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.

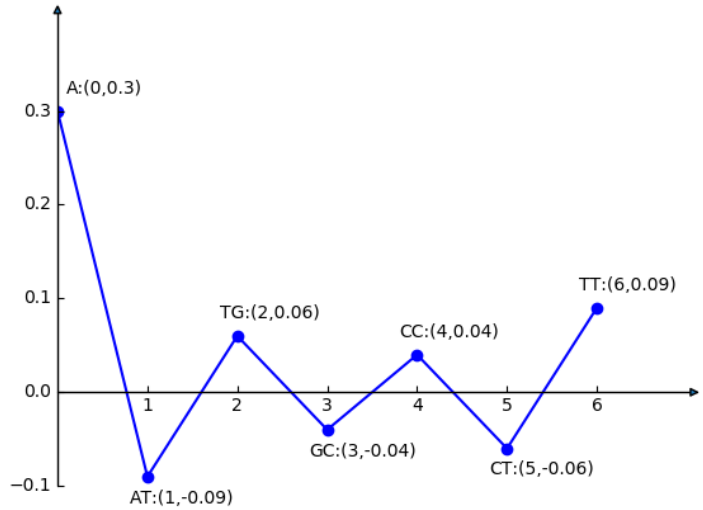

Figure 1. The graph corresponding to ATGCCTT

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

corresponding points on the zigzag curve, that is, x_i equals to i, and y_i is the number corresponding to the $(i+1)$-th nucleotide of the sequence based on the setting in Table 1. We can define a symmetric matrix E with elements E_{ij} given by

$$E_{ij} = \begin{cases} \dfrac{(y_i - y_j)^2}{|i-j|} & i \neq j \\ 0 & i = j \end{cases}.$$

So from linear algebra results, the matrix E is symmetric and then has real eigenvalues, so that there is the maximum of these eigenvalues. Hence, a DNA sequence derives a zigzag curve, which decides a number.

On the other hand, we could interchange the basic numbers corresponding to bases A and T to get another curve, so do C and G. Hence given a DNA sequence, we could get four curves, then four numbers λ_1, λ_2, λ_3 and λ_4. Hence by defining

$$\vec{V} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4),$$

we can obtain a four-dimensional vector for the DNA sequence. Roughly speaking, we define a novel DNA map from the space of DNA sequences to the four-dimensional Euclidean space. Please note that the target of the map here is not the space of representations which differs from that in (Randic 2004 [14]).

If given two sequences with two corresponding vectors $\vec{V}_1$ and $\vec{V}_2$ respectively, then the Euclidean distance d of these two vectors can be read as the dissimilarity between them, where

$$\mathrm{d} = \|\vec{V}_1 - \vec{V}_2\|.$$

Consequently, the smaller the distance d is, the more similar the two sequences are. On the contrary, the more dissimilar the two sequences are.

3. Results

In this section, we focus on analyzing two plant genes, ribulose bisphosphate carboxylase small chain gene and large chain gene, using the above measure d.

Let us first consider the coding sequences of ribulose bisphosphate carboxylase small chain gene (rbcS), whose information is listed in Table 5 in (Liu 2018b [11]). The gene comes from eleven flowering plants: Arachis duranensis (Dura),Arachis ipaensis (Ipae), Hevea brasiliensis (Rubber), Manihot esculenta (Cassava), Populus euphratica (Poplar), Nicotiana tabacum (Tobacco), Nicotiana tomentosiformis (Tome), Helianthus annuus (Sunflower), Brachypodium distachyon (Brome), Oryza brachyantha (Sina) and Musa acuminata (Banana). The last three species are from the Monocots, while the rest eight species belong to the Eudicots. The phylogenetic tree using our method reconstructed by UPGMA is given in Figure 2. From this figure, we can observe that the eleven species of plants fall into two groups successfully, and the coding DNA sequences of rbcS in Brome-Sina, Dura-Ipae, and Tobacco-Tome are close to each other.

