A Novel Method for Alignment-free DNA Sequence Similarity Analysis Based on the Characterization of Complex Networks

Jie Zhou, Pianyu Zhong and Tinghui Zhang

Guangdong Key Laboratory of Computer Network, School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

ABSTRACT: Determination of sequence similarity is one of the major steps in computational phylogenetic studies. One of the major tasks of computational biologists is to develop novel mathematical descriptors for similarity analysis. DNA clustering is an important technology that automatically identifies inherent relationships among large-scale DNA sequences. The comparison between the DNA sequences of different species helps determine phylogenetic relationships among species. Alignment-free approaches have continuously gained interest in various sequence analysis applications such as phylogenetic inference and metagenomic classification/clustering, particularly for large-scale sequence datasets. Here, we construct a novel and simple mathematical descriptor based on the characterization of cis sequence complex DNA networks. This new approach is based on a code of three cis nucleotides in a gene that could code for an amino acid. In particular, for each DNA sequence, we will set up a cis sequence complex network that will be used to develop a characterization vector for the analysis of mitochondrial DNA sequence phylogenetic relationships among nine species. The resulting phylogenetic relationships among the nine species were determined to be in agreement with the actual situation.

KEYWORDS: DNA sequence, cis sequence network, similarity analysis, phylogenetic tree

CITATION: Zhou et al. A Novel Method for Alignment-free DNA Sequence Similarity Analysis Based on the Characterization of Complex Networks. *Evolutionary Bioinformatics* 2016:12 229–235 doi: 10.4137/EBO.S40474.

TYPE: Original Research

RECEIVED: June 28, 2016. **RESUBMITTED:** September 04, 2016. **ACCEPTED FOR PUBLICATION:** September 06, 2016.

ACADEMIC EDITOR: Liuyang Wang, Associate Editor

PEER REVIEW: Two peer reviewers contributed to the peer review report. Reviewers' reports totaled 964 words, excluding any confidential comments to the academic editor.

FUNDING: The research was supported in part by Natural Science Foundation of Guangdong Province (2015A030308017). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jiezhou@scut.edu.cn

Introduction

It is of great importance for biologists to analyze and understand the structure and function of a large volume of genomic DNA sequences.[1,2] However, it is very difficult to obtain biological information directly from large DNA sequences.[3] Determination of sequence similarity is one of the major steps in computational phylogenetic studies. During evolution, both nucleotide mutation and gene rearrangement occurred. One of the major tasks of computational biologists is to develop novel mathematical descriptors for similarity analysis.[1] DNA clustering is an important technology that automatically finds inherent relationships in large-scale collections of DNA sequences. However, the quality of DNA clustering resulting from this technique may still need to be significantly improved. The comparison between the DNA sequences of different species helps determine the phylogenetic relationship among various species.[2]

In recent years, various approaches have been proposed for generating DNA sequence information.[1–9] In 1990, Jeffrey[4] proposed a method known as Chaos Game Representation (CGR) of gene structures. The Chaos Game is essentially an iterated function scheme, similar to that used in the representation of the Sierpinski triangle obtained by plotting a sequence of points generated in a somewhat random fashion.[4,10] CGR is a novel holistic approach that provides a visual image of a DNA sequence that is quite different from the traditional linear arrangement of nucleotides. CGR of DNA sequences has received widespread attention for searching global patterns in long DNA sequences.[10–12] Sequence alignment has been frequently used as a powerful tool for comparing two DNA sequences.[13] However, with the divergence of species over time, subsequence rearrangements occurring during evolution make sequence alignment similarity scores less reliable.[1] Alignment-free methods for more efficient comparison and classification of DNA sequences than sequence alignment were developed in the past decade.[6] Graphical representations of DNA sequences are also powerful alignment-free tools for the analysis of DNA sequences.[1] The first three-dimensional (3D) geometric representation for DNA sequences was presented by Hamori and Ruskin.[14] Later, 2D,[15–20] 3D,[21–25] 4D,[26] 5D,[27] and 6D[28] representations of DNA sequences were developed. These methodologies represent DNA as matrices that were associated with the selected geometrical objects, as well as vectors that were composed of invariants of matrices that were used to compare DNA sequences. Qi et al.[1] introduced a novel method based on a graph theory to represent

