Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 249 (2007) 681–690
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

New 3D graphical representation of DNA sequence based on dual nucleotides

Xiao-Qin Qi*, Jie Wen, Zhao-Hui Qi

Department of Computer and Information Engineering, Shijiazhuang Railway Institute, Hebei, Shijiazhuang 050043, People's Republic of China

Received 22 January 2007; received in revised form 26 August 2007; accepted 27 August 2007
Available online 1 September 2007

Abstract

We introduce a 3D graphical representation of DNA sequences based on the pairs of dual nucleotides (DNs). Based on this representation, we consider some mathematical invariants and construct two 16-component vectors associated with these invariants. The vectors are used to characterize and compare the complete coding sequence part of beta globin gene of nine different species. The examination of similarities/dissimilarities illustrates the utility of the approach.

Keywords: 3D dual nucleotide curve; 3D quantitative characterization; Euclidean distance; Complete coding sequence; Similarities/dissimilarities

1. Introduction

The number of biological sequences is rapidly increasing in the biological database. It is one of the challenges for bio-scientists to analyze mathematically the large volume of biological sequence data. It is good to use the graphic representation to study complicated biological systems because it can provide an intuitive picture and help people gain useful insights. Similar graphical approaches have also been used to deal with a wide variety of biological problems. For instance, various graphic approaches have been successfully used to study enzyme-catalyzed system (see, e.g., King and Altman, 1956; Chou et al., 1979; Chou and Forsen, 1980; Chou and Liu, 1981; Zhou and Deng, 1984, Chou, 1989, 1990; Kuzmic et al., 1992; Lin and Neet, 1990), protein folding kinetics (Chou, 1990, 1993), condon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), HIV reverse transcriptase inhibition mechanisms (Althaus et al., 1993a–c) and base frequency distribution in the anti-sense strands (Chou and Zhang, 1996). Recently, the images of cellular automata were also used to represent biological sequences (Xiao et al., 2005a), predict protein subcellular location (Xiao et al., 2006a), investigate HBV virus gene missense mutation (Xiao et al., 2005b) and HBV viral infections (Xiao et al., 2006b), as well as analyze the fingerprint of SARS coronavirus (Wang et al., 2005).

As for an important part of graphical techniques, graphical representations of DNA sequences have been proposed by several authors (Zhang, 1991; Nandy, 1994; Nandy and Nandy, 2003; Liao and Wang, 2004; Randic et al., 2003, Liu et al., 2006; Zhang and Chen, 2006). Some of them, for example Nandy' graphical representation (Nandy, 1994), are accompanied by some loss of visual information associated with crossing and overlapping of the curve with itself. In order to avoid the limitations related to crossing and overlapping, Liao (Liao and Wang, 2004) and Randic (Randic and Vracko, 2000; Randic et al., 2003) 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.

Moreover, the dinucleotide analysis has also been tried by several previous authors. Randic (2000) proposed a condensed representation of DNA based on pairs of nucleotides. This approach can offer fast, qualitative comparisons of DNA and allow quantitative comparisons of DNA from different sources. Wu et al. (2003) and Liu et al. (2006) proposed their analysis approaches based on neighboring nucleotides of DNA sequence, which reveal

*Corresponding author.
E-mail addresses: xiao_papers@yahoo.com.cn (X.-Q. Qi), zhqi.papers@yahoo.com.cn (Z.-H. Qi).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.08.025

the biology information hidden between the dual nucleotides (DNs). Qi and Qi (2007) also suggest a dinucleotide analysis method to reveal the biology information of DNA sequences.

Recently, Qi and Fan (2007) proposed a 3D graphical representation of DNA sequence based on a pair of nucleotides. Based on similar research object (3D graphical representation of DNA sequence based on a pair of nucleotides), in this paper we introduce a new 3D graphical representation (3D-DN curve) of DNA primary sequences, in which there is also no loss of information in the transfer of data from a DNA sequence to its mathematical representation. Our representation is different from that of PN-curve (Qi and Fan, 2007). The two papers are highly dissimilar with respect to each other in many aspects: the methods and contents of research, the map used to construct graphical representation, the graphical curve and numerical invariants characterizing DNA sequences. The introduced representation is simple and direct, and gives us more biology information based on DNs.

