International Journal of Hybrid Information Technology
Vol.8, No.1 (2015), pp.73-84
http://dx.doi.org/10.14257/ijhit.2015.8.1.06

Characterization and Similarity Analysis of DNA Sequences Considering Codon Degeneracy

Yong-Bin Zhao[1,2], Zhao-Hui Qi[2]* and Ai-Ping Yan[3]

[1]State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China
[2]School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[3]SiFang College, Shijiazhuang Tiedao University, Hebei, China
zhqi_wy2013@163.com

Abstract

We introduce a new 2D graphic representation of DNA sequences considering codon degeneracy. Derived from the graphic representation, a multi-component vector is proposed to characterize quantitatively DNA sequences. Then we use the graphic representation and the vector to perform the phylogenetic analysis on two datasets, the complete coding sequences of the β -globin genes of 11 species and the HA genes of 16 influenza A (H1N1) isolates. The experimental assessment demonstrates the efficiency of the proposed method.

Keywords: *Graphic representation; DNA sequence; Similarity analysis; Codon degeneracy*

1. Introduction

In recent years, graphical methods have emerged as a powerful tool for dealing with a wide variety of biological problems. For example, various graphic methods have been successfully used to study enzyme-catalyzed system [1-9], base frequency distribution in the anti-sense strands [10], protein folding kinetics [11,12], condon usage [13,14], biological sequence representation [15], protein subcellular location [16], HBV viral infections [17], the fingerprint of SARS coronavirus [18], and so on. As for an important graphic tool for biological problems, graphic representation methods of DNA sequence have recently become a useful approach to the biologists in studying the characterization and similarity analysis of DNA sequence. For example, some graphic methods in the literature [19-22] have been successfully used to study the similarity comparison of DNA sequences. But some of them have their own limitation. A major limitation is that graphic representation of DNA sequence is accompanied by some loss of information associated with crossing and overlapping of the resulting curve by itself. More details about the limitation have been discussed in a review by Nandy *et al.* [19]. In [19], authors took a comprehensive comparison of graphic representation methods of DNA sequence. The review gives good suggestion and plays an important role in the road ahead of developing different graphic approaches. Then, people began to take new attempt. For example, Bielinska *et al.* in [20] introduces a new graphic representation of the DNA sequences, which is called as 2D-dynamic graphs. The method removes most of the degeneracy from a previous model (Nandy plots) [21] and becomes a robust quantitative technique to detect regions of interest in DNA sequence. Later, in [22] Bielinska *et al.* perform similarity studies of DNA sequences based on descriptors derived from their 2D-dynamic graphs introduced in the previous paper [20]. They propose a set of descriptors, and draw a conclusion that different descriptors applied to the same pair of sequences lead to different similarity relations due to the high degree of complexity of DNA/RNA

ISSN: 1738-9968 IJHIT

sequences. Recently, more novel graphic representation methods are proposed to study similarities/dissimilarities of DNA sequences [23-29]. These novel methods play an important role in analyzing DNA sequence by graphic methods.

In the past years the nucleotide triplet codons analysis has also been tried by several authors. Balaban *et al.* in [30] consider a novel representation of a DNA sequence representing the standard genetic code. Wang and Zhang in [31] outline a procedure giving so-called asymmetry of direct-complementary triplets (ADCT) scheme to describe a DNA sequence. Graphic representations based on nucleotide triplet codons are also used to characterize protein sequences. For instance, graphic approaches were successfully used to study graphic representation of protein sequences [32].

