New 2-D graphical representation of DNA sequences

Jiasong Wang[a,b] and Weiqun Wang[a,*]

[a]Department of Human Nutrition, Kansas State University, Manhattan, KS 66506, USA and [b]Department of Mathematics, Nanjing University, Nanjing 210008, China

*Corresponding author. 785-532-0153, fax: 785-532-3132, email: wwang@ksu.edu

Abstract

In this paper we present a novel 2-D graphical representation of DNA sequences in the first quadrant Cartesian coordinate system. This representation has been mathematically proved to have a zero length of circuit, i.e., without any degeneracy. Given any point considered in a DNA sequence, the number of A, T, G, and C from the starting point could be calculated based upon an iterative comparison method. Furthermore, this graphical representation, in comparison with our and others previous methods, shows a unique application to the universal genetic code with double stochastic matrices in nature.

Keywords: 2-D graphical representation; DNA sequences; Degeneracy; Genetic code; Double stochastic matrices.

1. Introduction

Various 2-D graphical representations of DNA sequences as previously reported by Gates (1985), Nandy (1994), and Leong and Morgenthaler (1995) have provided an easy approach to visualize DNA strand, but there was a high degeneracy. Guo et al. (2001) and later Liu et al. (2002) reported a 2-D graphical representation with lower or nondegeneracy. We previously demonstrated a 2-D graphical representation of DNA strand using a two-quadrant Cartesian coordinates system resolved sequences' degeneracy (Yau et al., 2003). In this paper we report a novel 2-D graphical representation of DNA sequences by using the first quadrant of the Cartesian coordinate plane that is mathematically proved non-degeneracy. A practicable formula has been developed to calculate the frequencies of A, T, G, and C from the starting point to any considered point. A unique application of this positive numerical description of the nucleotides to the genetic codons has been further performed.

2. Methods and results

2.1. To construct a new graphical representation of DNA sequences in the first quadrant of the Cartesian coordinate plane

The unit vectors representing four nucleotides A, G, C, and T are as follows: $(1,0)\rightarrow$A, $(\sqrt{3}/2,1/2)\rightarrow$G, $(1/2,\sqrt{3}/2)\rightarrow$C, $(0,1)\rightarrow$T, that means OA lies on x-axis, the angle between OG and x-axis is 30 degrees, the angle between OC and x-axis is 60 degrees, OT stands up on y-axis, and |OA|=|OG|=|OC|=|OT|=1.

As shown in Figure 1, we set purines (A and G) under the bisector OQ of the first quadrant and pyrimidines (T and C) above the bisector OQ of the first quadrant.

To get the graphical representation of a DNA strand, we assume that S_1, S_2, …, S_n stand for a DNA sequence length at n, where S_i belongs to {A,T,C,G}. The sequence of the points, P_1, P_2, …, P_n will then be constructed as the vector $P_{i-1}P_i$ that corresponds to S_i, where P_0 is the origin, and $|P_{i-1}P_i|=1$. If S_i=A, then $P_{i-1}P_i$ is parallel to x-axis; if S_i=T, then $P_{i-1}P_i$ is parallel to y-axis; if S_i=G, then the angle between $P_{i-1}P_i$ and the ray from P_{i-1} parallel to x-axis is 30 degrees; if S_i==C then the angle between $P_{i-1}P_i$ and the ray from P_{i-1} parallel with x-axis is 60 degrees.

To get the numerical sequence of the points, P_1, P_2, …, P_n corresponding to S_1, S_2, …, S_n, we introduce a two dimensional array x(i) i=1, 2, …, n, and y(i), i=1, 2, …, n, and P_i =(x(i), y(i)). If S_i =A, then $P_i = P_{i-1}$+(1,0); if S_i =G, then $P_i = P_{i-1}$+($\sqrt{3}$/2,1/2); if S_i =C, then $P_i = P_{i-1}$+(1/2,$\sqrt{3}$/2) and if S_i =T, then $P_i = P_{i-1}$+(0,1), where i=1, 2, …, n, and P_0 =(0,0). The MATLAB code for the numerical sequence (x(i),y(i)) and the graphical representation of the DNA strand has been included in the appendix. A computational result of the first exon of both human and mouse β-globin gene by this DNA graph representation was shown in Figure 2.

