Physica A 456 (2016) 256–270

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

DUC-Curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment

Yushuang Li [a], Qian Liu [a], Xiaoqi Zheng [b,*]

[a] College of Science, Yanshan University, Qinhuangdao 066004, China

[b] Department of Mathematics, Shanghai Normal University, Shanghai 200234, China

HIGHLIGHTS

- A highly compact and simple 2D graphical representation of DNA sequences is constructed.
- This graphical representation could directly detect nucleotide, di-nucleotide compositions and microsatellite structure from DNA sequences.
- This model can be applicable to DNA sequence alignment.
- The obtained reliable results show that the proposed method is very effective in biological sequence comparison.

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form 18 February 2016
Available online 2 April 2016

Keywords:
Graphical representation
DNA sequence
DUC-Curve
Sequence alignment
Sequence comparison

ABSTRACT

A highly compact and simple 2D graphical representation of DNA sequences, named DUC-Curve, is constructed through mapping four nucleotides to a unit circle with a cyclic order. DUC-Curve could directly detect nucleotide, di-nucleotide compositions and microsatellite structure from DNA sequences. Moreover, it also could be used for DNA sequence alignment. Taking geometric center vectors of DUC-Curves as sequence descriptor, we perform similarity analysis on the first exons of β-globin genes of 11 species, oncogene TP53 of 27 species and twenty-four Influenza A viruses, respectively. The obtained reasonable results illustrate that the proposed method is very effective in sequence comparison problems, and will at least play a complementary role in classification and clustering problems.

1. Introduction

Graphical techniques have been widely applied in visualizing functional relationships of complicated processes, and simplifying scientific notations. About 30 years ago this approach has been used as a novel tool in the field of biological sequence analysis, construction of phylogenetic trees and identification of protein coding regions. Nowadays it has extended from DNA, RNA to protein and proteomics, also developed from qualitative and pictorial representations to quantitative and numerical characterizations. Randić and cooperators introduced the term "graphical bioinformatics" to emphasize this research direction of bioinformatics [1].

Most early contributions of DNA graphical representations were lattice paths based on vectorial representations of four bases, as for example Refs. [2–5]. These models revealed considerable information about sequences, including distributions of bases, distances between particular bases, global sequence patterns and evolutionary divergences. However, these

* Corresponding author. Tel.: +86 021 64324284; fax: +86 021 64324284.
E-mail address: xqzheng@shnu.edu.cn (X. Zheng).

http://dx.doi.org/10.1016/j.physa.2016.03.061
0378-4371/

methods also suffered from the loss of information due to the overlap of the lattice path. In order to eliminate, or to minimize, the degeneracy of above representations, some researchers modified the original model by changing representation of bases, or the angles between the basis vectors [6–9]. Another earliest contribution of graphical representations was chaos game representation (CGR) based on chaotic dynamics theory [10]. This method produced a picture of a DNA sequence in both local and global patterns, and raised a set of questions about the structure of DNA sequences. Just inspired by CGR, Randić [11] and Stan et al. [12] dealt with similarity analysis for DNA sequences, and Pal et al. [13] examined multifractal behavior in power law cross correlation between any pair of nucleotide sequences of unequal lengths.

From the beginning of the twentieth century, graphical techniques have received important breakthroughs and a huge number of new forms have emerged. One class of typical achievements is spectral representations based on several horizontal lines [14–18]. These models avoided the self-overlapping of the graphs and possessed common advantages of graphical representations, but their spatial consumptions become very large with the long DNA sequence. Bielińska-Wąz [19] presented four-component spectral representation and overcame this drawback by introducing a resolution parameter. Also, zigzag curves with the compact form, as put forward by Refs. [20–24] just make up for the deficiency of the spectral representation. Besides the above three types of representations, there are many works which skillfully utilized the classification of nucleotides to the graphical techniques and effectively revealed physical and chemical properties of given sequences, as for example Refs. [15–17,24–26]. Recently, many new models adequately used dual nucleotides [27–29] or nucleotide triplet codons [30–33] to design graphical representations. At the same time, He [34] and Hou [35] combined Gray code and CMI coding respectively to develop new ways of graphical representation. Here, for more details about progresses of graphical representation, we highly recommend reviews [1,36–38].

Motivated by the UC-Curve, a highly compact 2D graphical representation of protein sequences proposed recently [39], we here introduce a simple and impact 2D graphical representation of DNA sequence, named DUC-Curve, through assigning four nucleotides to a unit circle with a cyclic order. The simple composition of DNA sequence makes DUC-Curve more advantageous than UC-Curve. Reasonable phylogenetic relationships for the first exons of β-globin genes of 11 species, oncogene TP53 of 27 species and twenty-four Influenza A viruses (see Tables 1–3) were established by geometric center vectors of DUC-Curves with low computational cost. Another advantage of DUC-Curve is its usage in sequence alignment, which seldom happens for available graphical models except for the 'four line' representation [14,40] and the graphical simple alignment tree (GSA tree) [41]. However, although the above two methods exhibited alignment information, such as locations and amounts of mismatches or gaps, they were not able to display concrete mismatch styles in two sequences. We take advantage of the feature of DUC-Curve and further modify it to obtain "extended DUC-Curve", which successfully solves the aforementioned problem.

2. DUC-Curve

2.1. Construction of DUC-Curve

Given a DNA sequence $G = g_1 g_2 \ldots g_n$, define a function φ that maps each nucleic base g_i in the sequence to one point $(x_i,\ y_i)$ over the circumference of a unit circle as follows

$$\varphi(g_i) = \begin{cases} \left(\dfrac{i}{n+1},\ \sqrt{1-\left(\dfrac{i}{n+1}\right)^2}\right) & \text{if } g_i = \text{A} \\ \left(\dfrac{i}{n+1},\ -\sqrt{1-\left(\dfrac{i}{n+1}\right)^2}\right) & \text{if } g_i = \text{C} \\ \left(-\dfrac{i}{n+1},\ \sqrt{1-\left(\dfrac{i}{n+1}\right)^2}\right) & \text{if } g_i = \text{G} \\ \left(-\dfrac{i}{n+1},\ -\sqrt{1-\left(\dfrac{i}{n+1}\right)^2}\right) & \text{if } g_i = \text{T} \end{cases} \qquad i = 1, 2, \ldots, n. \tag{1}$$

According to the definition of the function φ, all bases 'A' correspond to points distributing in the first quadrant, bases 'G' in the second quadrant, bases 'T' in the third quadrant and bases 'C' in the fourth quadrant. Next, connect the adjacent points in $\varphi(G)$ by lines and then obtain a simple graphical representation of the sequence. Since all points in this graphical representation of DNA sequence are arranged over the circumference of a Unit Circle, we call it the 'DUC-Curve'. Take the first exons of β-globin genes of *Bovine* and *Goat* for example.

Bovine:
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAG