Available online at www.sciencedirect.com

ScienceDirect

Journal of Theoretical Biology 243 (2006) 555–561

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

PNN-curve: A new 2D graphical representation of DNA sequences and its application☆

Xiao Qing Liu[a,*], Qi Dai[b], Zhilong Xiu[a], Tianming Wang[b]

[a]Department of Bioscience and Biotechnology, School of Environmental and Biological Science and Technology, Dalian University of Technology, Dalian 116024, PR China
[b]Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PR China

Received 13 May 2006; accepted 20 July 2006
Available online 24 July 2006

Abstract

We introduce a novel 2D graphical representation of DNA sequences based on the pairs of the neighboring nucleotides (PNNs). Then we get the PNNs' distributions and obtain a y-M. The construction of the PNN-curve has some important advantages (1) It avoids loss of information and the PNN-curve standing for DNA sequences does not overlap or intersect with itself. (2) The novel 2D representation is more sensitive. The utility of this method can be illustrated by the examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of eleven different species in Table 2.

Keywords: DNA; Graphical representation; y-M; PNN-curve; Similarity

1. Introduction

The number of DNA sequences as strings of four nucleotides: A (adenine), C (cytosine), G (guanine), T (thymine) is growing rapidly in the DNA database. But it is difficult to obtain information from DNA sequences directly. Therefore, many kinds of methods have been proposed to characterize the DNA sequences.

Several researchers introduced an alternative way to compare the DNA sequences, based on a set of invariants of DNA sequences, rather than directly using string comparison. Graphical representation of DNA sequences provides a simple way of viewing, sorting, and comparing various gene structures (see Chi and Ding, 2005; Dai et al., 2006a; Gates, 1986; Guo et al., 2001; Guo and Nandy, 2003; Hamori and Ruskin, 1983; Hamori, 1989; He and Wang, 2002; Li and Wang, 2003; Liao and Wang, 2004; Liao et al., 2005; Nandy, 1994a,b; Nandy and Nandy, 2003; Randić et al., 2003; Randić, 2004; Wu et al., 2003; Yao and Wang, 2004). Nandy (1994a) presented a graphical representation by assigning A (adenine), G (guanine), T (thymine), and C (cytosine) to four direction $(-x), (+x), (-y), (+y)$, respectively. Such a representation of DNA sequences is accompanied with (1) some loss of visual information associated with crossing and overlapping of the curve with itself; (2) an arbitrary decision with respect to the choice of the direction for the bases. Recently several authors have outlined different graphical representations of DNA sequences based on 2D, 3D or 4D (see Chi and Ding, 2005; Dai et al., 2006a; Li and Wang, 2003; Liao and Wang, 2004; Liao et al., 2005; Nandy and Nandy, 2003; Randić et al., 2003; Randić, 2004; Wu et al., 2003; Yao and Wang, 2004), but some representations (see Guo and Nandy, 2003; Wu et al., 2003; Nandy and Nandy, 2003) are also accompanied with some loss of information due to overlapping and crossing of the curve with itself. Then they constructed the D/D, L/L and high order matrices to extract the leading eigenvalues as the invariants to characterize the DNA sequences, so their computation is very complex. Furthermore, they regarded a DNA

☆Work is supported by National Natural Science Foundation of China (10571019).

*Corresponding author. Tel.: +86 04184749327.

E-mail addresses: lxq785348@163.com (X.Q. Liu), zhlxiu@dlut.edu.cn (Z.L. Xiu), wangtm@dlut.edu.cn (T.M. Wang).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.07.018

sequence as a random sequence consisting of A, G, C and T, which may ignore some information hidden between the neighboring nucleotides.

In this paper, we introduce a new method to construct 2D graphical representation (PNN-curve) of DNA primary sequences based on the pairs of the neighboring nucleotides (PNNs), which can give us more information on the DNA sequences. We also get the PNNs' distributions and the new invariants to numerically characterize the DNA primary sequences.

2. PNN-curve: 2D graphical representation of DNA sequences

2.1. Construction of the PNN-curve

Given a DNA primary sequence, there are 16 kinds of the pairs of the neighboring nucleotides. For notational convenience, let PNN denote a pair of the neighboring nucleotides. Then we design the 4 × 4 cells and systems, and give a novel 2D graphical representation of DNA sequences. From the knowledge of biology, we know that the four DNA bases can be classified as $R = \{A, G\}$ and $Y = \{C, T\}$, $M = \{A, C\}$ and $K = \{G, T\}$, $W = \{A, T\}$ and $S = \{G, C\}$ according to their chemical properties. Take the second classification for example, we design the 4 × 4 cells and system. If $G = g_1g_2, \ldots, g_n$ is an arbitrary DNA primary sequence, we arrange 16 PNNs in a 4 × 4 cell. The first and second columns are labelled in the order MM and MK (AA, AC, AG, AT, CA, CC, CG, CT); the third and fourth columns are labelled in the order KM and KK (GA, GC, GG, GT, TA, TC, TG, TT). These points are separated by one unit which is shown in Fig. 1(a). All the 4 × 4 cells are arranged in the horizontal direction, and the adjacent cells are separated by one unit, the corresponding system is represented in Fig. 1(b). For example, let $S = ATGGTGC$ be the first seven bases of the coding sequence of the Human's β-globin gene. We assign the PNN AT to the position AT of the first 4 × 4 cell, ..., the PNN GC to the position GC in the sixth cell, and connect the adjacent positions one by one, the PNN-curve of the sequences S is obtained in Fig. 1(c).