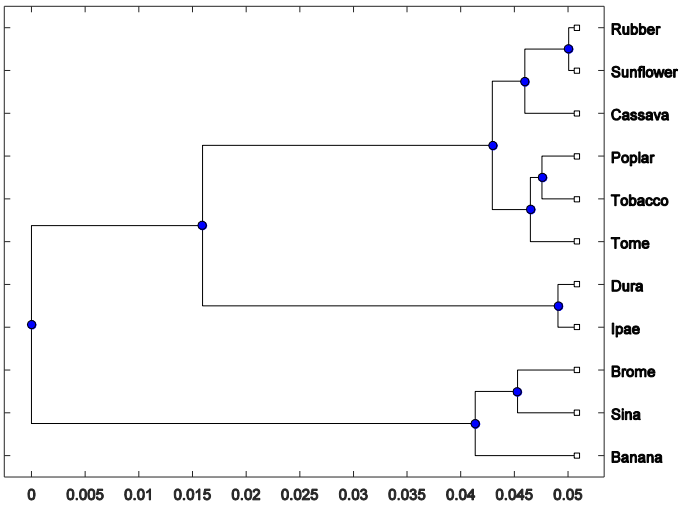

Figure 2. The phylogenetic tree of rbcS gene of 11 flowering plants constructed by our method with UPGMA.

DOI https://doi.org/2024-JICS-22406 | Generated on 2025-04-05 22:12:11

In order to further examine the effectiveness of our method, we study medium length sequences, the coding sequences of ribulose bisphosphate carboxylase large chain gene (rbcL) of eleven flowering plants, whose NCBI information is listed in Table 3. Note that Wheat, Rice, Onion and Duckweed belong to the Monocots, and others are from the Eudicots.

Accession number	Gene ID	Gene	Abbreviation	Name of species	Length(bp) of CDS
NC_002762.1	803091	rbcL	Wheat	Triticum aestivum	1434
NC_008155.1	4126887	rbcL	Rice	Oryza sativa Indica Group	1455
NC_024813.1	20355305	rbcL	Onion	Allium cepa	1440
NC_015891.1	11030084	rbcL	Duckweed	Spirodela polyrhiza	1461
NC_001879.2	800513	rbcL	Tobacco	Nicotiana tabacum	1434
NC_018117.1	13230328	rbcL	Jimsonweed	Datura stramonium	1434
NC_010361.1	5952063	rbcL	Primrose	Oenothera biennis	1428
NC_022404.1	17083101	rbcL	Tallowwood	Eucalyptus microcorys	1428
NC_002202.1	2715621	rbcL	Spinach	Spinacia oleracea	1428
NC_023357.1	18251513	rbcL	Corn cockle	Agrostemma githago	1428
NC_007977.1	4055709	rbcL	Sunflower	Helianthus annuus	1458

Table 3. NCBI information for rbcL gene of 11 flowering plants. CDS, coding DNA sequence.

The phylogenetic tree is shown in Figure 3. From the figure, we can discover that the 11 flowering plants are divided into two groups successfully, and the CDS of Spinach-Corn cockle, Tobacco-Jimsonweed, and Primrose-Tallowwood are close to each other.

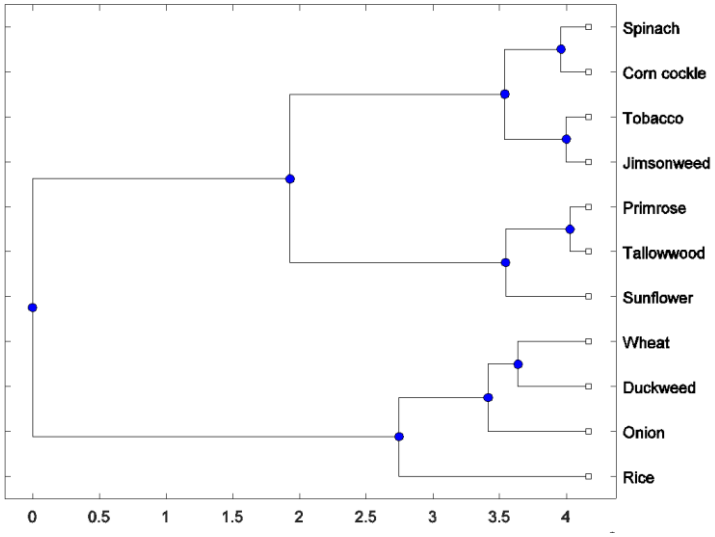

Figure 3. The phylogenetic tree of rbcL gene of 11 flowering plants constructed by our method with UPGMA.

In order to compare the method in (Liu 2018a), we apply their method to reconstruct the phylogenetic tree, which is shown in Figure 4. From this figure, we can see that this method fails to cluster this species into two classes. But except Jimsonweed, the rest ten species belong to their right classes.