DNA sequences mathematically for similarity analysis, in which they set up a weighted directed graph, whose adjacency matrix yields a representative vector. Two distance measurements for representative vectors are then defined to assess the similarity/dissimilarity of DNA sequences. However, to keep one-to-one mapping between a DNA sequence and its corresponding weighted directed multigraph, the weight of an arc between two nucleotides must retain a sufficient number of significant digits for large DNA sequences to increase its computational complexity. Hou et al.[2] proposed a new graphical representation method that adopted the telecommunication Coded Mark Inversion coding to represent four nucleotides, namely, A, G, C, and T. Their approach considers not only the sequences' structure but also the chemical structure of the DNA sequence. Jeong et al.[3] proposed a noble encoding approach for measuring the degree of similarity/dissimilarity between different species. Their approach preserves the physiochemical properties, positional information, and the codon usage bias of nucleotides. Recently, Thankachan et al.[7] presented ALFRED, an alignment-free distance computation method, which solves the generalized common substring search problem via exact computation. Pizzi[8] presented MissMax, an exact algorithm for the computation of the longest common substring with mismatches between each suffix of two sequences. Kumar et al.[9] introduced a 36-dimensional Periodicity Count Value that represented a particular nucleotide sequence by adapting a stochastic model. However, most alignment-free methods may lose the structural and functional information of DNA sequences because these mainly utilize feature extractions. Therefore, these may not fully reflect the actual differences among DNA sequences. Most of these techniques utilize genes that consist of several hundreds to thousands of bases to illustrate the effectiveness of their methods. However, for a large volume of genomic DNA sequences that comprises tens of thousands to millions of bases, the problem is computational complexity.

In most cases, it is necessary to identify a proper way to represent DNA sequences. In this paper, we construct a novel mathematical descriptor based on the characterization of complex networks. In particular, for each DNA sequence, a complex network is created and used to generate a 60-dimensional characterization vector that in turn would be employed in the analysis of the mitochondrial DNA sequence phylogenetic relationships among nine species. The present approach is capable of assessing tens to hundreds of thousands of bases of DNA sequences.

Complex Network Construction for DNA Sequence

The central dogma of molecular biology states that genetic information is transcribed from the DNA to the messenger RNA, which in turn in used in protein translation. A code of three nucleotide codes for an amino acid and various combinations are used to specify each of the 20 different amino acids occurring in living organisms.

Our method is mainly based on the central dogma. In addition, the similarity between two sequences based on alignment is completely based on the ordering of nucleotides. The basic idea of our model is to first construct the complex networks of a DNA sequence, which in turn are used in generating a characterization vector and integrated into a clustering algorithm for the analysis of phylogenetic relationships among various species.

To clarify the definition of different complex networks for a DNA sequence, we take the sequence shown in Figure 1 as an example. In accordance with the central dogma and considering computational complexity, for each DNA sequence, we will construct five complex networks called dinucleotide (trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide) cis sequence networks. The vertices of these cis sequence network are dinucleotides (trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides), and the edges of those networks are dinucleotide (trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide) cis sequence pairs. The vertex numbers of these cis sequence networks are $2^4 = 16$, $3^4 = 81$, $4^4 = 256$, $5^4 = 625$, and $6^4 = 1{,}296$, respectively. For the example sequence in Figure 1, the five complex network construction processes are illustrated in Figure 2. In Figure 2, in constructing a cis sequences network, each line has two cis sequences that represent two vertices between which there exists an edge. For example, in Figure 2A, cis sequence TG and CC are two vertices with an edge between TG and CC. Figure 3A–E shows the five nucleotide cis sequence networks of the example sequence that were laid out using Cytoscape 3.3.0.

Characterization Vector Construction for Complex Networks

Measurements of the topology of complex networks are essential for their characterization, analysis, classification, modeling, and validation. In a review, Costa et al.[31] surveyed the characterization of the main existing complex networks. The characterizations included distance-based measurements, clustering coefficients, assortativity, entropies, centrality,

TGCCATGCAATTCGTACTGCATGGCGAGATTCCTCTACCACTCTATAAATAGAGGCTCGATAAGTGTC
TTCACTTCATCAAATCAAAGAAATTGAGTATTAGCACGTATGGATATGGATGCTGCAAACAGGCTTT
CTGCAATTGCTGCCGAAATGGGGCAACTTCAAAATGAAATTCAAGAGCACCGTAGAGTGCTAAACTT
CCTTTTGAGGAGTGTTAGAACAATGGATCCTGCAAGGAAAGAAGCGCGCATTCGCGCTACCAGAGA

Figure 1. Example sequence.

