Original Article

2D GRAPHICAL REPRESENTATION OF DNA SEQUENCE BASED ON HORIZON LINES FROM A PROBABILISTIC VIEW

REPRESENTAÇÃO GRÁFICA 2D DA SEQÜÊNCIA DE DNA BASEADA EM LINHAS DE HORIZONTE A PARTIR DE UMA VISÃO PROBABILÍSTICA

Huili LIU
Guangdong Provincial Key Laboratory of Protein Function and Regulation in Agricultural Organisms, College of Life Sciences, South China Agricultural University, Guangzhou, 510642, China
Email: liuhuili@scau.edu.cn

ABSTRACT: In this study, we propose a new two-dimensional graphical representation of DNA sequence based on a choice of four horizon lines. The 2D representation is constructed in a probabilistic framework. Following the new approach, we perform the similarity analysis among coding sequences of the first exon of beta-globin gene from eleven species. Our results coincide with current biological analyses. We also compare our method with some existing DNA sequence comparison algorithms and find that ours is more intuitive and effective.

KEYWORDS: Zigzag curve. Horizon lines. Numerical characterization. Similarity analysis.

INTRODUCTION

With the rapid development of sequencing technology, more and more DNA data has been acquired. It has currently been a big challenge for scientists to analyze the DNA sequences quickly and effectively. One important step in this topic is to graphically represent the DNA sequence, such that it keeps the information of primary data as more as possible. A large number of biologists, computer scientists and mathematicians applied solid computational tools to represent the biological sequences such as DNA, RNA and protein. Among of them, the graphical representation provides a simple and efficient visualization way, which has been used for numerically compare the biological sequences.

More than thirty years ago, Hamori and Ruskin (HAMORI; RUSKIN, 1983; HAMORI, 1985) introduced the first 3D graphical representation H-curve of DNA sequence. Up to now, many other multi-dimensional graphical representations of DNA sequence were followed (NANDY et al., 2006; JIN et al., 2017), including Z-curves (ZHANG; ZHANG, 1994), 4D graphical representations (CHI; DING, 2005; LIAO et al., 2005; TANG et al., 2010) and 6D model (LIAO; WANG, 2004). In particular, Gates (1985), Nandy (1994) and Leong and Morgenthaler (1995) map DNA sequence to a random walk in the (x, y) plane using four unit vectors to represent the four bases along corresponding axis directions respectively. However, those representations may have degeneracy and loss of information. In order to overcome these two problems and analyze genes, many other representations were studied. For example, 2D or 3D representations were discussed (GUO et al., 2001; RANDIC et al., 2003b; a; YAU et al., 2003; YAO et al., 2005; ZHANG et al., 2005; QI et al., 2007; QI; FAN, 2007; CAO et al., 2008; YU et al., 2009; ZHANG, 2009; CAO et al., 2010; YU et al., 2010; YU et al., 2010; XIE; MO, 2011; YU et al., 2011; HUANG; WANG, 2012; YU et al., 2013; YU et al., 2014; ZHANG et al., 2014; ZOU et al., 2014; HUANG; YU, 2016; LI et al., 2016; LI et al., 2016).

Specifically, Randic et al (2003a; b) introduced a 2D graphical representation of DNA sequence based on four horizontal lines and performed similarity analysis by a proposed L/L matrix. Furthermore, Yu et al. (2011) applied probabilistic methods with the help of graphical representation to compare the DNA sequences. Motivated by their works, we assign a nucleotide of DNA sequence as a point on one of the horizon lines. Then using this representation in a probabilistic framework as a new descriptor, the similarity/dissimilarity of the first exon of beta-globin gene of eleven species was studied.

MATERIAL AND METHODS

Yu et al (2011) proposed a 2D graphical representation of DNA sequence by defining a probability distribution of it, such that each nucleotide of the sequence has a number as an assigned probability and the sum of the probability equals to one. In their setting, the same nucleotide may have different probabilities, which depends on the location in the DNA sequence. In this section,

Received: 20/09/17
Accepted: 20/02/18

we will explain a 2D probabilistic representation based on horizon lines.

In contrast to representation in Yu et al.'s work, each nucleotide is indicated by a number as follows:
0.3 → A,
-0.3 → T,
0.2 → C,
-0.2 → G.

Here A and T are corresponding to the same number up to a sign for differing in the graph, so as C and G, since A-T and C-G are two base pairs. Similar to (RANDIC et al., 2003a; b), we have four horizon lines paralleling to *x*-axis with *y* values 0.3, -0.3, 0.2, -0.2 respectively and each nucleotide of a DNA sequence is mapped to a point on one of these lines such that a representation curve is derived by connecting these points one by one. For example, the representation of sequence TGCAC can be shown in Table 1, and the corresponding graphical representation of this sequence is drawn in Figure 1.