2.2. To prove non-degeneracy in this new 2-D graphical representation of DNA strand

We assume that the number of nucleotides forming a degeneracy is n, and f_A, f_G, f_C, and f_T are the frequencies corresponding to the number of appearances of A, G, C and T in the circuit, respectively. Hence, $f_A + f_G + f_C + f_T = n$. Because f_AA, f_GG, f_CC, and f_TT form a circuit, the following equation holds:

$$f_A(1,0) + f_G(\sqrt{3}/2,1/2) + f_C(1/2,\sqrt{3}/2) + f_T(0,1) = 0$$

i.e.

$$2f_A + \sqrt{3}\, f_G + f_C = 0 \qquad [1]$$

$$f_G + \sqrt{3} f_C + 2 f_T = 0 \qquad [2]$$

Clearly [1] and [2] hold if, and only if, $f_A = f_G = f_C = f_T = 0$. So, $n = f_A + f_G + f_C + f_T = 0$, which implies no circuit exists in the graphical representation, i.e. there is no degeneracy in this representation.

Given any point, P=(x,y), on the graphical representation, the frequencies of f_A, f_G, f_C and f_T, representing the number of A, G, C and T, respectively, from the starting point, should meet with the equations as follows:

$$2f_A + \sqrt{3}\ f_G + f_C = 2x \qquad [3]$$

$$f_G + \sqrt{3} f_C + 2\ f_T = 2y \qquad [4]$$

Because 2x and 2y are irrational numbers as m+$\sqrt{3}$n, where m and n are integers, [3] and [4] can be rewritten as:

$$2f_A + \sqrt{3}\ f_G + f_C = m_x + \sqrt{3} n_x, \qquad [5]$$

$$f_G + \sqrt{3} f_C + 2\ f_T = m_y + \sqrt{3} n_y. \qquad [6]$$

By solving [5] and [6] we obtain:

$f_G = n_x$, $f_C = n_y$, $f_A = (m_x - n_y)/2$, $f_T = (m_y - n_x)/2$.

To get m_x, n_x and m_y, n_y from 2x and 2y, we design an iterative comparison method to compare the decimal parts of 2x with multiples of the decimal of $\sqrt{3}$ many times to get n_x, then $m_x = 2x - \sqrt{3}\ n_x$. Similarly, we can obtain m_y and n_y, then get f_A, f_G, f_C and f_T.

2.3. To apply this new representation to the genetic codons

Recently, He (2004) developed three genetic code-based matrices by using three genetic attribute equivalences. For example, 64 genetic codons could be plugged into one biperiodical table as GG(i,j) (i,j=1, 2, … 8):

$$GG = \left| \begin{array}{cccccccc} \text{CCC} & \text{CCA} & \text{CAC} & \text{CAA} & \text{ACC} & \text{ACA} & \text{AAC} & \text{AAA} \\ \text{CCU} & \text{CCG} & \text{CAU} & \text{CAG} & \text{ACU} & \text{ACG} & \text{AAU} & \text{AAG} \\ \text{CUC} & \text{CUA} & \text{CGC} & \text{CGA} & \text{AUC} & \text{AUA} & \text{AGC} & \text{AGA} \\ \text{UCC} & \text{UCA} & \text{UAC} & \text{UAA} & \text{GCC} & \text{GCA} & \text{GAC} & \text{GAA} \\ \text{CUU} & \text{CUG} & \text{CGU} & \text{CGG} & \text{AUU} & \text{AUG} & \text{AGU} & \text{AGG} \\ \text{UCU} & \text{UCG} & \text{UAU} & \text{UAG} & \text{GCU} & \text{GCG} & \text{GAU} & \text{GAG} \\ \text{UUC} & \text{UUA} & \text{UGC} & \text{UGA} & \text{GUC} & \text{GUA} & \text{GGC} & \text{GGA} \\ \text{UUU} & \text{UUG} & \text{UGU} & \text{UGG} & \text{GUU} & \text{GUG} & \text{GGU} & \text{GGG} \end{array} \right| \qquad [7]$$

then, three genetic code based matrices could be obtained by using G=U=0, A=C=1 to represent amino-mutating absence/present (0,1)-equivalence, C=U=1, A=G=2 to represent pyrimidine/purine ring-based (1,2)-equivalence, or A=U=2, C=G=3 to represent hydrogen bonds-based (2,3)-equivalence. Among the three genetic code based matrices, first two are stochastic matrices and the last one is a double stochastic matrix.

We extend the idea of He's to our 2-D numerical representation. The numerical representations of four nucleotides in 2-D graphical representation of DNA strand can be considered as two 1-D numerical representations of A, G, C and U in RNA.

i.e.