A		B		C		D		E	
TG	CC	TGC	CAT	TGCC	ATGC	TGCCA	TGCAA	TGCCAT	GCAATT
CC	AT	CAT	GCA	ATGC	AATT	TGCAA	TTCGT	GCAATT	CGTACT
AT	GC	GCA	ATT	AATT	CGTA	TTCGT	ACTGC	CGTACT	GCATGG
GC	AA	ATT	CGT	CGTA	CTGC	ACTGC	ATGGC	GCATGG	CGAGAT
AA	TT	CGT	ACT	CTGC	ATGG	ATGGC	GAGAT	CGAGAT	TCCTCT
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Figure 2. Examples of nucleotide cis sequence network constructions. (**A**) Dinucleotide cis sequences network construction, (**B**) trinucleotide cis sequences network construction, (**C**) tetranucleotide cis sequences network construction, (**D**) pentanucleotide cis sequences network construction, and (**E**) hexanucleotide cis sequences network construction.

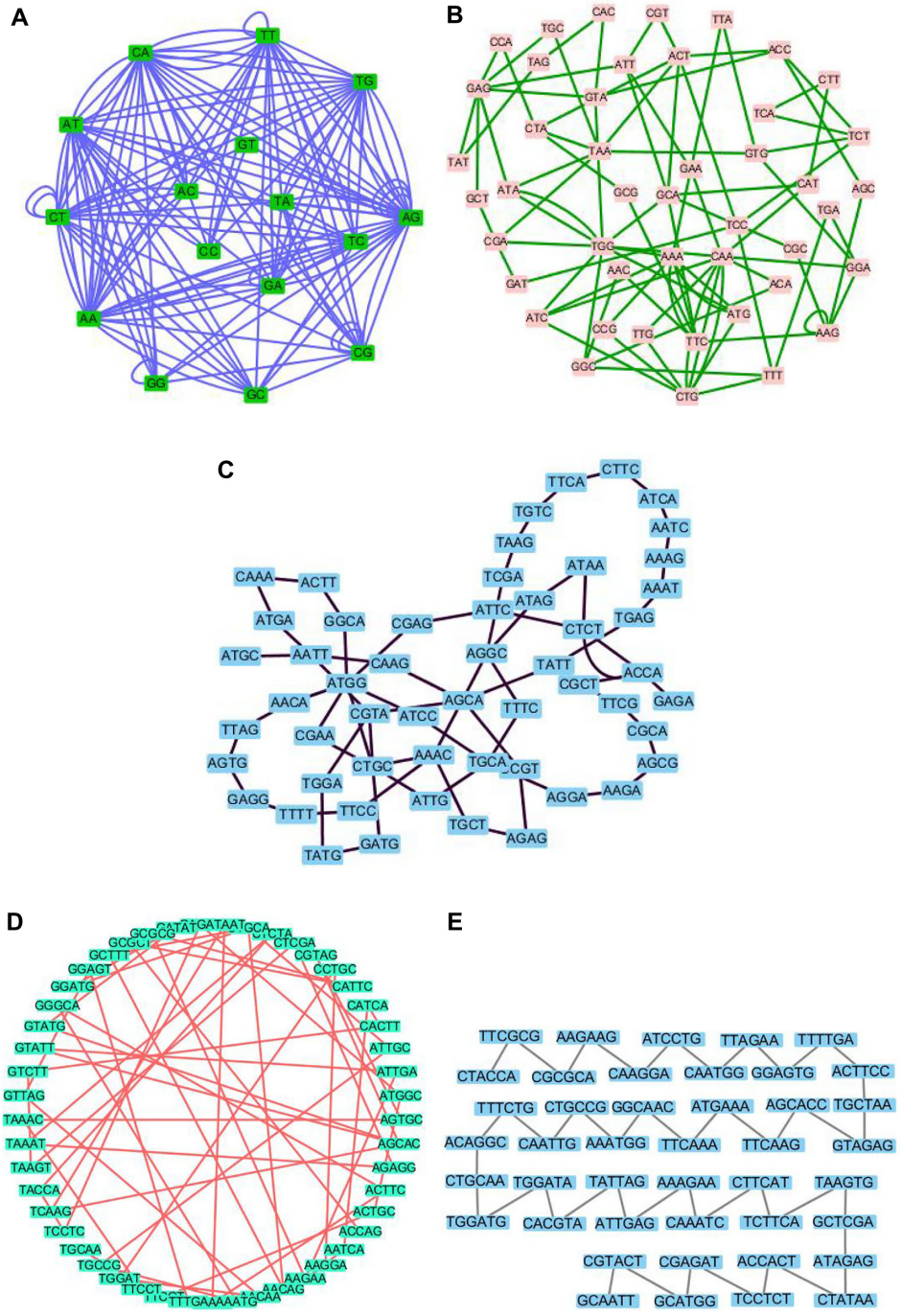

Figure 3. Five nucleotide cis sequence networks of the example sequence. (**A**) Dinucleotide cis sequences network, (**B**) trinucleotide cis sequences network, (**C**) tetranucleotide cis sequences network, (**D**) pentanucleotide cis sequences network, and (**E**) hexanucleotide cis sequences network.

subgraphs, spectral analysis, community-based measurements, hierarchical measurements, and fractal dimensions.

In this paper, for each of the five nucleotide cis sequence networks of a given DNA sequence, we calculated 15 global characterizations (Table 1). All the global characterizations of the complex networks used were calculated using Cytoscape 3.3.0[29] and networkX 1.11.[30] In the calculation of characterizations, we treated all the networks as undirected. We refer interested readers to the survey[31] for the definitions of the 15 global characterizations of complex network listed in Table 1. Some characteristics of some complex networks are the same among nine species. For example, among nine species, the values of connected components characteristic of dinucleotide cis sequences complex network are the same. Therefore, we omitted this characteristic in dinucleotide cis sequences complex network. We totally deleted 15 characteristics shared by all complex networks in different complex networks. After deleting those complex network characterizations of each DNA sequence, a 60-dimensional vector was generated (Supplementary material) and used in the subsequent analyses.