In this paper, we propose a new graphic representation of the DNA sequence considering codon degeneracy. The representation considers the conversion from triplet codons to amino acids. It establishes the relation between DNA sequence and all the triplet codons. Such a representation of DNA sequence is accompanied by some loss of information because of the synonyms of codons. The tendency for similar amino acids to be represented by related codons minimizes the effects of mutations. It increases the probability that a single random base change will result in no amino acid substitution or in one involving amino acids of similar character. The proposed graphic representation method just utilizes the loss of information (the synonyms of codons) to reveal numerical characterization of DNA sequences from the views of evolution history and mutation effect. For example, we consider two hypothetical sequences shown below in which we grouped nucleic acids in triplets: (1) TTT GCT TCG CTC and (2) TTC GCT TCG CTG. There are two mutations from segment (1) to segment (2): T to C and C to G. By the approaches of comparison of DNA sequences proposed in [19-29], the sequence (1) and (2) are distinct each other. However, the single random base change does not result in amino acid substitution. The mutations of TTT to TTC and CTC to CTG, that is to say, have no any effect on amino acid. Considering the effect of mutations between sequences, we think the sequence (1) and (2) are uniform. Related amino acids often have related codons, minimizing the effects of mutation. As for organism, the ability to minimizing effects of mutation has importance role in keeping genetic characteristic. Here, we propose the 2D graphic representation of DNA sequences considering codon degeneracy.

2. 2D Graphic Representation of DNA Sequences Considering Codon Degeneracy

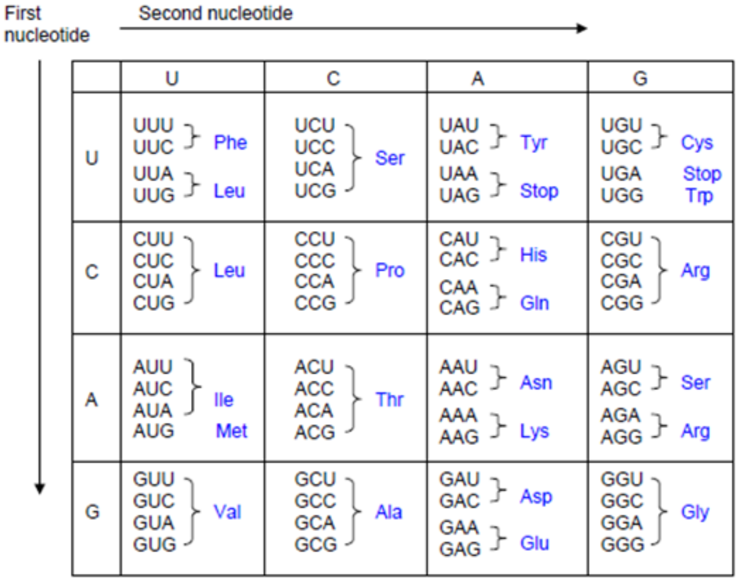
In this paper, we construct 2D graphic representation of DNA sequences based on triplets of nucleotide bases instead of individual nucleotide bases. The consideration has the following reasons.

The sequence of a coding strand of DNA consists of triplets of nucleotide bases corresponding to the amino acid of a protein read from N-terminus to C-terminus. There are 64 codons. Each of these codons has a specific meaning in protein synthesis (61 codons represent amino acids; 3 codons cause the termination of protein synthesis). The standard genetic codes are summarized in Figure 1, where the letter U standing for uracil in RNA is equal to the letter T denoting thymine in DNA. The number of codons (61) is more than amino acids (20). Almost all amino acids are represented by more codons. These codons representing the same or related amino acids tend to be similar in sequence. The tendency for similar amino acids to be represented by related codons minimizes the effects of mutations. This is called as the codon degeneracy. It increases the probability that a single random base change will not result in amino acid substitution. For example, a mutation of CUC to CUG has no effect on the corresponding codon, since they code the same amino acid *Leu* (*leucine*).

In detail, let $S = s_1 s_2 \cdots s_n$ be an arbitrary DNA sequence. Then based on the standard genetic codons in Figure 1, we define a mapping ϕ , which mappings s into a plot set, as the following,