$(1,0) \rightarrow \text{A}, \quad (\sqrt{3}/2, 1/2) \rightarrow \text{G}, \quad (1/2, \sqrt{3}/2) \rightarrow \text{C}, \quad (0,1) \rightarrow \text{U}$

for x axis projections: $\quad 1 \rightarrow \text{A}_x,\ \sqrt{3}/2 \rightarrow \text{G}_x,\ 1/2 \rightarrow \text{C}_x,\ 0 \rightarrow \text{U}_x \qquad [8]$

for y axis projections: $\quad 0 \rightarrow \text{A}_y,\ 1/2 \rightarrow \text{G}_y,\ \sqrt{3}/2 \rightarrow \text{C}_y,\ 1 \rightarrow \text{U}_y \qquad [9]$

By using two 1-D numerical representations of x and y axis projections, we compute the genetic code based matrices GGx and GGy, respectively, and find if

$$GG4 = GGx + GGy, \qquad [10]$$

then

$$GG4 = \begin{vmatrix} p & q & q & r & q & r & r & s \\ q & p & r & q & r & q & s & r \\ q & r & p & q & r & s & q & r \\ q & r & r & s & p & q & q & r \\ r & q & q & p & s & r & r & q \\ r & q & s & r & q & p & r & q \\ r & s & q & r & q & r & p & q \\ s & r & r & q & r & q & q & p \end{vmatrix}$$

where p = 4.0981, q = 3.7321, r = 3.3660 and s = 3.0000. The sum of each column's entries in GG4 is equal to the sum of each row's entries, i.e. 28.3923. Therefore, GG4 is a double stochastic matrix.

Due to the linearity of the genetic code formula [10], the computation of GG4 resulted in $A_x + A_y = U_x + U_y = 1$, and $C_x + C_y = G_x + G_y = (1+\sqrt{3})/2$. It is obvious that a double stochastic matrix for the genetic code property (He, 2004) should be obtained as long as A = U and C = G in despite of any assigned numbers.

4. Discussion and conclusion

Graphical representation of DNA sequence may provide a simple way of viewing, sorting and comparing various gene structures. The previous studies of the 2-D graphical representations of DNA sequence used four quadrants and placed four nucleotides along with four axes of coordinate system usually had high degeneracy. In order to overcome the defect of degeneracy, we recently presented a graphical representation of NDA strand without degeneracy by using two quadrant Cartesian coordinates system (Yau et al., 2003).

In comparison with our and others previous methods, we present a new 2-D graphical representation of DNA sequence by performing data in the first quadrant of Cartesian coordinate system in this paper. We proved that there was not any degeneracy. With an iterative comparison method, the number of A, G, C and T in any point in the sequence from the starting point can be calculated. Furthermore, the number of purines and pyrimidines could be easily compared via visualizing the total points separated by the bisector OQ. If we get a graphical representation of one DNA sequence, its symmetric graph along the OQ line should be its complementary strand. Finally, the usage of this graphical representation in the first quadrant may provide a unique approach from DNA sequence to RNA universal genetic code. A potential application with this new 2-D representation method into 64 genetic codons has been performed to generate a double stochastic matrix, which appears to be similar as the repeated genetic code informatics in nature.

Acknowledgements

The authors would like to thank Dr S. Yau, Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, and Dr S. Yang, Department of Statistics, Kansas State University, for their invaluable discussions and suggestions. This study was supported in part by grants from the Terry C. Johnson Center for Basic Cancer Research, Kansas State University, the NIH COBRE Award 1P20 RR15563 and Kansas State matching support, and the NIH 1R01 CA106397-01. This is a journal paper No. 04-343-J of the Kansas Agriculture Experiment Station, Kansas State University.

References

Gates, M. A., 1985. Simple DNA sequence representations. *Nature 316*, 219.

Guo, X., Randic, M., Basak, S. C., 2001. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Let. 350*, 106-112.

He, M., 2004. Genetic code, attributive mappings and stochastic matrices. *Bull. Math. Biol. 66*, in press.

Leong, P. M., Morgenthaler, S., 1995. Random walk and gap plot of DNA sequences. *Comput. Applic. Biosci. 11*, 503-507.

Liu, Y., Guo, X., Xu, J., Pan, L., Wang, S., 2002. Some notes on 2-D graphical representation of DNA sequence. *J. Chem. Inf. Omput. Sci. 42*, 529-533.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure: 1.methodology and application to globin genes. *Curr. Sci. 66*, 309-313.

Yau, S., Wang, J., Niknejad, A., Lu, C., Jin, N., Ho, Y., 2003. DAN sequence representation without degeneracy. *Nucleic Acids Res. 31*, 3078-3080.

