SCIENTIFIC REPORTS

OPEN

One novel representation of DNA sequence based on the global and local position information

Received: 11 January 2018
Accepted: 27 April 2018
Published online: 15 May 2018

Zhiyi Mo[1], Wen Zhu[2], Yi Sun[2], Qilin Xiang[2], Ming Zheng[1], Min Chen[3] & Zejun Li[3]

One novel representation of DNA sequence combining the global and local position information of the original sequence has been proposed to distinguish the different species. First, for the sufficient exploitation of global information, one graphical representation of DNA sequence has been formulated according to the curve of Fermat spiral. Then, for the consideration of local characteristics of DNA sequence, attaching each point in the curve of Fermat spiral with the related mass has been applied based on the relationships of neighboring four nucleotides. In this paper, the normalized moments of inertia of the curve of Fermat spiral which composed by the points with mass has been calculated as the numerical description of the corresponding DNA sequence on the first exons of beta-global genes. Choosing the Euclidean distance as the measurement of the numerical descriptions, the similarity between species has shown the performance of proposed method.

The graphical and numerical representation of DNA, RNA or protein sequences has become the popular strategies to analyze the evolutionary relationship between species. As the availability of varies gene data for different species, the comparison of different organisms that own unique genetic information involves in mathematics, biology, physics, informatics and so on. Many researchers have focused on the issue of representation of gene sequence, as seen in[1–31], so the study of representation of gene sequence is significant and beneficial.

Hamori and Ruskin[32] first proposed the H-curve, the graphical representation of nucleotide sequence, which is convenient for the visual analysis and comprehension of the DNA sequences. Following them, further researches of representation of DNA sequence were carried[33–44]. For example, Zhang[45] proposed a five-color map visualization of DNA sequences named ColorSquare. Jafarzadeh[1] constructed the C-curve with no loss of information. And Aram[5] introduced a new graphical representation of the DNA sequences which called spider representation. Moreover, Bielinska-Waz[10] represented the sequence with a set of discrete lines which referred to as the B−spectrum. Unfortunately, owing to the high degeneracy and loss of information and the need of a lot of space in the transformation of DNA sequence to graphical representation, the performances of many methods are not satisfactory as expected.

To solve those problems, we present one novel representation of DNA sequence based on global and local position information. Distinct from previous reports, the more effective representation is obtained and the possible effect caused by different length of DNA sequence is restrained by new method. In detail, the novel concept of representation of DNA sequence involves (1) formulating the graphical representation of DNA sequence according to the curve of Fermat spiral which remaining the global position information of the original sequence, (2) taking the local position information of DNA sequence into consideration according to attach each point in the curve of Fermat spiral with the related mass, (3) the normalized moments of inertia of the curve of Fermat spiral which composed by the points with mass has been calculated as the description of the corresponding DNA sequence on the first exons of beta-global genes.

Graphical representation of DNA sequence

In order to make full use of global information of DNA sequence, the original DNA sequence is divided into four subsequences constituted by A, C, G or T that four point sets correspondingly can be obtained by the position of nucleotide in the original DNA sequence. Thus, each nucleotide in the subsequence corresponds to one point in the set. With the operation by distributing each point set to the curve of Fermat spiral, four corresponding

[1]School of Information and Electronic Engineering, Wuzhou University, Wuzhu, China. [2]College of Computer Science and Electronic Engineering, Hunan University, Hunan, China. [3]College of Computer and Information Science, Hunan Institute of Technology, Hengyang, China. Correspondence and requests for materials should be addressed to W.Z. (email: syzhuwen@163.com)

curves which means the graphical representation of DNA sequence can be plotted. The reason that we choose the Fermat spiral instead of the circle as the distribution curve of subsequence is that the curve of Fermat spiral is the monotonically increasing functions in the polar coordinate system which can remaining the information of position of the original sequence.