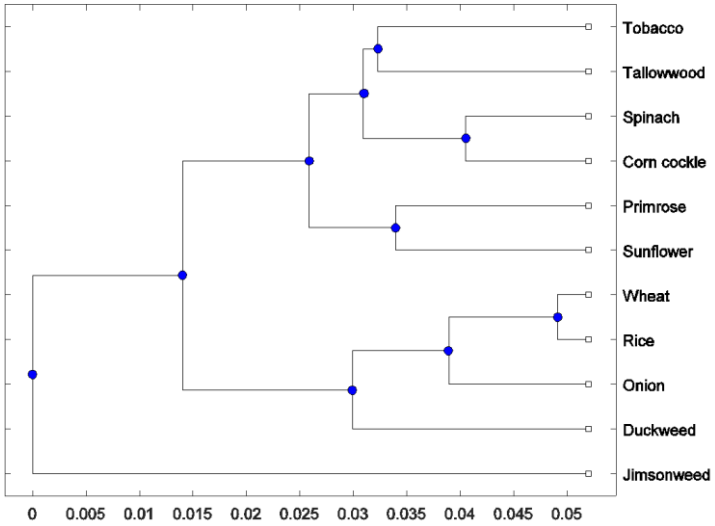

Figure 4. The phylogenetic tree reconstructed by the method in (Liu 2018a) with UPGMA.

4. Discussions

In (Randic et al. 2003 [15]), they used the 2D graphical representation of DNA sequence based on horizontal lines, and defined some related matrices, and then computed their eigenvalues, to analyze the coding sequence of the first exon of human beta-globin gene. Later in (Liu 2018a [10]), the probability view was involved. Our method combines their ideas and defines a new matrix E with elements E_{ij}. For $i \neq j$, the value of the element is defined by the ratio of the square of the difference of y-coordinates and the distance of x-coordinates. Noting that all y-coordinates are small, and then the square of their difference is smaller. Soas the distance of x-coordinates is very large, then their ratio is very small. To some extent, two nucleotides with longer distance have smaller impact with each other. Except above mentioned method, there are many other cluster methods, for instance (Yu CL, Deng, et al. 2014 [26]; Yu CL, He, et al. 2014 [27]; Siegel et al. 2015 [18]).

As its applications, we study the similarities among CDS of rbcS and rbcL. According to the CDS of rbcS, Figure 3 in (Liu 2018b [11]) shows that the maximum likelihood (ML) and the neighbor-joining (NJ) methods based on MSA could not divide these species into two correct classes (Eudicots and Monocots). In Figure 2, our method can do it correctly, which is also supported by the results of Figure 2 in (Liu 2018b [11]). Moreover the species Poplar is close to Tobacco, which is consistent with the results by ML and NJ based on MSA, while Figure2 in (Liu 2018b [11]) shows that it is close to Rubber. For the CDS of rbcL, Figure 3 shows that our method can put the species into right groups, but Figure 4 by the method in (Liu 2018a [10]) shows that Jimsonweed is an exception. Comparing these two figures, we can see that the most results for the other ten species support each other. All these imply the effectiveness of our proposed method.

In this work, we develop an alternative map from the space of DNA sequences to four-dimensional Euclidean space by combing the algebraic, geometrical and probabilistic views. The ingredients of our method may be applied to studying other biological sequences. But here we need to compute the maximal eigenvalue of the symmetric matrix E, which is not fast enough for a large amount of very long sequences. In future research, we try to develop our method to improve efficiency and analyze more biological data.

Acknowledgements

We express thanks to our tutor Xu-Qian Fan for his patient guidance. Thanks to the editor and anonymous reviewers for their comments and constructive reviews.

References

[1] Deng M, Yu C, Liang Q, He RL, Yau SS. 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. Plos One. 6(3):e17293.

[2] Gates MA. 1985. Simpler DNA sequence representations. Nature. 316(6025):219.

[3] Gong W, Fan X-Q. 2019. A geometric characterization of DNA sequence. Physica A: Statistical Mechanics and its Applications. 527.

[4] Hamori E. 1985. Novel DNA sequence representations. Nature. 314(6012):585-586.

[5] Hamori E, Ruskin J. 1983. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J Biol Chem. 258(2):1318-1327.