In addition, for a large volume of DNA sequences, the characterizations of each of the five nucleotide cis sequence networks revealed a short characteristic path length, one connected component, and low diameter and density, and each of the networks displayed common characteristics of a biological network such as scale free and small words, thereby suggesting that the inferred networks were not random and presented features of complex networks. The five nucleotide cis sequence networks of the example sequence shown in Figure 1 do not present the characteristic of scale-free small words [for example, Figure 1E is just a path] because the example sequence is short.

Table 1. Fifteen global characterizations of complex network.

Cluster coefficient	Characteristic path length
Connected components	Number of self loops
Diameter	Multi edge node pairs
Radius	Global efficiency
Centralization	Harmonic mean
Average number of neighbors	Transitivity
Density	Central point dominance
Network heterogeneity	

Materials

In the present study, we used mitochondrial DNA sequences for verifying the validity of our method because compared to autosomal and sex chromosome DNA sequences mitochondrial DNA sequences are relatively short. Mitochondrial DNA sequences were downloaded from ftp://ftp.ensembl.org/pub/current_fasta/.8[32] The mitochondrial sequence of a total of nine species were downloaded and shown in Table 2.

From above, five nucleotide cis sequence complex networks and 60-dimensional vector associated with each of the nine species' mitochondrial DNA sequences were generated (Supplementary material).

Distance Measurements for Similarity Calculations

Each of the DNA sequences was mapped to a characterization vector in the 60-dimensional linear space. The comparison between DNA sequences is now reduced to the comparison between these 60-dimensional vectors. We applied two standard measurements: one is the distance between two 60-dimensional vectors and the other is the similarity/dissimilarity among the different species. For two species s and h, we denote the representative vectors by $V_s = \{v_1^s, v_2^s, \cdots, v_{60}^s\}$ and $V_h = \{v_1^h, v_2^h, \cdots, v_{60}^h\}$, respectively.

We present the first similarity/dissimilarity matrix based on the Euclidean distance between two vectors V_s and V_h, which is based on the assumption that two DNA sequences are similar if the corresponding two vectors have similar magnitudes, ie,

$$d(V_s, V_h) = \sqrt{\sum_{i=1}^{60} (v_i^s - v_i^h)^2}.$$

The second similarity/dissimilarity matrix is based on the cosine of the angle included between vectors V_s and V_h, ie,

$$C(V_s, V_h) = \frac{\sum_{i=1}^{60} v_i^s \cdot v_i^h}{\sqrt{\sum_{i=1}^{60} (v_i^s)^2 \cdot \sum_{i=1}^{60} (v_i^h)^2}}$$

Table 2. Length of the mitochondrial DNA sequences of nine species.