We regard the DNA sequence as BS(base sequence) which is constituted by four subsequences of AS, CS, GS and TS. Concretely, the i-th nucleotide in BS is denoted as V_i^{BS}, i = 1, 2, ···, N_{BS}. It is obvious that the length of nucleotide in base sequence is equal to the total length of nucleotide in four subsequence, described as:

$$N_{BS} = N_{AS} + N_{CS} + N_{GS} + N_{TS} \tag{1}$$

where N_{BS}, N_{AS}, N_{CS}, N_{GS} and N_{TS} respectively denote the length of nucleotide in base, A, C, G and T subsequence. For the purpose of plotting the base curve of Fermat spiral corresponding to the base sequence, the coordinate of points in the polar coordinate system are calculated according to the information of position in the base sequence. For each point, calculated as:

$$\theta_{V_i^{BS}} = \frac{2\pi}{(L-1)} \times \left(L_{V_i^{BS}} - 1\right) \tag{2}$$

where $\theta_{V_i^{BS}}$ denotes the polar angle of nucleotide V_i^{BS} in the polar coordinate system; L is one constant which means the shortest length of DNA sequence for different species in the experience; $L_{V_i^{BS}}$ denotes the position of nucleotide V_i^{BS} in the base sequence which ranging from 1 to N_{BS}. The mathematical formula of the curve of Fermat spiral is described as:

$$\rho_{V_i^{BS}} = \sqrt{\theta_{V_i^{BS}}} \tag{3}$$

As for the nucleotides in the base sequence, the corresponding set of coordinate for each point in the polar coordinates are calculated as

$$\left\{p_{V_1^{BS}}\left(\theta_{V_1^{BS}}, \rho_{V_1^{BS}}\right), p_{V_2^{BS}}\left(\theta_{V_2^{BS}}, \rho_{V_2^{BS}}\right), \ \ldots, \ p_{V_i^{BS}}\left(\theta_{V_i^{BS}}, \rho_{V_i^{BS}}\right), \ \ldots, \ p_{V_{N_{BS}}^{BS}}\left(\theta_{V_{N_{BS}}^{BS}}, \rho_{V_{N_{BS}}^{BS}}\right)\right\}.$$

Correspondingly, four subsets can be obtained and plotted. As shown in Fig. 1, the graphical representation of the first exons of β-globin gene of human DNA gene is plotted.

Attaching each point with a mass

In order to make full use of carried information of DNA sequence, the local characteristics are taken into consideration to attach each point corresponding to the nucleotide in the base sequence with a mass. Since one of immediate 5′ neighbor nucleotide and two of immediate 3′ neighbor nucleotides were considered as the context to calculate the mass of point corresponding to the second nucleotide in the group, the times and the compactness that the second nucleotide occurs and arranges are considered as the criterion to confirm the mass of the second nucleotide in the group.

According to the times that the nucleotide same as the second position repeats in the group, four categories may be divided. As shown in the following the nucleotide being same as the second nucleotide is denoted as ***1*** and the nucleotide being different from the second nucleotide is denoted as ***0***.

(1) ***0100***
(2) ***0101, 1100, 0110***
(3) ***1101, 1110, 0111***
(4) ***1111***

For example, the nucleotide of second position occurs one time in the first category. And according to the analysis of the four categories, six situations are obtained by the compactness that the second nucleotide arranges. For example, as for the second category that the nucleotide of second position occurs two times, its first situation of ***0101*** in which the two ***1*** are separated by on***e 0*** and its second situation in which the two ***1*** are compactly arranged.

(1) ***0100***
(2) ***0101***
(3) ***1100, 0110***
(4) ***1101***
(5) ***1110, 0111***
(6) ***1111***

Therefore, the different mass in $\left\{\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\right\}$ is attached to the point corresponding to the nucleotide of second position. However, for the purpose of reducing the impact of DNA sequence which is too long, the mass of latter sequence after ***L*** are restrained as

