ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 4, 2019, *pp*.300-306

Similarity/dissimilarity analysis of flowering plant DNA sequences

Xiaoshun Xu [1+] and Zhongrui Gao [1]

[1] *Department of Mathematics, Jinan University, Guangzhou, 510632, China*
(Received November 06 2019, accepted December 28 2019)

Abstract. The multiple sequence alignment (MSA) is a usual tool in DNA sequence comparison. However, this method meets a hard challenge for a large number of long DNA sequences. To remedy this problem, we propose a new way for DNA sequence comparison based on a novel DNA map. The method is that, by assigning a dinucleotide to a number, we construct a new graphical representation of DNA sequences based on horizontal lines. We further utilize the maximal eigenvalue of a related matrix to derive a mathematical descriptor for a DNA sequence. We also perform the similarity/dissimilarity analysis among the coding sequences of ribulose bisphosphate carboxylase small chain gene and large chain gene of flowering plants. The results indicate that this method can be reliable in the comparison of the flowering plant DNA sequences.

Keywords: Graphical representation, Dinucleotide, DNA map, Similarity/dissimilarity analysis, Phylogenetic tree, Flowering plants.

1. Introduction

There are so huge of biological data with growing rapidly, which catch more and more scientists' attention to analyze them. But the classical multiple sequence alignment (MSA) method has to face a so-called NP-hard problem for a large amount of data (Wang and Jiang 1994 [20]; Deng et al. 2011 [1]). For overcoming this barrier, many works focused on providing suitable alignment-free sequence comparison methods, for more details, please see (Jin et al. 2017 [6]; Zielezinski et al. 2017 [33]; Ren et al. 2018 [17]) and the references therein.

For example, for DNA sequence comparison, a basic procedure of alignment-free method is to use a suitable measure to compute the distance between feature vectors which are obtained from the representation of DNA sequences, so that the similarity/dissimilarity results can be derived. Accordingly, the choice of the measure and feature vector for DNA sequence is very important for this purpose. One way of the representation of DNA sequences is based on the graphs which were firstly introduced by Hamori and Ruskin (Hamori and Ruskin 1983 [5]; Hamori 1985 [4]), and then followed by (Gates 1985 [2]; Nandy 1994 [12]; Zhang R and Zhang 1994 [30]; Leong and Morgenthaler 1995 [7]; Yau et al. 2003 [22]; Yu JF et al. 2009 [29]; Zhang ZJ 2009 [31]; Tang et al. 2010 [19]; Yu CL et al. 2010 [28]; Yu CL, Deng, et al. 2011 [25]; Liao et al. 2013 [9]; Zhang ZJ et al. 2014 [32]; Zou et al. 2014 [34]; Li et al. 2016 [8]; Panas et al. 2018 [13]; Gong and Fan 2019 [3]). A parallel problem is to deal with the protein sequences, please refer (Yau et al. 2008 [23]; Wu et al. 2010 [21]; Randic et al. 2011 [16]; Yu CL, Cheng, et al. 2011 [24]) and the references therein.

Up to now, many mathematical tools have been applied in this topic. In particular, the works (Yu CL, Deng, et al. 2011 [25]; Liu 2018a [10], 2018b [11]) used the basic probabilistic quantities. Following their steps, we will provide a map from the space of DNA sequences to the 4-dimensional Euclidean space based on six horizontal lines. Once getting the feature vectors, we can study the similarity/dissimilarity of two plant genes respectively.

2. Methods

In this section, we first modify the graphical representation of DNA sequence which was introduced in (Liu 2018b [11]) based on joint probability, then define a new matrix, and get the feature vector of a sequence, so as to get the distance between two DNA sequences finally.

On account of the fact that A, T and C, G are two base pairs, Liu (Liu 2018a [10], 2018b [11]) assigned A, T and C, G to the same probability respectively, and got 2D graphical representations there. As discussed in (Liu 2018a [10]), each nucleotide is indicated by a number as follows.

[+] Corresponding author. *E-mail address: xiaoshunxu0808@stu2018.jnu.edu.cn.*

$$0.3 \to \text{A},\ -0.3 \to \text{T},$$
$$0.2 \to \text{C},\ -0.2 \to \text{G}.$$

Please note that A and T have the same absolute value which could be regarded as the probability, so do C and G.

From these setting, we can set a number to a dinucleotide as listed in Table 1 in a joint probability framework. For instance, the number of AC is $0.3 \times 0.2 = 0.06$, so do the others. There are 4x4=16 dinucleotides, and just six numbers corresponding to them. The details are as follows.

Dinucleotide	Number	Dinucleotide	Number
AA, TT	0.09	CC, GG	0.04
AC, CA	0.06	CT, TC	-0.06
AT, TA	-0.09	CG, GC	-0.04
AG, GA	-0.06	GT, TG	0.06

Table 1. Correspondence of numbers and dinucleotides

Now we want to present a 2D graphical representation of a DNA sequence. For example, given a DNA sequence, ATGCCTT, we read it as A, AT, TG, GC, CC, CT and TT. So its representation is as follows.

sequence	x-coordinate	y-coordinate
A	0	0.3
AT	1	-0.09
TG	2	0.06
GC	3	-0.04
CC	4	0.04
CT	5	-0.06
TT	6	0.09

Table 2. Representation of sequence ATGCCTT

Figure 1 provides the graph corresponding to this sequence.

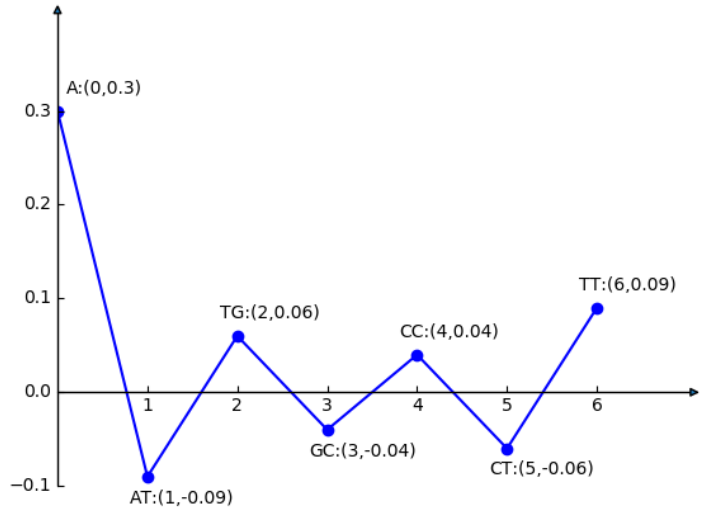

Figure 1. The graph corresponding to ATGCCTT

Actually, besides the first nucleotide, we set the product of the assigned numbers, which could be read as the joint probability up to a sign, to the corresponding dinucleotide, then put the point on the corresponding horizontal line, and finally connect these points to get the representation curve for this sequence.

As in (Liu 2018b [11]), we can omit the first point of the corresponding zigzag curve when the DNA sequences have the same first nucleotide. For a DNA sequence of length $n+1$, let (x_i, y_i) be the