2. 3D graphical representation of DNA sequences based on dual nucleotides

Given a DNA primary sequence, there are 16 kinds of the pairs of the neighboring nucleotides. These pairs can be classified as four categories based on their chemical properties: purine-DN {AG, GA}/pyrimidine-DN {CT, TC}, amino-DN {AC, CA}/keto-DN {GT, TG}, weak-H bond DN {AT, TA}/strong-H bond DN {CG, GC} and repeat-DN {AA, CC, GG, TT}. Then we design a 4×4 matrix and give a new 3D graphical representation of DNA sequences. We arrange 16 DNs in a 4×4 matrix according to the above four categories. The matrix is

$$\begin{bmatrix} \mathrm{AG} & \mathrm{GA} & \mathrm{CT} & \mathrm{TC} \\ \mathrm{AC} & \mathrm{CA} & \mathrm{GT} & \mathrm{TG} \\ \mathrm{AT} & \mathrm{TA} & \mathrm{CG} & \mathrm{GC} \\ \mathrm{AA} & \mathrm{CC} & \mathrm{GG} & \mathrm{TT} \end{bmatrix}.$$

Every element of the matrix has a corresponding index (i,j), $i=0,1,2,3$; $j=0,1,2,3$. Based on the index, we assign one DN as follows:

$$\begin{aligned} &(0,0,0)\to \mathrm{AG},\ (0,1,0)\to \mathrm{GA},\ (0,2,0)\to \mathrm{CT},\ (0,3,0)\to \mathrm{TC},\\ &(1,0,0)\to \mathrm{AC},\ (1,1,0)\to \mathrm{CA},\ (1,2,0)\to \mathrm{GT},\ (1,3,0)\to \mathrm{TG},\\ &(2,0,0)\to \mathrm{AT},\ (2,1,0)\to \mathrm{TA},\ (2,2,0)\to \mathrm{CG},\ (2,3,0)\to \mathrm{GC},\\ &(3,0,0)\to \mathrm{AA},\ (3,1,0)\to \mathrm{CC},\ (3,2,0)\to \mathrm{GG},\ (3,3,0)\to \mathrm{TT}. \end{aligned}$$

That is to say, we assign every DN to its corresponding index (x,y), respectively, while the corresponding curve extending along with z-axes. In detail, let $G=g_1g_2\ldots$ be an arbitrary DNA primary sequence. Then we define a map ϕ as follows:

$$\phi(g_ig_{i+1})=\begin{cases} (0,0,i) & \text{if } g_ig_{i+1}=\mathrm{AG},\\ (0,1,i) & \text{if } g_ig_{i+1}=\mathrm{GA},\\ (0,2,i) & \text{if } g_ig_{i+1}=\mathrm{CT},\\ (0,3,i) & \text{if } g_ig_{i+1}=\mathrm{TC},\\ (1,0,i) & \text{if } g_ig_{i+1}=\mathrm{AC},\\ (1,1,i) & \text{if } g_ig_{i+1}=\mathrm{CA},\\ (1,2,i) & \text{if } g_ig_{i+1}=\mathrm{GT},\\ (1,3,i) & \text{if } g_ig_{i+1}=\mathrm{TG},\\ (2,0,i) & \text{if } g_ig_{i+1}=\mathrm{AT},\\ (2,1,i) & \text{if } g_ig_{i+1}=\mathrm{TA},\\ (2,2,i) & \text{if } g_ig_{i+1}=\mathrm{CG},\\ (2,3,i) & \text{if } g_ig_{i+1}=\mathrm{GC},\\ (3,0,i) & \text{if } g_ig_{i+1}=\mathrm{AA},\\ (3,1,i) & \text{if } g_ig_{i+1}=\mathrm{CC},\\ (3,2,i) & \text{if } g_ig_{i+1}=\mathrm{GG},\\ (3,3,i) & \text{if } g_ig_{i+1}=\mathrm{TT}. \end{cases}$$

The map ϕ maps G into a plot set. For example, the corresponding plot set of the sequence ATGGTGCACC is {(2, 0, 1), (1, 3, 2), (3, 2, 3), (1, 2, 4), (1, 3, 5), (2, 3, 6), (1, 1, 7), (1, 0, 8), (3, 1, 9)}. The corresponding plot set is called as characteristic plot set. The curve connected all plots of the characteristic plot set in turn is called 3D-DN curve. In Table 1 and Fig. 1, we show the corresponding coordinates and the 3D graphical representation of the sequence, respectively.

From the construction of the 4×4 matrix, we know that their designs are not unique. There are 16 kinds of DNs, so they have 16! combinations. But we design the 4×4 matrix based on the classifications of nucleotides. In this paper, we only consider the above 4×4 matrix to illustrate our scheme.

3. Numerical characterization of DNA sequences

Given a DNA sequence with the length of N. Based on the definition of the map ϕ, we can have a set of points (x_i,y_i,z_i), $i=1,2,\ldots,N-1$, and the correspondence

Table 1
Cartesian 3D coordinates for the sequence ATGGTGCACC of the coding sequence of the first exon of human β-globin gene

Base	DNs	x	y	z
1	AT	2	0	1
2	TG	1	3	2
3	GG	3	2	3
4	GT	1	2	4
5	TG	1	3	5
6	GC	2	3	6
7	CA	1	1	7
8	AC	1	0	8
9	CC	3	1	9