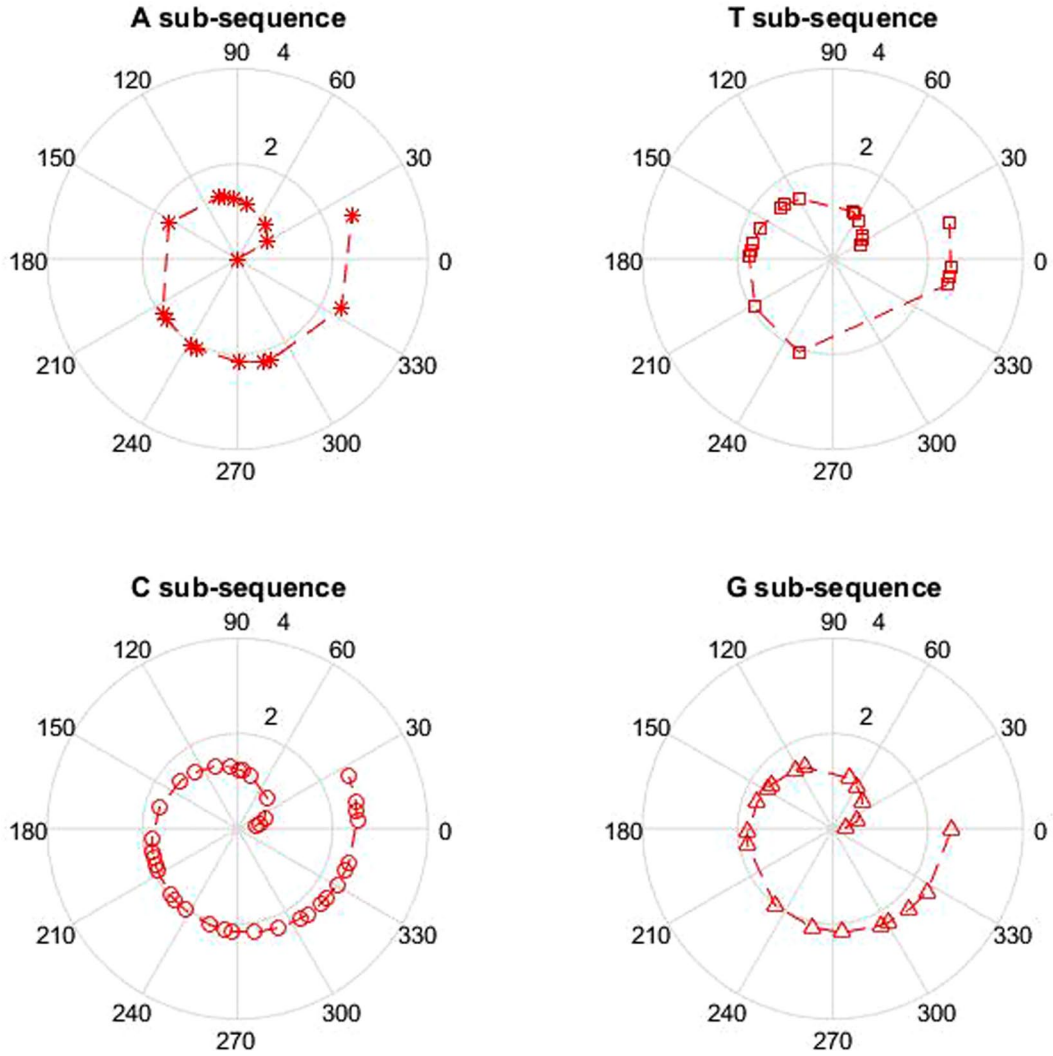

Figure 1. The graphical representation of human gene. From left to right and from top to bottom, the graphical representations are respectively for A, C, G and T subsequence.

$$\begin{cases} \widetilde{m_{V_i^{BS}}} = m_{V_i^{BS}}, & if\ L_{V_i^{BS}} \le L \\ \widetilde{m_{V_i^{BS}}} = \varepsilon \times m_{V_i^{BS}}, & if\ L_{V_i^{BS}} > L \end{cases} \tag{4}$$

where $\widetilde{m_{V_i^{BS}}}$ denotes the mass of the point corresponding to the nucleotide of V_i^{BS} after restraint; ε denotes the scale of constraint which is one constant of ***0.0375*** in experiment. So the points corresponding to the positions later than L own bigger polar radius but smaller mass; on the one hand, this characteristic can restrain the difference on length of dissimilar species; on other, it also can reserve the smaller difference on length of similar species.