$$\phi(s_i s_{i+1} s_{i+2}) \begin{cases} (i,0) & if \quad s_i s_{i+1} s_{i+2} = TAA, TAG, TGA \\ (i,1) & if \quad s_i s_{i+1} s_{i+2} = ATG \\ (i,2) & if \quad s_i s_{i+1} s_{i+2} = GCT, GCC, GCA, GCG \\ (i,3) & if \quad s_i s_{i+1} s_{i+2} = CGT, CGC, CGA, CGG, AGA, AGG \\ (i,4) & if \quad s_i s_{i+1} s_{i+2} = AAT, AAC \\ (i,5) & if \quad s_i s_{i+1} s_{i+2} = GAT, GAC \\ (i,6) & if \quad s_i s_{i+1} s_{i+2} = TGT, TGC \\ (i,7) & if \quad s_i s_{i+1} s_{i+2} = CAA, CAG \\ (i,8) & if \quad s_i s_{i+1} s_{i+2} = GAA, GAG \\ (i,9) & if \quad s_i s_{i+1} s_{i+2} = GGT, GGC, GGA, GGG \\ (i,10) & if \quad s_i s_{i+1} s_{i+2} = ATT, ATC, ATA \\ (i,11) & if \quad s_i s_{i+1} s_{i+2} = CAT, CAC \\ (i,12) & if \quad s_i s_{i+1} s_{i+2} = TTA, TTG, CTT, CTC, CTA, CTG \\ (i,13) & if \quad s_i s_{i+1} s_{i+2} = AAA, AAG \\ (i,14) & if \quad s_i s_{i+1} s_{i+2} = TTT, TTC \\ (i,15) & if \quad s_i s_{i+1} s_{i+2} = CCT, CCC, CCA, CCG \\ (i,16) & if \quad s_i s_{i+1} s_{i+2} = TCT, TCC, TCA, TCG, AGT, AGC \\ (i,17) & if \quad s_i s_{i+1} s_{i+2} = ACT, ACC, ACA, ACG \\ (i,18) & if \quad s_i s_{i+1} s_{i+2} = TGG \\ (i,19) & if \quad s_i s_{i+1} s_{i+2} = TAT, TAC \\ (i,20) & if \quad s_i s_{i+1} s_{i+2} = GTT, GTC, GTA, GTG \end{cases}$$

where i denotes the i -th base of sequence S . Then based on the mapping ϕ we can maps any DNA sequence S into a plot set. For example, the corresponding plot set of the sequence ATGGTGCACC is {(1, 1), (2, 18), (3, 9), (4, 20), (5, 6), (6, 2), (7, 11), (8, 17)}. Connection all plots of the plot set we can get a 2D curve. Here, for convenience, the curve is named as CCD-curve (CCD: Considering Codon Degeneracy).



First nucleotide ↓ | Second nucleotide →

	U	C	A	G
U	UUU UUC } Phe UUA UUG } Leu	UCU UCC UCA UCG } Ser	UAU UAC } Tyr UAA UAG } Stop	UGU UGC } Cys UGA Stop UGG Trp
C	CUU CUC CUA CUG } Leu	CCU CCC CCA CCG } Pro	CAU CAC } His CAA CAG } Gln	CGU CGC CGA CGG } Arg
A	AUU AUC AUA } Ile AUG Met	ACU ACC ACA ACG } Thr	AAU AAC } Asn AAA AAG } Lys	AGU AGC } Ser AGA AGG } Arg
G	GUU GUC GUA GUG } Val	GCU GCC GCA GCG } Ala	GAU GAC } Asp GAA GAG } Glu	GGU GGC GGA GGG } Gly

