\theta_i = \phi(S_i) + arcsin\left\{\frac{r_{i-1}}{r_i} sin[\theta_{i-1} - \phi(S_i)]\right\} \quad (i = 2, \cdots n) \quad (13)$$

We should pay attention to that the length and direction of vector $\overrightarrow{P_{i-1}P_i}$ vary with the position parameter $i$ and base parameter $S_i$, $i = 1, 2, \cdots n$ i.e. the fold lines $P_0P_1P_2\cdots P_n$ include almost of the information about the corresponding DNA sequences. It is especially important that these graphical representations of DNA sequences do not have degeneracy, because $P_1, P_2, \cdots P_n$ locate at concentric circles, where the radius sequence $\{r_i\}$ is an increasing sequence, the pole O is the center of circles, and the directed walk

$P_0P_1P_2\cdots P_n$ must not form a circuit. In actual applications, we may select appropriate sequence numbers $\{r_i\}$ and parameters $\phi_i(i = 1, 2, 3, 4)$ according to the discussed problem. For example, if $\{r_i\}$ be taken $i/(i + 9)$, $\phi_i = (n - 1)\pi/3$ $(i = 1, 2, 3, 4)$, then, in the presence of electronic computer, the graphical representations of DNA sequences could be plotted easily in a quarter of a unit circle as per the recursive formula (13). Based on the graphical representations of DNA sequences, some new invariant can be introduced to some numerical characterization of DNA sequences. We will propose a new measure of the dispersion of DNA graphs that can be used to quantify the differences between two or more graphs of genes of various organisms. The technique is an improvement over the existing ones in that differences in sequence length are normalized out for a more acceptable comparison between sequences of differing lengths. It also appears that, once standardized, the proposed scheme may help in rapid identifying and retrieving of molecular sequences from electronic sequence libraries and in studying molecular phylogency in evolutionary time scales.

## ACKNOWLEDGMENT

This work is partially supported by a grant from the National Science Foundation of China. The authors sincerely appreciate the encouraging comments from the Editor of the journal on this paper. They also wish to thank an anonymous referee of this paper who provided many useful and constructive suggestions for the improvement of the paper. Finally, thanks to all the authors that appear in the references.

## REFERENCES AND NOTES

(1) Hamori, Ruskin. J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Bio. Chem.* **1983**, *258*, 1318−1327.

(2) Gates, M. A. A simple way to look at DNA. *J. Theor. Biol.* **1986**, *119*, 319−328.

(3) Nandy. A. A new graphical representation and analysis of DNA sequence structure I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309−313.

(4) Nandy. A. Graphical analysis of DNA sequence structure:*III*. Indications of evolutionary distinctions and characteristics of intros and exons. *Curr. Sci.* **1996**, *70*, 611−668.

(5) Leong. P. M.; Morgenthaler. S. Random walk and gap plots of DNA sequences. *Comput. Applic. Biosci.* **1995**, *11*, 503−507.

(6) Nandy, A.; Nandy, P. Graphical analysis of DNA sequence structure *II*: Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Curr. Sci.* **1995**, *68*, 75−85.

(7) Blatsdell. B. E.; Campbel. A. M.; Karlin. S. Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5854−5859.

(8) Randic. M.; Vracko. M.; Nandy. A.; Basak. S. C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−1244.

(9) Guo, X.; Randic, M.; Basak, S. C. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy, *Chem. Phys. Lett.* **2002**, *350*, 106−112.

CI010017G