Numerical Representation

For the widespread application of the moment of inertia in many gene numerical representation method[10,11,15,16], the normalized moments of inertia for each massive sub-curve of Fermat spiral are calculated as the numerical representation of formal DNA sequence in this paper. To the convenience of calculation, the transformation of polar coordinates to plane coordinates is performed:

$$\begin{cases} x_{V_i^{BS}} = \rho_{V_i^{BS}} \times cos\ \theta_{V_i^{BS}} \\ y_{V_i^{BS}} = \rho_{V_i^{BS}} \times sin\ \theta_{V_i^{BS}} \end{cases} \tag{5}$$

k	Species	Gene ID	N
1	Human	U01317	92
2	Gorilla	X61109	93
3	Chimpanzee	X02345	105
4	Rat	X06701	92
5	Mouse	V00722	93
6	Lemur	M15734	92
7	Rabbit	V00882	92
8	Goat	M15387	86
9	Bovine	X00376	86
10	Opossum	J03643	92
11	Gallus	V00409	92

Table 1. The first exons of β-globin gene of different species.

Species	r_{AS}	r_{CS}	r_{GS}	r_{TS}
Human	1.6674	1.7921	1.7233	1.7689
Gorilla	1.6674	1.7921	1.7233	1.7727
Chimpanzee	1.6885	1.8085	1.7239	1.7845
Rat	1.6074	1.7858	1.8242	1.7462
Mouse	1.5943	1.7480	1.8407	1.7921
Lemur	1.6781	1.6659	1.8149	1.8046
Rabbit	1.6454	1.7145	1.8690	1.7809
Goat	1.5416	1.6808	1.8246	1.8476
Bovine	1.5416	1.5929	1.8056	1.8471
Opossum	1.5693	1.6713	1.9416	1.7398
Gallus	1.7986	1.6879	1.9639	1.7140

Table 2. The numerical representation of DNA sequence.

Species	Human	Gorilla	Chimp	Rat	Mouse	Lemur	Rabbit	Goat	Bovine	Opossum	Gallus
Human	0	0.0038	0.0309	0.1198	0.1470	0.1604	0.1670	0.2114	0.2616	0.2696	0.2983
Gorilla	0.0038	0	0.0292	0.1206	0.1465	0.1596	0.1668	0.2100	0.2604	0.2701	0.2991
Chimp	0.0309	0.0292	0	0.1365	0.1620	0.1707	0.1782	0.2281	0.2804	0.2870	0.2988
Rat	0.1198	0.1206	0.1365	0	0.0631	0.1513	0.0987	0.1602	0.2282	0.1684	0.2583
Mouse	0.1470	0.1465	0.1620	0.0631	0	0.1208	0.0683	0.1031	0.1763	0.1393	0.2582
Lemur	0.1604	0.1596	0.1707	0.1513	0.1208	0	0.0832	0.1443	0.1608	0.1792	0.2131
Rabbit	0.1670	0.1668	0.1782	0.0987	0.0683	0.0832	0	0.1355	0.1844	0.1209	0.1940
Goat	0.2114	0.2100	0.2281	0.1602	0.1031	0.1443	0.1355	0	0.0900	0.1618	0.3215
Bovine	0.2616	0.2604	0.2804	0.2282	0.1763	0.1608	0.1844	0.0900	0	0.1921	0.3433
Opossum	0.2696	0.2701	0.2870	0.1684	0.1393	0.1792	0.1209	0.1618	0.1921	0	0.2324
Gallus	0.2983	0.2991	0.2988	0.2583	0.2582	0.2131	0.1940	0.3215	0.3433	0.2324	0

Table 3. Similarity/dissimilarity matrix under the Euclidean distance.

Since the point $p_{V_i^{BS}}(\theta_{V_i^{BS}}, \rho_{V_i^{BS}})$ in the polar coordinates is transformed to point $p_{V_i^{BS}}(x_{V_i^{BS}}, y_{V_i^{BS}})$ in the plane coordinates, the center of mass for the massive curve of Fermat spiral in the plane coordinates system is calculated as:

$$\begin{cases} \widetilde{x_{V^{BS}}} = \dfrac{1}{N_{BS}} \sum_{i=1}^{N_{BS}} m_{V_i^{BS}} \times x_{V_i^{BS}} \\ \widetilde{y_{V^{BS}}} = \dfrac{1}{N_{BS}} \sum_{i=1}^{N_{BS}} m_{V_i^{BS}} \times y_{V_i^{BS}} \end{cases} \tag{6}$$