NUMBER	SPECIES	LENGTH (bp)	NUMBER	SPECIES	LENGTH (bp)
1	*Tetraodon nigroviridis*	16,462	6	*Gorilla gorilla*	16,412
2	*Oryzias latipes*	16,714	7	*Pan troglodytes*	16,554
3	*Bos taurus*	16,338	8	*Mus musculus*	16,299
4	*Equus caballus*	16,660	9	*Rattus norvegicus*	16,313
5	*Ovis aries*	16,616			

We will use different similarity/dissimilarity metrics for numerical analysis and determine phylogenetic analysis to show the evolutionary proximity and distance among different species. We further analyze the similarity/dissimilarity of the resulting complex networks in terms of the local characterizations of topological coefficients.

Results and Discussion

Similarity analysis of global characterizations. Numerical representation of biomolecular sequences facilitates application of conventional pattern recognition tools and techniques such as various classifiers and clustering techniques to group and classify sequences.[9] By our method, each of the five nucleotide cis sequence networks can be represented by a 60-dimensional vector using 15 global characterizations, which can then be used in calculating the similarity/dissimilarity matrix using the two measurements of defining distances between two characterization vectors as well as in a clustering algorithm for the analysis of phylogenetic relationships of nine species based on their mitochondrial DNA sequences. We evaluated and applied the proposed characterizations of complex networks model similarity measure to the Matlab clustering to perform clustering. Using the similarity/dissimilarity matrix, we generate the phylogenic tree using the functions "pdist", "linkage", and "dendrogram" in Matlab 7.12.0 (R2011a).

Tables 3 and 4 show the upper triangular part of the similarity/dissimilarity matrix among nine species listed in Table 1 based on two measurements. Figures 4 and 5 are the phylogenic trees for the nine species constructed applying the similarity/dissimilarity matrix of Tables 3 and 4 and Matlab as earlier described. Tables 3 and 4 show the digital in the first row, and the first column are the labels of species listed in Table 2. In Figures 4 and 5, the corresponding digitals on the horizontal axis are the labels of species shown in Table 2. The results of the phylogenic tree reflect the quality of the similarity matrix that efficiently extracts evolutionary information from DNA sequences with our encoding method. For the Euclidean distance (Table 3), the smaller the distance is, the more similar the two sequences are, whereas for the cosine of vector-included angle (Table 4), the larger the distance is, the more similar the two sequences are. Figures 4 and 5 show more similar species pairs such as *Tetraodon nigroviridis* and *Oryzias latipes*, *Bos taurus* and *Equus caballus*, *Gorilla gorilla* and *Pan troglodytes*, and *Mus musculus* and *Rattus norvegicus*, which were also in agreement with the actual situation. One can also find that the two phylogenic trees of these nine species have similar topologies.

Table 3. Upper triangular part of similarity/dissimilarity matrix based on the Euclidean distance of vectors.

	1	2	3	4	5	6	7	8	9
1	0	85.06559	121.8426	123.2042	132.7759	136.498	116.8607	223.4014	200.215
2	0	0	204.6468	204.7177	215.1477	212.3071	200.7845	307.5472	284.6062
3	0	0	0	23.06878	67.05954	99.11729	40.3144	120.4128	86.46105
4	0	0	0	0	62.58154	101.2624	37.98756	123.2114	88.81361
5	0	0	0	0	0	72.55671	32.33285	100.3644	88.23187
6	0	0	0	0	0	0	73.92698	118.1972	112.355
7	0	0	0	0	0	0	0	110.2725	87.37712
8	0	0	0	0	0	0	0	0	46.35559
9	0	0	0	0	0	0	0	0	0

Table 4. Upper triangular part of similarity/dissimilarity matrix based on the cosine of a vector-included angle.

	1	2	3	4	5	6	7	8	9
1	1	0.999418	0.998967	0.998368	0.996763	0.996361	0.998082	0.99256	0.995654
2	0	1	0.997032	0.996216	0.993468	0.993522	0.995451	0.987908	0.991999
3	0	0	1	0.999871	0.998988	0.997535	0.999706	0.996287	0.998545
4	0	0	0	1	0.999017	0.997267	0.999652	0.996414	0.998759
5	0	0	0	0	1	0.99853	0.999717	0.998892	0.999756
6	0	0	0	0	0	1	0.998493	0.997919	0.998505
7	0	0	0	0	0	0	1	0.997957	0.999481
8	0	0	0	0	0	0	0	1	0.99932
9	0	0	0	0	0	0	0	0	1