Table 1. Representation of sequence TGCAC.

sequence	*x*-coordinate	*y*-coordinate
T	1	-0.3
G	2	-0.2
C	3	0.2
A	4	0.3
C	5	0.2

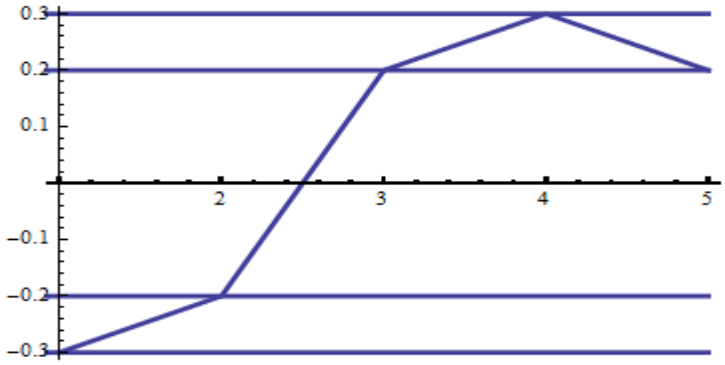

Figure 1. The graph corresponding to TGCAC.

This representation also has no loss of information and degeneracy, that is to say, the curve has no circuits and the mapping between the DNA sequences and the curves of graphical representation is one-to-one (YAU et al., 2003).

RESULTS

Based on the representation of DNA sequence, Randic *et al* (2003a; b) used matrices to provide numerical characterizations of DNA sequences which could be used to make similarity analysis of DNA sequences. They proposed an L//L matrix to characterize a DNA sequence with 12-component vectors, and obtained the similarity result by computing the Euclidean length of the difference of the vectors. This technical of mapping a zigzag curve to a vector is very effective for DNA similarity analysis. In this section, we characterize a sequence as a 4D vector. For two DNA sequences, we also compute the Euclidean length of difference of two 4D vectors derived from these two sequences, which reflects the similarity/dissimilarity between them.

For a DNA sequence of length *n*, we have a corresponding zigzag curve based on the assignment of bases as described in Table 1. Let (x_i, y_i) be the point corresponding to the *i*-th nucleotide of the sequence, then we can define the elements of matrix *E* evolving *y*-coordinate only, as follows:

$$E_{ij} = \begin{cases} \dfrac{y_i - y_j}{i - j} & \mathrm{i} \neq \mathrm{j} \\ 0 & \mathrm{i} = \mathrm{j} \end{cases}.$$

So the matrix *E* is symmetric, and has real eigenvalues such that the maximum of these eigenvalues exits.

For a DNA sequence, we have four difference graphs by interchanging assignment values corresponding to bases A and T, similar to bases C and G. These four curves are symmetry with respect to the x-axis. For each choice of the assignment for the basic four nucleotides, we can get a number by taking the maximal eigenvalue. So we get four

numbers d_1, d_2, d_3, d_4, then a 4D vector by assigning

$$\vec{P} = (d_1, d_2, d_3, d_4)$$

If we have two sequences with the corresponding descriptors $\vec{P}_1$ and $\vec{P}_2$ respectively, then similarity indexes between these two sequences can be defined by the Euclidean distance of these two vectors:

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

Thus the smaller *d* reflects that the DNA sequences are more similar.

Now we use this method to study the similarities among the coding sequences of the first exon of beta-globin genes of eleven species based on the data information listed in (JIN *et al.*, 2017) from GenBank. The result is shown in Table 2. In this table, we can see that (1) Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee are most similar. (2) Among the similarity of Human to other ten species, Human-Gorilla is the minimum, while Human-Gallus is the maximum. (3) Opossum, Lemur, Mouse, Rabbit, Rat, Gorilla, Bovine and Chimpanzee achieve the maximum at Human or Gallus.

Table 2. The similarity result (1.0e-2) for the coding sequences of the first exon of beta-globin gene of 11 species.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	7.338	6.235	14.558	5.487	5.009	10.118	9.364	0.020	6.456	0.417
Goat		0	11.111	9.103	9.104	4.709	6.870	7.127	7.335	3.463	7.316
Opossum			0	15.267	6.104	6.995	10.998	9.057	6.222	8.520	6.072
Gallus				0	13.158	9.666	5.660	8.383	14.547	9.176	14.420
Lemur					0	4.974	7.674	7.729	5.470	7.242	5.134
Mouse						0	5.324	5.318	4.996	2.978	4.824
Rabbit							0	5.176	10.104	5.959	9.892
Rat								0	9.348	4.177	9.079
Gorilla									0	6.446	0.400
Bovine										0	6.293
Chimpanzee											0

In order to compare our method with other ways, we list several highly cited results about the similarity of human with other ten species, which it is shown in Table 3. As in (PENG; LIU, 2015), we normalize the index by Human-Goat ratio such that the result can be compared easier. From Table 3, most results indicate that Human-Gorilla and Human-Chimpanzee are more similar than other 8 species, which is consistent with our result.