So the ordinate of the center of mass is point $\widetilde{p_{V^{BS}}}(\widetilde{x_{V^{BS}}}, \widetilde{y_{V^{BS}}})$, the moment of inertia of the massive curve is described as:

Methods	Gorilla	Chimp	Rat	Mouse	Lemur	Rabbit	Goat	Bovine	Opossum	Gallus
Our work	0.0038	0.0309	0.1198	0.1470	0.1604	0.1670	0.2114	0.2616	0.2696	0.2983
Randic *et al.* 2003[33]	0.0210	0.0170	0.0430	0.0830	0.0870	0.0420	0.0610	0.0840	0.1480	0.1090
Dai *et al.* 2006[46]	0.0120	0.0155	0.0704	0.0543	0.0603	0.0287	0.0169	0.0276	0.1389	0.1146
Liu and Wang 2006[47]	0.3070	0.3101	0.4256	0.3089	0.3688	0.2968	0.4341	0.4172	0.3805	0.4479
Liao *et al.* 2013[2]	0.1651	0.4688	0.9202	0.6024	1.0110	0.7453	0.6010	0.6320	1.3710	1.5932
Jafarzadeh *et al.* 2013[1]	0.0330	0.0920	0.2160	0.1630	0.1940	0.1240	0.1650	0.2210	0.1940	0.1940
Bielinska-Waz *et al.* 2017[10]	0.0056	0.0314	0.1838	0.2395	0.2497	0.1844	0.1276	0.0872	0.3904	0.4687

Table 4. Similarity/dissimilarity between Human and other species with different methods.

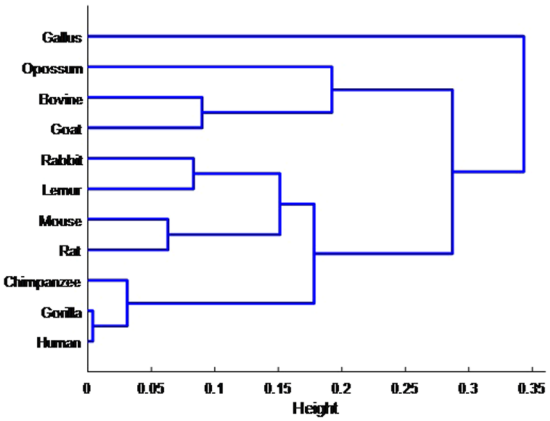

Figure 2. Cluster dendrogram.

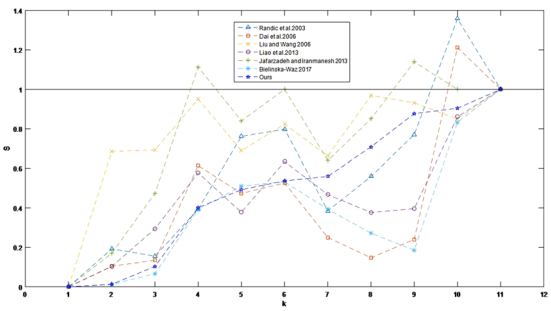

Figure 3. Similarity values of human-other species with different methods.

$$M_{BS} = \sum_{i=1}^{N_{BS}} m_{V_i^{BS}} \times distance\left(p_{V_i^{BS}}, \widetilde{p_{V^{BS}}}\right) \tag{7}$$

where distance $\left(p_{V_i^{BS}}, \widetilde{p_{V^{BS}}}\right)$ denotes the squared distance, calculated as:

$$distance\left(p_{V_i^{BS}}, \widetilde{p_{V^{BS}}}\right) = \left(x_{V_i^{BS}} - \widetilde{x_{V_i^{BS}}}\right)^2 + \left(y_{V_i^{BS}} - \widetilde{y_{V_i^{BS}}}\right)^2 \tag{8}$$