Figure 1. All the Triplet Codons

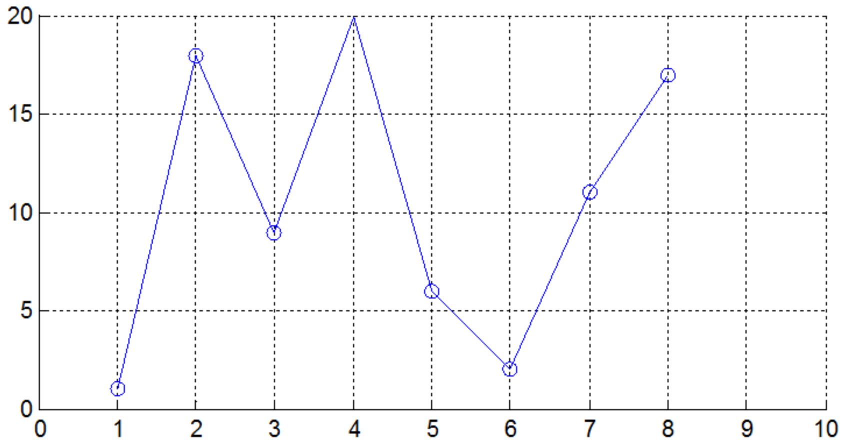


Figure 2. CCD-curve of the Sequence ATGGTGCACC

In Figure 2, we show the 2D graphic representation of the sequence. The CCD-curve has two main properties as follows.

Property 1. *There is no circuit from DNA sequence to CCD-curve*

Proof. We suppose that there exits one or more circuits in CCD-curve. Then there is at least two points overlapping each other, (x_i, y_i) and (x_j, y_j), $i \neq j$. On the other hand, according to the mapping ϕ if $x_i = x_j$ and $y_i = y_j$, the index i and j are equal. This contradicts $i \neq j$. Hence, there is exists no circuit in CCD-curve.

Property 2. *There is only one 2D graphic representation corresponding to a given DNA sequence, not vice versa.*

Proof. This property shows that for a given DNA sequence there is a unique CCD-curve correspondingly, not vice versa. Let (x_i, y_i) be the coordinate of the i-th trinucleotide of a DNA sequence. Then according to the mapping ϕ we can get the following some equations,

$(1)\begin{cases} x_i = i \\ y_i = 0, \end{cases}$ $(2)\begin{cases} x_i = i \\ y_i = 1, \end{cases}$ $(3)\begin{cases} x_i = i \\ y_i = 2, \end{cases}$

…

$(20)\begin{cases} x_i = i \\ y_i = 19, \end{cases}$ $(21)\begin{cases} x_i = i \\ y_i = 20. \end{cases}$

From the 20 equations we find that every (x_i, y_i) is different from each other. Furthermore, according to the mapping ϕ each trinucleotide has a unique one of the 20 equations correspondingly. Hence, for any one trinucleotide there is a unique coordinate correspondingly. It means that there must be only one unique CCD-curve for a given DNA sequence.

At the same time, according to the mapping ϕ we can find that any one point (x_i, y_i) of a CCD-curve corresponds to one or more trinucleotide because of the codon degeneracy. Hence, for a given CCD-curve there are one or more DNA sequences. These DNA sequences are thought as the same evolution characteristic when we consider the codon degeneracy.

3. Application

3.1. Sequence Invariants derived from the CCD-curve

In order to find some of the invariants sensitive to the form of the directed curve, we use another mathematical object, an N-dimension vector (may be also looked on as one-dimension matrix), to represent a DNA sequence. The mathematical object is directly derived from the CCD-curve. The initial developments in characterizing DNA sequences using matrix methods were by Randić and Vracko [33] and Randić *et al.* [34] where they used 2D and 3D graphic representation of DNA sequences to generate descriptor matrices. The matrix technique can capture the essence of the base composition and distribution of the sequence in a quantitative manner which would facilitate sequence identification and comparison of similarities and dissimilarities of different sequences. Similarly, the N-dimension vector has similar function. In this paper we will use vector method to characterize DNA sequences.

We associate the CCD-curve with a characteristic vector V . The V is a 21-dimension vector, whose element i is the y -coordinate subtracted directly from the i -th vertex of CCD-curve.