Table 3. The similarity indexes between human and other species. All indexes are normalized to Human-Goat ratio.

Methods	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Our work	1	0.85	1.98	0.75	0.68	1.38	1.28	0.00	0.88	0.06
Chi and Ding (2005)	1	3.71	0.82	2.73	0.69	0.50	0.48	0.07	3.59	0.58
Randic et al. (2003a)	1	2.43	1.79	1.43	1.37	0.69	0.70	0.34	1.38	0.28
Zhang (2009)	1	2.49	2.42	1.05	0.93	1.12	1.11	0.55	0.76	2.01
Yu et al. (2009)	1	1.14	1.12	1.07	0.78	0.77	0.86	0.27	0.78	0.41
Xie and Mo (2011)	1	8.20	14.54	6.65	18.87	4.76	13.94	0.54	0.93	0.08

DISCUSSION

Based on the facts that A-T and C-G are two base pairs, we choose four basic horizon lines where the points corresponding to DNA sequence lie in. Then we use a new matrix E to analyze the similarity/dissimilarity among eleven species by their coding sequences of the first exon of beta-globin gene. We use the same probability for a base even in different species, which is different from the methods used in (YU et al., 2011; ZHANG; CHEN, 2011). Roughly speaking, they regard a DNA sequence as an union of disjoint events such that the sum of the probability equals one, while we regard a DNA sequence as a random result, a sequence of probability. On the other hand, we focus the change of the probability of the sequence, and consider the matrix E just involving y-coordinates, in-depending x-coordinates. Thus the expression of E is much simpler than the L/L matrix in (RANDIC et al., 2003a; b), which shows that the cost of computing is reduced and our method is much faster.

In Table 3, one can see that, by our method, (1) the index of Human-Gallus is the maximum; (2) the indices of Human-Gorilla and Human-Chimpanzee are the two minimums. In fact, (2) is consistent with the results in (RANDIC et al., 2003a; YU et al., 2009; XIE; MO, 2011), but (1) is not true for other methods, such as in the results of (PENG; LIU, 2015). Among these 11 species, only Gallus is not mammalian, while Human, Gorilla and Chimpanzee belong to Primates, thus our result about Human-Gallus is more convincing. It implies the power of our new method.

In this study, we develop a new method to combine the geometrical and probabilistic information to analyze and compare the DNA sequences. Actually, some similar works with protein sequence have also been proposed (YAU et al., 2008; YU et al., 2011). Our approach can also extended to other biological sequences such as protein sequence. For example, we may consider 20 amino acids as 20 vectors instead of 4 nucleotides as 4 vectors here. Thus, further studies may be needed to decide what combination of 20 amino acid vectors to compare protein sequences. Furthermore, our method with sequence can be extended to study the two-dimensional structures of DNA or RNA as the fact that sequence determines structure and structure determines function. Our study provides an intuitive and efficient tool for DNA sequence comparison studies, which will be used to study more biological data in the near future.

ACKNOWLEDGEMENTS

This research was partially supported by the National Natural Science Foundation of China (31200218). The author would like to thank Sun Y. for valuable discussion.

RESUMO: Neste estudo, propomos uma nova representação gráfica bidimensional da sequência de DNA baseada na escolha de quatro linhas horizontais. A representação 2D é construída em uma estrutura probabilística. Seguindo a nova abordagem, realizamos a análise de similaridade entre as seqüências codificantes do primeiro exon do gene da beta-globina de onze espécies. Nossos resultados coincidem com as análises biológicas atuais. Também comparamos nosso método com alguns algoritmos de comparação de sequências de DNA existentes e descobrimos que o nosso é mais intuitivo e eficaz.

PALAVRAS-CHAVE: Curva em zigue-zague. Linhas de horizonte. Caracterização numérica. Análise de similaridade.

REFERENCES

CAO, Z.; LI, R. F.; CHEN, W. Y. A 3D Graphical Representation of DNA Sequence Based on Numerical Coding Method. **International Journal of Quantum Chemistry,** v. 110, n. 5, p. 975-980, 2010.