The normalized moment of inertia is described as:

$$r_{BS} = \sqrt{\frac{M_{BS}}{\sum_{i=1}^{N_{BS}} m_{V_i^{BS}}}} \tag{9}$$

There r_{BS}, one 4-dimensional vector $r_{BS} = [r_{AS}, r_{CS}, r_{GS}, r_{TS}]$, denotes the numerical representation of DNA sequence consisted of A, T, C and G subsequences. Following, the similarity distance between species is calculated according to the Euclidean measurement:

$$S(\alpha, \beta) = \left[\sum \left|r_{BS}^{\alpha} - r_{BS}^{\beta}\right|^2\right]^{\frac{1}{2}} \quad (10)$$

where r_{BS}^{α} and r_{BS}^{β} respectively denote the numerical representations of species α and β. So $S(\alpha, \beta)$ denotes the similarity distance between vectors r_{BS}^{α} and r_{BS}^{β} in the 4-dimensional space.

Results and Discussion

We test the performance of proposed method in the standard dataset that popular in the field of the DNA representation research, as seen in Table 1, the first exons of β-globin gene of different species. According to Eq. (9), Table 2 shows the numerical representations of DNA sequence for each target species. After obtaining the numerical representation consisted of 4-dimensional vectors, Table 3 shows the similarity/dissimilarity between pairs of species according the description of Eq. (10).

For the comparison, Table 4 shows the similarity/dissimilarity between Human and other species in some other methods similarly taking the Euclidean distance as the measurement. From Table 4, finding that most listed methods[1,2,10,46,47] also make the same conclusion that Gorilla are the most similar species to Human and Chimp is the next similar species to Human except method[33] which make the similar conclusion that Chimp is the most similar species to Human and Gorilla is the next similar species to Human. Besides, some listed methods[2,10,47] also make the same conclusion that Gallus is the most dissimilar species to Human.

Two most distinct dendrogram corresponding to the Euclidean measures is plotted in Fig. 2. As seen, the similar cluster pairs are respectively as Human-Gorilla(same cluster result in[1,2,10,40,47,48]), Rat-Mouse(same result in[10,49]), Lemur-Rabbit(same cluster result in[10]), Goat-Bovine(same cluster result in[10,33,40,47–49]), Human-Gorilla-Chimpanzee (same cluster result in[1,10,33,40,46,48,49]).

Normalizing $S^{human-gallus} = 1$ to the convenience of the visualization for results in other paper[1,2,10,33,46,47] which similarly using the Euclidean measurement. As shown in Fig. 3, different methods perform different results that may be useful with different consideration.

In conclusion, the paper presents a novel method to extract the characteristic of the DNA sequence with the graphical and numerical operations which can effectively achieve the similarity/dissimilarity comparison of different species. In this method, the distribution of sequence to the curve of Fermat spiral remains the global position information successfully and the attachment of the mass to the point remains the local position information successfully. Specifically in our result, the group of Rat-Mouse- Lemur- Rabbit is more similar to the group of Human- Gorilla-Chimpanzee compared with the group of Goat- Bovine-Opossum which may be helpful to the exploration of the evolutionary relationship between species. Moreover, the similar pairs that obtained by our method illustrate the performance of proposed representation of DNA sequence.