$V = (v_0, v_1, \cdots v_{20})$

$v_0 = 0$,

$v_1 = \underbrace{1 + 1 + \cdots + 1}_{n_1}$,

$v_2 = \underbrace{2 + 2 + \cdots + 2}_{n_2}$, $\cdots$,

$v_{20} = \underbrace{20 + 20 + \cdots + 20}_{n_{20}}$

The parameters $n_1, n_2, \cdots, n_{20}$ denote the number of the corresponding trinucleotides of the mapping ϕ , respectively. For example, for the formula $v_2 = \underbrace{2 + 2 + \cdots + 2}_{n_2}$, the v_2 is the sum of the y -coordinate in the CCD-curve whose value is 2. The n_2 is the number of the trinucleotides, GCT, GCC, GCA and GCG. Then based on the characteristic vector V

we give a evolution distance computing scheme about two characteristic vectors V_a and V_b . The computing formula is the following,

$$D(V_a, V_b) = \frac{1 - V_a \square V_b \Big/ \|V_a\| \, \|V_b\|}{2}$$

$$= \frac{1 - \sum_{i=1}^{21} v_a(i) \times v_b(i) \Bigg/ \sqrt{\sum_{i=1}^{21} [v_a(i)]^2 \times \sum_{i=1}^{21} [v_b(i)]^2}}{2}$$

Then we use the distance $D(V_a, V_b)$ of two vectors to represent the distances of two CCD-curves. According to the Property 1 and Property 2 of CCD-curve, the distance $D(V_a, V_b)$ can be used to describe the evolution distance between the DNA sequences S_a and S_b . Two DNA sequences would be considered relatively similar if the $D(V_a, V_b)$ is small.

3.2. Similarity Analysis

Firstly, we discuss the graphic characterization of DNA sequences based on CCD-curve. In order to facilitate the quantitative comparison of different DNA sequences, we apply the proposed method to the complete coding sequences of the β -globin genes with an examination of similarities/dissimilarities among 11 species of Table 1: Human (U01317.1), Goat (M15387.1), North American opossum (J03642.1), Gallus (V00409.1), Lemur (M15734.1), House mouse (V00726.1), Rabbit (V00882.1), Norway rat (X06701.1), Gorilla (X61109.1), Bovine (X00376.1) and Chimpanzee (X02345.1). For simplification, in Table 1 we only list the primary DNA sequences of the complete coding sequences of part species.

The results of the examination are listed in Table 2. Observing Table 2, we find that the distances among Human, Gorilla and Chimpanzee are the smallest ones. This shows that Human-Gorilla, Human- Chimpanzee and Gorilla-Chimpanzee are the most similar species pairs. Rather, Gallus and Opossum show strong dissimilarity among the 11 species because of the larger computing distance. In addition, the pair Rat-Mouse also show the closer computing distance. The results are similar to that reported in the literatures [23-24, 35]. This is not accidental, but shows that these species have close evolution relationship.

Table 1. Complete Coding Sequences of β -globin Genes of 11 Species

Species	*Complete coding sequence*
Human	ACCESSION, U01317.1; ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTAC CCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTG TTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTA GTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTG AGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCA ACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGT GCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAA GTATCACTAA
Goat	ACCESSION, M15387.1 ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAA GTGGATGAAGTTGGTGCTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCCTGG ACTCAGAGGTTCTTTGAGCACTTTGGGGACTTGTCCTCTGCTGATGCTGTTATGA ACAATGCTAAGGTGAAGGCCCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACG GCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCTCAGCTGAGTGAGCTGC ACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGC TGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGC

	TGAGTTTCAGAAGGTGGTGGCTGGTGTTGCCAATGCCCTGGCCCACAGATATCA CTAA
Opossum	ACCESSION, J03642.1; For simplification, only first 30 bases are listed; ATGGTGCACTTGACTTCTGAGGAGAAGAAC
Gallus	ACCESSION, V00409.1; For simplification, only first 30 bases are listed; ATGGTGCACTGGACTGCTGAGGAGAAGCAG
Lemur	ACCESSION, M15734.1; For simplification, only first 30 bases are listed; ATGACTTTGCTGAGTGCTGAGGAGAATGCT
Mouse	ACCESSION, V00726.1; For simplification, only first 30 bases are listed; ATGGTGCACCTGACTGATGCTGAGAAGTCT
Rabbit	ACCESSION, V00882.1; For simplification, only first 30 bases are listed; ATGGTGCATCTGTCCAGTGAGGAGAAGTCT
Rat	ACCESSION, X06701.1; For simplification, only first 30 bases are listed; ATGGTGCACCTAACTGATGCTGAGAAGGCT
Gorilla	ACCESSION, X61109.1; For simplification, only first 30 bases are listed; ATGGTGCACCTGACTCCTGAGGAGAAGTCT
Bovine	ACCESSION, X00376.1; For simplification, only first 30 bases are listed; ATGCTGACTGCTGAGGAGAAGGCTGCCGTC
Chimpanzee	ACCESSION, X02345.1; For simplification, only first 30 bases are listed; ATGGTGCACCTGACTCCTGAGGAGAAGTCT