CAO, Z.; LIAO, B.; LI, R. F. A group of 3D graphical representation of DNA sequences based on dual nucleotides. **International Journal of Quantum Chemistry,** v. 108, n. 9, p. 1485-1490, 2008. https://doi.org/10.1002/qua.21698

CHI, R.; DING, K. Q. Novel 4D numerical representation of DNA sequences. **Chemical Physics Letters,** v. 407, n. 1-3, p. 63-67, 2005. https://doi.org/10.1016/j.cplett.2005.03.056

GATES, M. A. Simpler DNA sequence representations. **Nature,** v. 316, n. 6025, p. 219, 1985. https://doi.org/10.1038/316219a0

GUO, X. F.; RANDIC, M.; BASAK, S. C. A novel 2-D graphical representation of DNA sequences of low degeneracy. **Chemical Physics Letters,** v. 350, p. 106-112, 2001. https://doi.org/10.1016/S0009-2614(01)01246-5
HAMORI, E. Novel DNA sequence representations. **Nature,** v. 314, n. 6012, p. 585-6, 1985. https://doi.org/10.1038/314585a0
https://doi.org/10.1038/314585b0

HAMORI, E.; RUSKIN, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. **Journal of Biological Chemistry,** v. 258, n. 2, p. 1318-27, 1983.

HUANG, H. H.; YU, C. Clustering DNA sequences using the out-of-place measure with reduced n-grams. **Journal of Theoretical Biology,** v. 406, p. 61-72, 2016. https://doi.org/10.1016/j.jtbi.2016.06.029

HUANG, Y. J.; WANG, T. M. New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis. **International Journal of Quantum Chemistry,** v. 112, n. 6, p. 1746-1757, 2012. https://doi.org/10.1002/qua.23157

JIN, X.; JIANG, Q.; CHEN, Y.; LEE, S. J.; NIE, R.; YAO, S.; ZHOU, D.; HE, K. Similarity/dissimilarity calculation methods of DNA sequences: A survey. **Journal of Molecular Graphics and Modelling,** v. 76, p. 342-355, 2017. https://doi.org/10.1016/j.jmgm.2017.07.019

LEONG, P. M.; MORGENTHALER, S. Random walk and gap plots of DNA sequences. **Computer Applications in the Biosciences,** v. 11, n. 5, p. 503-7, 1995. https://doi.org/10.1093/bioinformatics/11.5.503

LI, C.; FEI, W. C.; ZHAO, Y.; YU, X. Q. Novel graphical representation and numerical characterization of DNA sequences. **Applied Sciences-Basel,** v. 6, n. 3, 2016.

LI, Y. S.; LIU, Q.; ZHENG, X. Q. DUC-Curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment. **Physica a-Statistical Mechanics and Its Applications,** v. 456, p. 256-270, 2016.

LIAO, B.; TAN, M. S.; DING, K. Q. A 4D representation of DNA sequences and its application. **Chemical Physics Letters,** v. 402, n. 4-6, p. 380-383, 2005. https://doi.org/10.1016/j.cplett.2004.12.062

LIAO, B.; WANG, T. M. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. **Journal of Chemical Information and Computer Sciences,** v. 44, n. 5, p. 1666-1670, 2004. https://doi.org/10.1021/ci034271f

NANDY, A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. **Current Science,** v. 66, p. 309-314, 1994.

NANDY, A.; HARLE, M.; BASAK, S. C. Mathematical descriptors of DNA sequences: development and applications. **Arkivoc,** v. 9, p. 211-238, 2006.

PENG, Y.; LIU, Y. W. An Improved Mathematical Object for Graphical Representation of DNA Sequences. **Current Bioinformatics,** v. 10, n. 3, p. 332-336, 2015. https://doi.org/10.2174/1574893610031507231355559

QI, X. Q.; WEN, J.; QI, Z. H. New 3D graphical representation of DNA sequence based on dual nucleotides. **Journal of Theoretical Biology,** v. 249, n. 4, p. 681-90, 2007. https://doi.org/10.1016/j.jtbi.2007.08.025

QI, Z. H.; FAN, T. R. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. **Chemical Physics Letters,** v. 442, n. 4-6, p. 434-440, 2007. https://doi.org/10.1016/j.cplett.2007.06.029

RANDIC, M.; VRACKO, M.; LERS, N.; PLAVSIC, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. **Chemical Physics Letters,** v. 371, n. 1-2, p. 202-207, 2003a. https://doi.org/10.1016/S0009-2614(03)00244-6