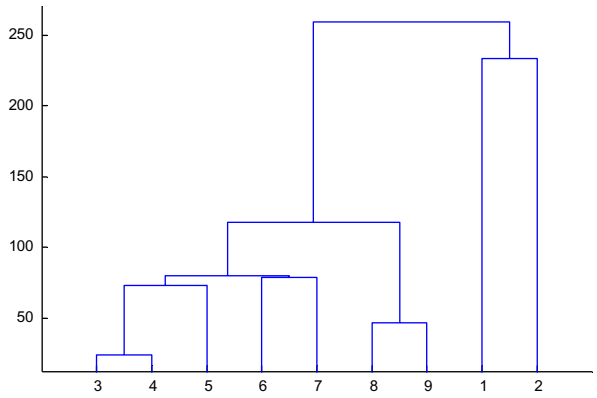

Figure 4. The phylogenic tree of nine species using the upper triangular part of similarity/dissimilarity matrix based on Euclidean distance of vectors.

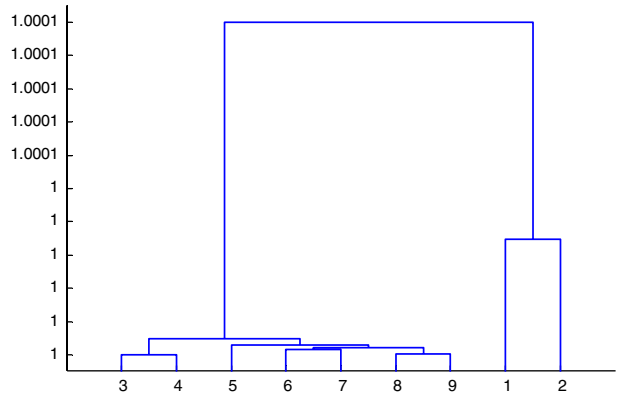

Figure 5. The phylogenic tree of nine species using the upper triangular part of similarity/dissimilarity matrix based on cosine of vector-included angle.