Table 2 Similarity/dissimilarity matrix for the complete coding sequences of the β -globin genes among 11 species

Species	*Human*	*Goat*	*Opossum*	*Gallus*	*Lemur*	*Mouse*	*Rabbit*	*Rat*	*Gorilla*	*Bovine*	*Chimpanzee*
Human	0	0.0057	0.0063	0.0203	0.0054	0.0066	0.0061	0.0057	0.00097	0.0053	0.0010
Goat		0	0.0034	0.0269	0.0026	0.0091	0.0053	0.0093	0.00289	0.0027	0.0038
Opossum			0	0.0245	0.0042	0.0101	0.0050	0.0108	0.00530	0.0056	0.0061
Gallus				0	0.0318	0.0095	0.0353	0.0159	0.02443	0.0291	0.0278
Lemur					0	0.0155	0.0026	0.0150	0.00296	0.0045	0.0032
Mouse						0	0.0173	0.0021	0.00783	0.0102	0.0095
Rabbit							0	0.0172	0.00480	0.0055	0.0045
Rat								0	0.00728	0.0095	0.0078
Gorilla									0	0.0035	0.0003
Bovine										0	0.0040
Chimpanzee											0

Then we consider another application for the evolution relationship of HA genes of some influenza A (H1N1) isolates (Table 3). In April 2009, a novel influenza A (H1N1) viruses, A/California/07/2009 (H1N1)-like virus, emerged and showed a strong ability to transmit from human to human and spread worldwide [36]. From March 2009 to April 2010 the new infectious influenza virus have resulted in about 18,000 deaths around the world [37].

Table 3. HA Genes of 16 Influenza A (H1N1) Isolates

Strains	GenBank ID	Time
A/Mexico/4108/2009	GQ223112	2009/04/03
A/Nebraska/06/2009	GQ475885	2009/03/22
A/Michigan/07/2009	GQ476023	2009/04/08
A/Singapore/GP101/2009	CY068676	2009/03/20
A/Boston/95/2009	CY080610	2009/04/01
A/Alabama/WRAIR1235P/2009	CY100844	2009/04/01
A/Mexico/24062/2009	CY147955	2009/03/26
A/California/07/2009	KF009554	2009/04/09
A/Boston/71/2009	CY080929	2009/03/23
A/Florida/07/2009	GQ476125	2009/03/25
A/Mexico/3955/2009	GQ162194	2009/04/02
A/Mexico/4108/2009	GQ162170	2009/04/03
A/California/04/2009	GQ117044	2009/04/01
A/California/08/2009	FJ971076	2009/04/09
A/California/10/2009	FJ969511	2009/04/08

Here, we use the proposed method to analyze the influenza A (H1N1) isolates including new infectious isolates and old isolates. Table 3 shows the randomly selected 16 isolates from March 20 to April 10, 2009. We use the distance computing formula $D(V_a, V_b)$ to obtain all computing distances between any two isolates. Then based on all distances we use the *linkage* and *dendrogram* functions of Matlab to achieve their cluster analysis. The clustering results of the 16 isolates are shown in Figure 3. From Figure 3, we find the 16 H1N1 isolates are classified into two main clusters, Ⅰ and Ⅱ. The cluster Ⅰ is the old isolates. The cluster Ⅱ is the new infectious influenza isolates including A/California/07/2009 (H1N1) (KF009554, 2009/04/09). For comparison, we use MEGA version 6 [38] to conduct phylogenetic and molecular evolutionary analyses of the 16 isolates. The phylogeny tree is shown in Figure 4. From Figure 3 and Figure 4, we find the results are accordance with each other.