References

1. Jafarzadeh, N. & Iranmanesh, A. C-curve: a novel 3D graphical representation of DNA sequence based on codons. *Math Biosci.* **241**, 217–224 (2013).
2. Liao, B., Xiang, Q., Cai, L. & Cao, Z. A new graphical coding of DNA sequence and its similarity calculation. *Physica A.* **392**, 4663–4667 (2013).
3. Yang, X. & Wang, T. Linear regression model of short k-word: A similarity distance suitable for biological sequences with various lengths. *J Theor Biol.* **337**, 61–70 (2013).
4. Wąż, P. & Bielińskawąż, D. Non-standard similarity/dissimilarity analysis of DNA sequences. *Genomics.* **104**, 464–471 (2014).
5. Aram, V., Iranmanesh, A. & Majid, Z. Spider representation of DNA sequences. *J Comput Theor Nanos.* **11**, 418–420 (2014).
6. Liu, Y. W. & Peng, Y. A novel technique for analyzing the similarity and dissimilarity of DNA sequences. *Genet Mol Res.* **13**, 570–577 (2014).
7. Yin, C., Yin, X. E. & Wang, J. A novel method for comparative analysis of DNA sequences by Ramanujan-Fourier transform. *J Comput Biol.* **21**, 867–879 (2014).
8. Li, C., Fei, W. C., Zhao, Y. & Yu, X. Q. Novel Graphical Representation and Numerical Characterization of DNA Sequences. *Applied Sciences.* **6**, 63 (2016).
9. Xu, X. & Zhu, F. A New Method to Digitize DNA Sequence. *J Biosci Med.* **05**, 7–12 (2017).
10. Bielińskawąż, D. & Wąż, P. Spectral-dynamic representation of DNA sequences. *J Biomed Inform.* **72**, 1–7 (2017).
11. Panas, D., Wąż, P., Bielińskawąż, D., Nandy, A. & Basak, S. C. 2D-Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome. *MATCH Commun. Math Comput Chem.* **77**, 321–332 (2017).
12. Ma, T., Liu, Y., Dai, Q., Yao, Y. & He, P. A. A graphical representation of protein based on a novel iterated function system. *Physica A.* **403**, 21–28 (2014).
13. Li, Y., Liu, Q., Zheng, X. & He, P. A. UC-Curve: A highly compact 2D graphical representation of protein sequences. *Int. J Quantum Chem.* **114**, 409–415 (2014).
14. Yao, Y., Yan, S., Han, J., Dai, Q. & He, P. A. A novel descriptor of protein sequences and its application. *J Theor Biol.* **347**, 109–117 (2014).
15. Yao, Y. *et al.* Similarity/Dissimilarity Analysis of Protein Sequences Based on a New Spectrum-Like Graphical Representation. *Evol Bioinform Online.* **10**, 87–96 (2014).
16. Xu, S. C., Li, Z., Zhang, S. P. & Hu, J. L. Primary structure similarity analysis of proteins sequences by a new graphical representation. *SAR QSAR Environ Res.* **25**, 791–803 (2014).
17. El-Lakkani, A. & Mahran, H. An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation. *SAR QSAR Environ. Res.* **26**, 125–137 (2015).
18. Hou, W., Pan, Q. & He, M. A new graphical representation of protein sequences and its applications. *Physica A.* **444**, 996–1002 (2016).
19. Czerniecka, A., Bielińskawąż, D., Wąż, P. & Clark, T. 20D-dynamic Representation of Protein Sequences. *Genomics.* **107**, 16–23 (2016).
20. Ping, P., Zhu, X. & Wang, L. Similarities/dissimilarities analysis of protein sequences based on pca-fft. *J Biol Syst.* **25**, 1–17 (2017).
21. Hu, H., Li, Z., Dong, H. & Zhou, T. Graphical Representation and Similarity Analysis of Protein Sequences Based on Fractal Interpolation. *IEEE ACM T Comput Bi.* **14**, 182–192 (2017).
22. Liao, B., Liao, L., Wu, R. & Li, R. Construction of the phylogenetic tree by self-organizing map based on encoding sequence. *J Comput Theor Nanos.* **9**, 826–830 (2012).
23. Liao, B., Liao, B. Y., Lu, X. & Cao, Z. A Novel Graphical Representation of Protein Sequences and Its Application. *J Comput Chem.* **32**, 2539–2544 (2011).