Figure Legend:

Figure 1. The unit vectors representing A, T, G, and C in the first quadrant Cartesian coordinate system.

Figure 2. Two-dimensional graphs of both human (solid circle) and mouse β-globin exon-1 gene sequence (open circle). The β-globin sequences were obtained from NCBI GenBank at AF527577 and J00413 for human and mouse, respectively.

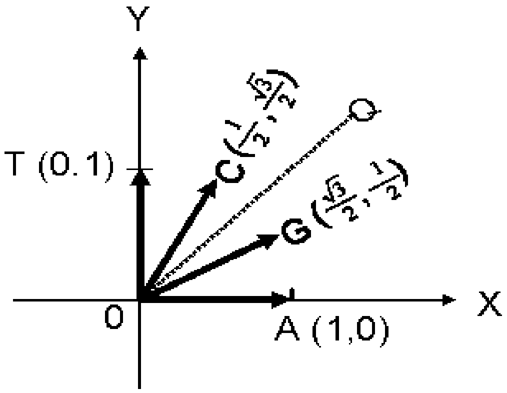
Y
T (0.1)
C($\frac{1}{2}$, $\frac{\sqrt{3}}{2}$)
Q
G($\frac{\sqrt{3}}{2}$, $\frac{1}{2}$)
0
A (1,0)
X

Figure 1

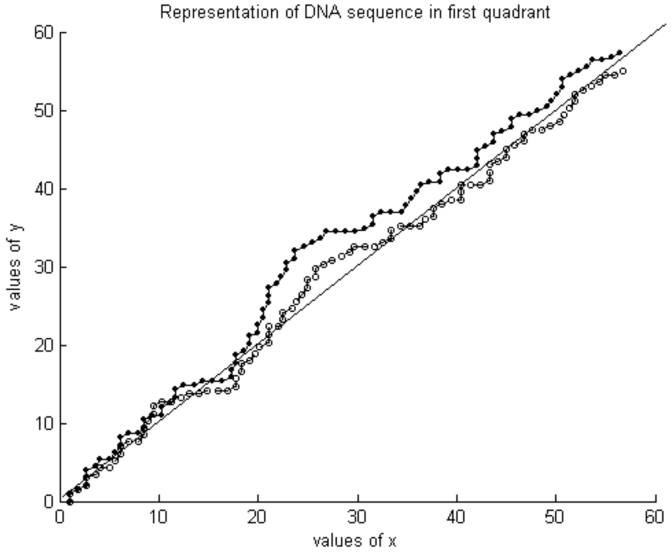
Representation of DNA sequence in first quadrant
values of y
values of x
0
10
20
30
40
50
60
0
10
20
30
40
50
60

Figure 2

Appendix

MATLAB code

```
Function        [x,y]=representation(DNA strand,n)
% This function computes the summations of numerical representation
% values of points, one by one, on the DNA strand in first quadrant,
% then, give the output of the graphical representation of it.
% The input array, DNA strand, stores a DNA strand, of length n, which needs input.
% Output array x stores the x-coordinates of the summations just introduced.
% Output array y stores the y-coordinates of the summations just introduced.
fid=fopen('DNA strand','r');
gene=fscanf(fid,'%1s',inf);
n=  ;
gene=upper(gene);
x=zeros(1,n);
y=zeros(1,n);
b=sqrt(3);
% First step computing
if strcmp(gene(1),’A’)==1
      x(1)=1;
      y(1)=0;
      elseif strcmp(gene(1),’T’)==1
      x(1)=0;
      y(1)=1;
      elseif strcmp(gene(1),’C’)==1
     x(1)=1/2;
      y(1)=b/2;
      elseif strcmp(gene(1),’G’)==1
      x(1)=b/2;
      y(1)=1/2;
   end
% Iterations
for m=2:n
   if strcmp(gene(m),'A')==1
      x(m)=x(m-1)+1;
      y(m)=y(m-1);
      elseif strcmp(gene(m),'T')==1
      x(m)=x(m-1);
      y(m)=y(m-1)+1;
      elseif strcmp(gene(m),'C')==1
      x(m)=x(m-1)+1/2;
      y(m)=y(m-1)+b/2;
      elseif strcmp(gene(m),'G')==1
      x(m)=x(m-1)+b/2;
      y(m)=y(m-1)+1/2;
   end
```

```
end
% Plot the graphical representation of DNA strand in first
quadrant
plot(x,y,'r*:')
title('Representation of DNA sequence in first quadrant')
xlabel('values of x')
ylabel('values of y')
```