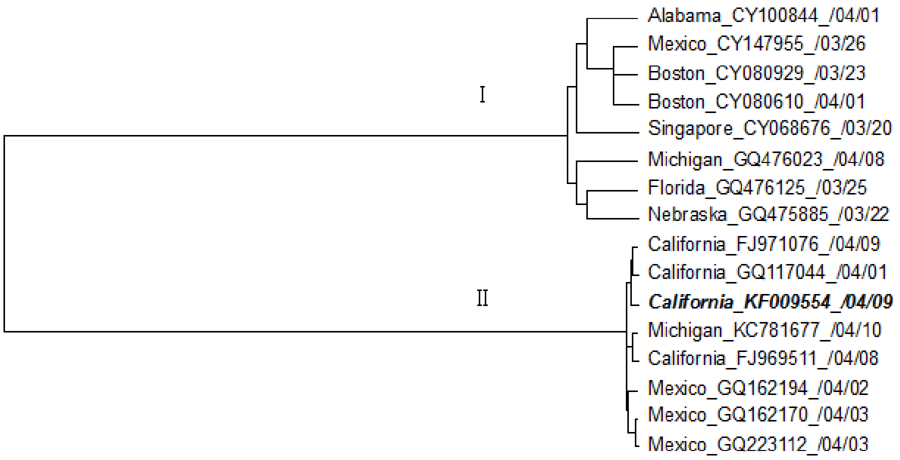


Figure 3. Clustering Results of the 16 Isolates of Table 2 by the Proposed Method

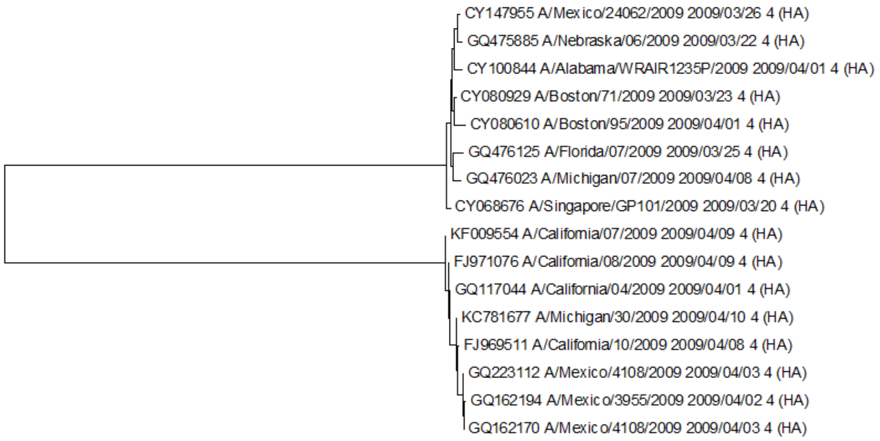

CY147955 A/Mexico/24062/2009 2009/03/26 4 (HA)
GQ475885 A/Nebraska/06/2009 2009/03/22 4 (HA)
CY100844 A/Alabama/WRAIR1235P/2009 2009/04/01 4 (HA)
CY080929 A/Boston/71/2009 2009/03/23 4 (HA)
CY080610 A/Boston/95/2009 2009/04/01 4 (HA)
GQ476125 A/Florida/07/2009 2009/03/25 4 (HA)
GQ476023 A/Michigan/07/2009 2009/04/08 4 (HA)
CY068676 A/Singapore/GP101/2009 2009/03/20 4 (HA)
KF009554 A/Califomia/07/2009 2009/04/09 4 (HA)
FJ971076 A/Califomia/08/2009 2009/04/09 4 (HA)
GQ117044 A/California/04/2009 2009/04/01 4 (HA)
KC781677 A/Michigan/30/2009 2009/04/10 4 (HA)
FJ969511 A/Califomia/10/2009 2009/04/08 4 (HA)
GQ223112 A/Mexico/4108/2009 2009/04/03 4 (HA)
GQ162194 A/Mexico/3955/2009 2009/04/02 4 (HA)
GQ162170 A/Mexico/4108/2009 2009/04/03 4 (HA)