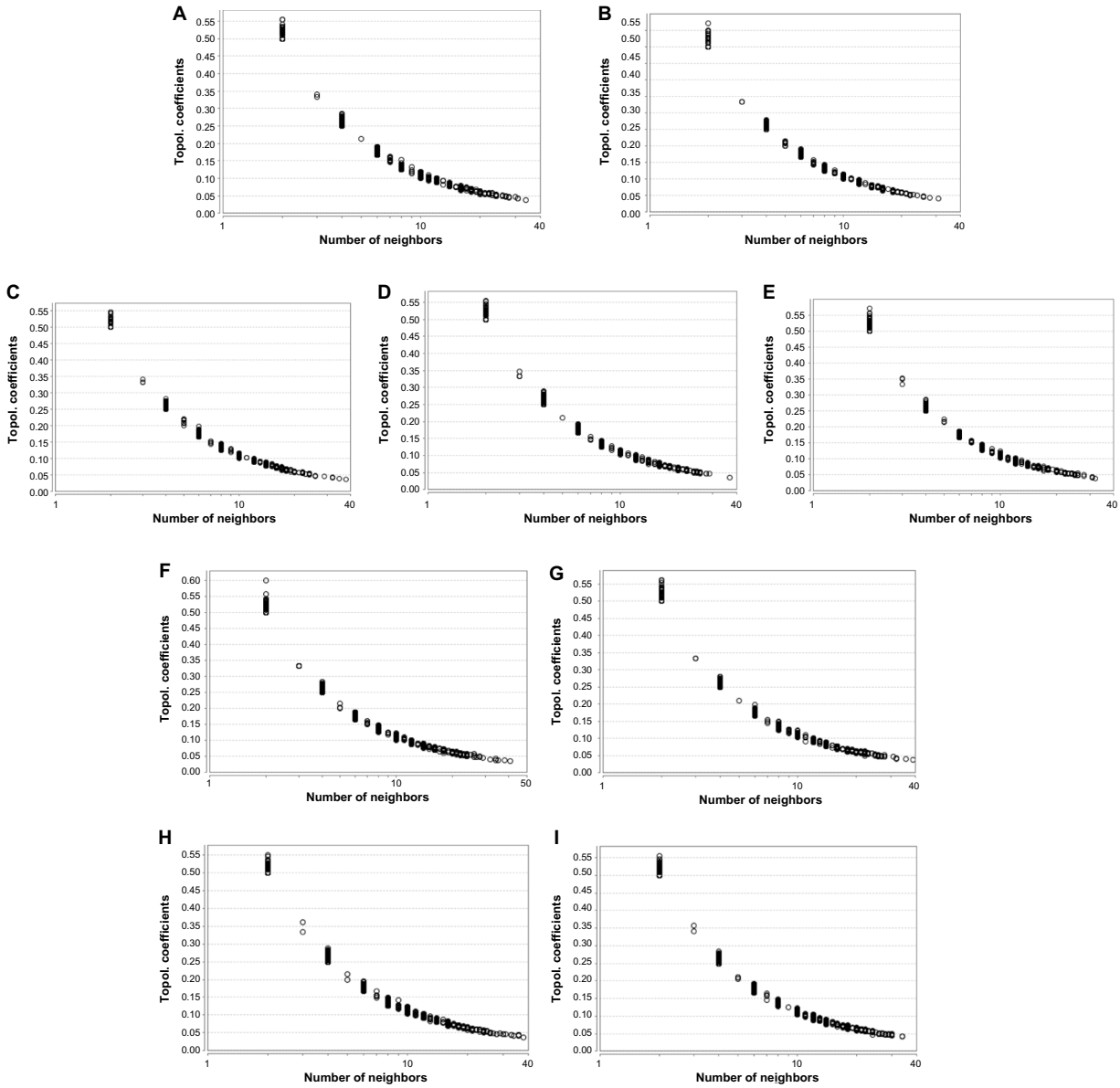

Figure 6. Topological coefficients of pentanucleotide cis sequence networks of nine species. (**A**) *Tetraodon nigroviridis* pentanucleotide, (**B**) *Oryzias latipes* pentanucleotide, (**C**) *Bos taurus* pentanucleotide, (**D**) *Equus caballus* pentanucleotide, (**E**) *Ovis aries* pentanucleotide, (**F**) *Gorilla gorilla* pentanucleotide, (**G**) *Pan troglodytes* pentanucleotide, (**H**) *Mus musculus* pentanucleotide, and (**I**) *Rattus norvegicus* pentanucleotide.

Similarity analysis of local characterizations. We analyze the similarity/dissimilarity of the local characterizations[33] of these complex networks. We consider the topological coefficients of pentanucleotide cis sequence networks of nine species as an example (Fig. 6).

From the topological coefficients of pentanucleotide cis sequence networks of nine species, the more similar species pairs include *Tetraodon nigroviridis* and *Oryzias latipes* [Fig. 6A and B], *Bos taurus* and *Equus caballus* [Fig. 6C and D], *Gorilla gorilla* and *Pan troglodytes* [Fig. 6F and G], and *Mus musculus* and *Rattus norvegicus* [Fig. 6H and I]. Figure 6E shows that the topological coefficients figure of *Ovis aries* is similar to those of *Bos taurus* [Fig. 6C] and *Equus caballus* [Fig. 6D], which is in agreement with Figure 4. The similarity/dissimilarity of the local characterizations of the topological coefficients of pentanucleotide cis sequence networks of nine species coincides with the results of global characterization.

Conclusions

We present a new method for the characterization of a complex networks model for alignment-free DNA sequence similarity analysis. This new approach is based on a code of three cis nucleotides in a gene that could code for an amino acid. The similarity/dissimilarity matrix for the nine complete mitochondrial DNA sequences belonging to different species is built, and the elements of the similarity matrix are used to construct phylogenic tree. When the mitochondrial DNA sequences were compared between different species, the results were in agreement with the actual situation.

Author Contributions

JZ conceived and designed the experiments and was a major contributor in writing the manuscript. PYZ and THZ analyzed the data and the result. All authors reviewed and approved of the final manuscript.

Supplementary Material

Characterizations of five complex cis sequence network of nine species.