RANDIC, M.; VRACKO, M.; LERS, N.; PLAVSIC, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. **Chemical Physics Letters,** v. 368, n. 1-2, p. 1-6, 2003b. https://doi.org/10.1016/S0009-2614(02)01784-0

TANG, X. C.; ZHOU, P. P.; QIU, W. Y. On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. **Chinese Science Bulletin,** v. 55, n. 8, p. 701-704, 2010. https://doi.org/10.1007/s11434-010-0045-2

XIE, G. S.; MO, Z. X. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. **Journal of Theoretical Biology,** v. 269, n. 1, p. 123-130, 2011. https://doi.org/10.1016/j.jtbi.2010.10.018

YAO, Y. H.; NAN, X. Y.; WANG, T. M. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. **Chemical Physics Letters,** v. 411, n. 1-3, p. 248-255, 2005. https://doi.org/10.1016/j.cplett.2005.06.040

YAU, S. S. T.; WANG, J. S.; NIKNEJAD, A.; LU, C.; JIN, N.; HO, Y. K. DNA sequence representation without degeneracy. **Nucleic Acids Research,** v. 31, n. 12, p. 3078-3080, 2003. https://doi.org/10.1093/nar/gkg432

YAU, S. S. T.; YU, C.; HE, R. A protein map and its application. **DNA and Cell Biology,** v. 27, n. 5, p. 241-250, 2008. https://doi.org/10.1089/dna.2007.0676

YU, C.; HERNANDEZ, T.; ZHENG, H.; YAU, S. C.; HUANG, H. H.; HE, R. L.; YANG, J.; YAU, S. S. T. Real Time Classification of Viruses in 12 Dimensions. **Plos One,** v. 8, n. 5, 2013. https://doi.org/10.1371/journal.pone.0064328

YU, C.; CHENG, S. Y.; HE, R. L.; YAU, S. S. T. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. **Gene,** v. 486, n. 1-2, p. 110-118, 2011. https://doi.org/10.1016/j.gene.2011.07.002

YU, C.; DENG, M.; YAU, S. S. T. DNA sequence comparison by a novel probabilistic method. **Information Sciences,** v. 181, n. 8, p. 1484-1492, 2011. https://doi.org/10.1016/j.ins.2010.12.010

YU, C.; HE, R. L.; YAU, S. S. T. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. **Journal of Theoretical Biology,** v. 348, p. 12-20, 2014. https://doi.org/10.1016/j.jtbi.2014.01.022

YU, C.; LIANG, Q. A.; YIN, C. C.; HE, R. L.; YAU, S. S. T. A novel construction of genome space with biological geometry. **DNA Research,** v. 17, n. 3, p. 155-168, 2010. https://doi.org/10.1093/dnares/dsq008

YU, J. F.; SUN, X.; WANG, J. H. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. **Journal of Theoretical Biology,** v. 261, n. 3, p. 459-468, 2009. https://doi.org/10.1016/j.jtbi.2009.08.005

YU, J. F.; WANG, J. H.; SUN, X. Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. **MATCH-Communications in Mathematical and in Computer Chemistry,** v. 63, n. 2, p. 493-512, 2010.

ZHANG, R.; ZHANG, C. T. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. **Journal of Biomolecular Structure and Dynamics,** v. 11, n. 4, p. 767-82, 1994. https://doi.org/10.1080/07391102.1994.10508031

ZHANG, Y. S.; CHEN, W. A new measure for similarity searching in DNA sequences. **MATCH-Communications in Mathematical and in Computer Chemistry,** v. 65, n. 2, p. 477-488, 2011.

ZHANG, Y. S.; LIAO, B.; DING, K. Q. On 2D graphical representation of DNA sequence of nondegeneracy. **Chemical Physics Letters,** v. 411, n. 1-3, p. 28-32, 2005. https://doi.org/10.1016/j.cplett.2005.06.005

ZHANG, Z. J. DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. **Bioinformatics,** v. 25, n. 9, p. 1112-1117, 2009. https://doi.org/10.1093/bioinformatics/btp130

ZHANG, Z. J.; LI, J. Y.; PAN, L. Q.; YE, Y. M.; ZENG, X. X.; SONG, T.; ZHANG, X. F.; WANG, E. K. A novel visualization of DNA sequences, reflecting GC-content. **MATCH-Communications in Mathematical and in Computer Chemistry,** v. 72, n. 2, p. 533-550, 2014.

ZOU, S.; WANG, L.; WANG, J. A 2D graphical representation of the sequences of DNA based on triplets and its application. **Eurasip Journal on Bioinformatics & Systems Biology,** v. 2014, n. 1, p. 1, 2014. https://doi.org/10.1186/1687-4153-2014-1