24. Liao, B., Liao, B., Sun, X. & Zeng, Q. A Novel method for similarity analysis and protein subcellular localization prediction. *Bioinformatics.* **26**, 2678–2683 (2010).
25. Li, X., Liao, B., Zeng, Q. & Luo, J. Protein functional class prediction using global encoding of amino acid sequence. *J Theor Biol.* **261**, 290–293 (2009).
26. Huang, G., Liao, B. & Li, R. Similarity studies of DNA sequences based on a new 2D graphical representation. *Biophys Chem.* **143**, 55–59 (2009).
27. Liao, B., Zeng, C., Li, F. & Tang, Y. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Dual Nucleotides. *MATCH Commun Math Co.* **59**, 647–652 (2008).
28. Yao, Y., Kong, F., Dai, Q. & He, P. A Sequence-Segmented Method Applied to the Similarity Analysis of Long Protein Sequence. *MATCH Commun Math Co.* **70**, 431–450 (2013).
29. He, P., Xu, S., Dai, Q. & Yao, Y. A generalization of CGR representation for analyzing and comparing protein sequences. *Int J Quantum Chem.* **116**, 476–482 (2016).
30. Dai, Q. *et al*. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. *BMC Bioinformatics.* **14**, 152 (2013).
31. Dai, Q. *et al*. Study of LZ-word distribution and its application for sequence comparison. *Journal of Theor Biol.* **336**, 52–60 (2103).
32. Hamori, E. & Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J Biol Chem.* **258**, 1318–1327 (1983).
33. Randić, M., Vračko, M., Lerš, N. & Plavšić, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem Phys Lett.* **371**, 202–207 (2003).
34. Wąż, P. & Bielińskawąż, D. 3D-dynamic representation of DNA sequences. *J Mol Model.* **20**, 2141 (2014).
35. Jeong, B. S., Bari, A. T. G., Rokeya, R. M., Jeon, S. & Lim, C. G. Codon-based encoding for DNA sequence analysis. *Methods.* **67**, 373–379 (2014).
36. Bari, A. T., Reaz, M. R., Islam, A. K., Choi, H. J. & Jeong, B. S. Effective Encoding for DNA Sequence Visualization Based on Nucleotide's Ring Structure. *Evol Bioinfrom.* **9**, 251–261 (2013).
37. Xie, X., Guan, J. & Zhou, S. Similarity evaluation of DNA sequences based on frequent patterns and entropy. *Bmc Genomics.* **16**, 1–10 (2015).
38. Yu, H. J. & Huang, D. S. Graphical Representation for DNA Sequences via Joint Diagonalization of Matrix Pencil. *IEEE J Biomed Health.* **17**, 503–511 (2013).
39. Hou, W., Pan, Q. & He, M. A novel representation of DNA sequence based on CMI coding. *Physica A.* **409**, 87–96 (2014).
40. Li, Y., Liu, Q. & Zheng, X. DUC-Curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment. *Physica A.* **456**, 256–270 (2016).
41. Yin, C. Representation of DNA sequences in genetic codon context with applications in exon and intron prediction. *J Bioinf Comput Biol.* **13**, 1550004 (2015).
42. Peng, Y. & Liu, Y. A Novel Numerical Characterization for Graphical Representations of DNA Sequences. *Mini-Rev Org Chem.* **12**, 534–539 (2015).
43. Cheng, J., Shan & Ping, S. 4D Graphical representation research of DNA sequences. *Int J Biomath.* **08**, 47–58 (2015).
44. Manoj, K. G., Rajdeep, N. & Manoj, M. A new adjacent pair 2D graphical representation of DNA sequences. *J Biol Syst.* **21**, 196–244 (2013).
45. Zhang, Z. *et al*. ColorSquare: A colorful square visualization of DNA sequences. *MATCH Commun Math Comput Chem.* **68**, 621–637 (2012).
46. Dai, Q., Liu, X. & Wang, T. A novel graphical representation of DNA sequences and its application. *J Mol Graph Model.* **25**, 340–344 (2006).
47. Liu, Y. & Wang, T. Related matrices of DNA primary sequences based on triplets of nucleic acid bases. *Chem Phys Lett.* **417**, 173–178 (2006).
48. Jin, X. *et al*. A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding. *Physica A.* **461**, 325–338 (2016).
49. Li, Y. & Xiao, W. Circular Helix-Like Curve: An Effective Tool of Biological Sequence Analysis and Comparison. *Comput Math Method M.* **2**, 1–12 (2016).

Acknowledgements

This study is supported by the National Nature Science Foundation of China (Grant Number: 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, 61502343, 61672214, 61672223 and 61772192), the Guangxi Natural Science Foundation (Grant Number: 2017GXNSFAA198148).

Author Contributions

Zhiyi Mo and Yi Sun wrote the main manuscript text, Qilin Xiang prepared Figures 1–2 and Tables 2–3. Besides, Ming Zheng collected the gene data and prepared Table 1, Min Chen and Zejun Li give the analyses of different methods and prepared the Figure 3 and Table 4. Wen Zhu advised on adding some comparison of similar gene group of different methods.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.