REFERENCES

1. Qi X, Wu Q, Zhang Y, Fuller E, Zhang C-Q. A novel model for DNA sequence similarity analysis based on graph theory. *Evol Bioinform Online*. 2011;7:149–58.
2. Hou W, Pan Q, He M. A novel representation of DNA sequence based on CMI coding. *Physica A*. 2014;409:87–96.
3. Jeong B-S, Golam Bari ATM, Rokeya Reaz M, Jeon S, Lim C-G, Choi H-J. Codon-based encoding for DNA sequence analysis. *Methods*. 2014;67:373–9.
4. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res*. 1990;18:2163–70.
5. Bao J, Yuan R, Bao Z. An improved alignment-free model for DNA sequence similarity metric. *BMC Bioinformatics*. 2014;15(321):1–15.
6. Yin C, Chen Y, Yau SS. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. *J Theor Biol*. 2014;359:18–28.
7. Thankachan SV, Chockalingam SP, Liu Y, Apostolico A, Aluru S. ALFRED: a practical method for alignment-free distance computation. *J Comput Biol*. 2016;23(6):452–60.
8. Pizzi C. MissMax: alignment-free sequence comparison with mismatches through filtering and heuristics. *Algorithms Mol Biol*. 2016;11:6.
9. Kumar R, Mishra BK, Lahiri T, et al. PCV: an alignment free method for finding homologous nucleotide sequences and its application in phylogenetic study. in Interdisciplinary Sciences: Computational Life Sciences, Dong-Qing Wei (Editor-in-Chief), Springer-Verlag, 2016:1–11.
10. Carl Leinbach L. Using CAS to show chaos game representations of DNA sequences. *Int J Technol Math Educ*. 2013;20(3):125–30.
11. Deng W, Luan Y. Analysis of similarity/dissimilarity of dna sequences based on chaos game representation. *Abstract and Applied Analysis*. 2013;2013:1–6.
12. Gutierrez JM, Rodriguez MA, Abramson G. Multifractal analysis of DNA sequences using a novel chaos-game representation. *Physica A*. 2001;300:271–84.
13. Mount DM. *Bioinformatics*: *Sequence and Genome Analysis*. 2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2004. NY. ISBN0-87969-608-7.
14. Hamori E, Ruskin J. H curves: a novel method of representation of nucleotide series especially suited for long DNA sequences. *J Biol Chem*. 1983;258:1318–27.
15. Guo X, Randić M, Basak SC. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem Phys Lett*. 2001;350:106–12.
16. Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequence based on novel 2-D graphical representation. *J Chem Inform Comput Sci*. 2003;371:202–7.
17. Randić M, Vračko M, Zupan J, Novic M. Compact 2-D graphical representation of DNA. *Chem Phys Lett*. 2003;373:558–62.
18. Randić M. Graphical representations of DNA as 2-D map. *Chem Phys Lett*. 2004;386:468–71.
19. Liu X, Dai Q, Xiu Z, Wang T. PNN-curve: a new 2D graphical representation of DNA sequences and its application. *J Theor Biol*. 2006;243:555–61.
20. Huang G, Liao B, Li Y, Liu Z. H curves: a novel 2D graphical representation for DNA sequences. *Chem Phys Lett*. 2008;462:129–32.
21. Liao B, Wang T. 3-D graphical representation of DNA sequences and their numerical characterization. *J Mol Struct (Theochem)*. 2004;681:209–12.
22. Qi X, Wen J, Qi Z. New 3D graphical representation of DNA sequence based on dual nucleotides. *J Theor Biol*. 2007;249:681–90.
23. Qi Z, Fan T. PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization. *Chem Phys Lett*. 2007;442:434–40.
24. Cao Z, Liao B, Li R. A group of 3D graphical representation of DNA sequences based on dual nucleotides. *Int J Quantum Chem*. 2008;108:1485–90.
25. Yu J, Sun X, Wang J. TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. *J Theor Biol*. 2009;261:459–68.
26. Chi R, Ding K. Novel 4D numerical representation of DNA sequences. *Chem Phys Lett*. 2005;407:63–7.
27. Liao B, Li R, Zhu W. On the similarity of DNA primary sequences based on 5-D representation. *J Math Chem*. 2007;42:47–57.
28. Liao B, Wang T. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping trinucleotides of nucleotide bases. *J Chem Inform Comput Sci*. 2004;44:1666–70.
29. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
30. Available at: https://pypi.python.org/pypi/networkx/
31. Costa LF, Rodrigues FA, Travieso G, Villas Boas PR. Characterization of complex networks: a survey of measurements. *Adv Phys*. 2007;56:167–242.
32. Yates A, Akanni W, Amode MR, et al. Ensembl 2016. *Nucleic Acids Res*. 2016;44(D1):D710–6.
33. Caraiani P. Characterizing emerging European stock markets through complex networks: from local properties to self-similar characteristics. *Physica A*. 2012;391:3629–37.