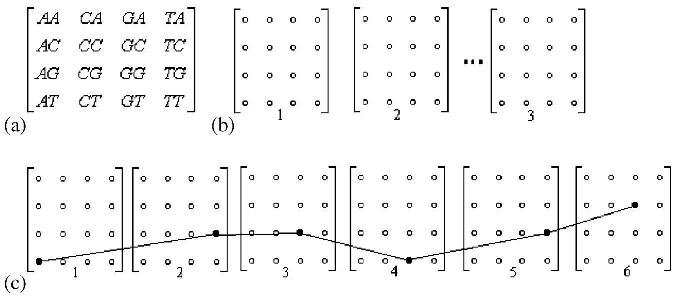

Fig. 1. The new graphical representation of the sequence ATGGTGC: (a) 4 × 4 cell; (b) system; (c) the PNN-curve of the sequence ATGGTGC.

Table 1
Cartesian 2D-coordinates for the sequence ATGGTGCACC of the coding sequence of the first exon of the human β-globin gene

Base	1	2	3	4	5	6
PNNs	AT	TG	GG	GT	TG	GC
x	1	8	11	15	20	23
y	1	2	2	1	2	3

From the construction of the 4 × 4 cells, we know that their designs are not unique. There are 16 kinds of PNNs, hence they have 16! combinations. But we design the 4 × 4 cells according to the classifications of nucleotides, there are only 48 combinations left. Here, we only consider the above 4 × 4 cells to illustrate our method. We can easily find a map between the PNNs and their positions in the coordinate system. If $G = g_1g_2, \ldots, g_n$ is an arbitrary DNA primary sequence. Then we define a map ψ as follows:

$$\psi(g_ig_{i+1}) = \begin{cases} (4 \times i - 3, 4) & \text{if } g_ig_{i+1} = AA, \\ (4 \times i - 3, 3) & \text{if } g_ig_{i+1} = AC, \\ (4 \times i - 3, 2) & \text{if } g_ig_{i+1} = AG, \\ (4 \times i - 3, 1) & \text{if } g_ig_{i+1} = AT, \\ (4 \times i - 2, 4) & \text{if } g_ig_{i+1} = CA, \\ (4 \times i - 2, 3) & \text{if } g_ig_{i+1} = CC, \\ (4 \times i - 2, 2) & \text{if } g_ig_{i+1} = CG, \\ (4 \times i - 2, 1) & \text{if } g_ig_{i+1} = CT, \\ (4 \times i - 1, 4) & \text{if } g_ig_{i+1} = GA, \\ (4 \times i - 1, 3) & \text{if } g_ig_{i+1} = GC, \\ (4 \times i - 1, 2) & \text{if } g_ig_{i+1} = GG, \\ (4 \times i - 1, 1) & \text{if } g_ig_{i+1} = GT, \\ (4 \times i, 4) & \text{if } g_ig_{i+1} = TA, \\ (4 \times i, 3) & \text{if } g_ig_{i+1} = TC, \\ (4 \times i, 2) & \text{if } g_ig_{i+1} = TG, \\ (4 \times i, 1) & \text{if } g_ig_{i+1} = TT. \end{cases}$$

Let us consider the above sequence (S) again, the corresponding coordinates of its nucleotides are illustrated in Table 1.

2.2. Comparison with other 2D graphical representations

In this section, we will show some advantages of the PNN-curve by comparing with other graphical representations.

(1) As we all know, a DNA sequence contains much information about the evolution. Recently some researchers (see Chi and Ding, 2005; Dai et al., 2006a; Gates, 1986; Guo et al., 2001; Guo and Nandy, 2003; Liao and Wang, 2004; Liao et al., 2005; Randić, 2004; Wu et al., 2003; Yao and Wang, 2004) regarded a DNA sequence as a random sequence consisting of A, G, C and T. But as pointed in Dai et al. (2006b), the neighboring bases can characterize the DNA sequences better, which is also shown in Randić (2004) and He and Wang (2002). Here, we give a novel graphical representation based on the PNNs, it maybe gives us more information which others may have ignored on the DNA primary sequences.