[6] Jin X, Jiang Q, Chen Y, Lee SJ, Nie R, Yao S, Zhou D, He K. 2017. Similarity/dissimilarity calculation methods of DNA sequences: A survey. J Mol Graph Model. 76:342-355.

[7] Leong PM, Morgenthaler S. 1995. Random walk and gap plots of DNA sequences. Computer Applications in the Biosciences Cabios. 11(5):503-507.

[8] Li YS, Liu Q, Zheng XQ. 2016. DUC-Curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment. Physica A. 456:256-270. English.

[9] Liao B, Xiang QL, Cai LJ, Cao Z. 2013. A new graphical coding of DNA sequence and its similarity calculation. Physica A. 392(19):4663-4667. English.

[10] Liu HL. 2018a. 2D Graphical Representation of DNA Sequence Based on Horizon Lines from a Probabilistic View. Biosci J. 34(3):1344-1350. English.

[11] Liu HL. 2018b. A Joint Probabilistic Model in DNA Sequences. Curr Bioinform. 13(3):234-240. English.

[12] Nandy A. 1994. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr Sci. 66:309-314.

[13] Panas D, Waz P, Bielinska-Waz D, Nandy A, Basak SC. 2018. An Application of the 2D-Dynamic Representation of DNA/RNA Sequences to the Prediction of Influenza A Virus Subtypes. MATCH Commun Math Comput Chem. 80(2):295-310. English.

[14] Randic M. 2004. Graphical representations of DNA as 2-D map. Chem Phys Lett. 386(4-6):468-471. English.

[15] Randic M, Vracko M, Lers N, Plavsic D. 2003. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 368(1-2):1-6. English.

[16] Randic M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D. 2011. Graphical representation of proteins. Chem Rev. 111(2):790-862. English.

[17] Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, Sun F. 2018. Alignment-Free Sequence Analysis and Applications. Annual Review of Biomedical Data Science. 1(1):93-114.

[18] Siegel K, Altenburger K, Hon Y-S, Lin J, Yu C. 2015. PuzzleCluster: A Novel Unsupervised Clustering Algorithm for Binning DNA Fragments in Metagenomics. Curr Bioinform. 10(2):225-231. English.

[19] Tang XC, Zhou PP, Qiu WY. 2010. On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chin Sci Bull. 55(8):701-704. English.

[20] Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. J Comput Biol. 1(4):337-348.

[21] Wu ZC, Xiao XA, Chou KC. 2010. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 267(1):29-34. English.

[22] Yau SST, Wang JS, Niknejad A, Lu C, Jin N, Ho YK. 2003. DNA sequence representation without degeneracy. Nucleic Acids Res. 31(12):3078-3080. English.

[23] Yau SST, Yu CL, He R. 2008. A protein map and its application. DNA Cell Biol. 27(5):241-250. English.

[24] Yu CL, Cheng SY, He RL, Yau SST. 2011. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. Gene. 486(1-2):110-118. English.

[25] Yu CL, Deng M, Yau SST. 2011. DNA sequence comparison by a novel probabilistic method. Inform Sciences. 181(8):1484-1492. English.

[26] Yu CL, Deng M, Zheng L, He RL, Yang J, Yau SST. 2014. DFA7, a new method to distinguish between intron-containing and intronless genes. Plos One. 9(7). English.

[27] Yu CL, He RL, Yau SST. 2014. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. J Theor Biol. 348:12-20. English.

[28] Yu CL, Liang QA, Yin CC, He RL, Yau SST. 2010. A novel construction of genome space with biological geometry. DNA Res. 17(3):155-168. English.

[29] Yu JF, Sun X, Wang JH. 2009. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J Theor Biol. 261(3):459-468. English.

[30] Zhang R, Zhang CT. 1994. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. Journal of

Biomolecular Structure & Dynamics. 11(4):767-782.

[31] Zhang ZJ. 2009. DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics. 25(9):1112-1117. English.

[32] Zhang ZJ, Li JY, Pan LQ, Ye YM, Zeng XX, Song T, Zhang XF, Wang EK. 2014. A novel visualization of DNA sequences, reflecting GC-content. MATCH Commun Math Comput Chem. 72(2):533-550. English.

[33] Zielezinski A, Vinga S, Almeida J, Karlowski WM. 2017. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 18. English.

[34] Zou S, Wang L, Wang J. 2014. A 2D graphical representation of the sequences of DNA based on triplets and its application. EURASIP J Bioinform Syst Biol. 2014(1):1.