Figure 4. Phylogeny Tree of the 16 Isolates of Table 2 by MEGA Version 6 [38]

4. Conclusions

In this paper, we present a 2D graphic representation of DNA sequence considering codon degeneracy. It allows visual inspection of DNA sequence and help in recognizing main similarities among different DNA sequences. This representation builds the function from DNA sequence to graphic curve. It utilizes the synonyms of codons to reveal numerical characterization of DNA sequences from the views of evolution history and mutation effect. Based on the graphic representation, we construct a 21-dimension vector. Then we use the vector to characterize and compare two sets of DNA sequences, the complete coding sequences of the β -globin genes of 11 species and the HA genes of 16 influenza A (H1N1) isolates. The results show that the proposed graphic method allows visual inspection of data based on the synonyms of codons, helping in recognizing the similarities of mutations among different DNA sequences.

Acknowledgements

This work supported by the National Natural Science Foundation of China (Grant No. 61272254), and by the Natural Science Foundation of Hebei Province, China (Project No. F2012210017), and by the Humanities and Social Sciences Research of Ministry of Education of China (Project name, The Origin, Propagation and Migration of Human Influenza Epidemic (1918-2010) from Space-time Perspective; Project No. 11YJCZH132).

References

[1] A. Cornish-Bowden, “Fundamentals of Enzyme Kinetics”, Chapter 4. Butterworths, London, (**1979**).

[2] K. C. Chou, “A new schematic method in enzyme kinetics”, European Journal of Biochemistry, vol. 113, (**1980**), pp. 195-198.

[3] K. C. Chou and S. Forsen, “Graphical rules for enzyme-catalyzed rate laws”, Biochemical Journal, vol. 187, (**1980**), pp. 829-835.

[4] K. C. Chou and W. M. Liu, “Graphical rules for non-steady state enzyme kinetics”, Journal of Theoretical Biology, vol. 91, (**1981**), pp. 637-654.

[5] D. Myers and G. Palmer, “Microcomputer tools for steady-state enzyme kinetics”, Bioinformatics (original: Computer Applied Bioscience) vol. 1, (**1985**), pp. 105-110.

[6] G. P. Zhou and M. H. Deng, “An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways”, Biochemical Journal, vol. 222, (**1984**), pp. 169-176.

[7] S. X. Lin and K. E. Neet, “Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy”, J. Biol. Chem., vol. 265, (**1990**), pp. 9670-9675.

[8] P. Kuzmic, K. Y. Ng and T. D. Heath, “Mixtures of tight-binding enzyme inhibitors”, Kinetic analysis by a recursive rate equation, Anal. Biochem., vol. 200, (**1992**), pp. 68-73.

[9] J. Andraos, “Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs”, Canadian Journal of Chemistry, vol. 86, (**2008**), pp. 342-357.

[10] K. C. Chou, C. T. Zhang and D. W. Elrod, “Do antisense proteins exist?”, Journal of Protein Chemistry, vol. 15, (**1996**), pp. 59-61.

[11] K. C. Chou, “Review: Applications of graph theory to enzyme kinetics and protein folding kinetics”, Steady and non-steady state systems, Biophysical Chemistry, vol. 35, (**1990**), pp. 1-24.

[12] K. C. Chou, “Graphic rule for non-steady-state enzyme kinetics and protein folding kinetics”, Journal of Mathematical Chemistry, vol. 12, (**1993**), pp. 97-108.

[13] K. C. Chou and C. T. Zhang, “Diagrammatization of codon usage in 339 HIV proteins and its biological implication”, AIDS Research and Human Retroviruses, vol. 8, (**1992**), pp. 1967-1976.

[14] C. T. Zhang and K. C. Chou, “Analysis of codon usage in 1562 E. Coli protein coding sequences”, Journal of Molecular Biology, vol. 238, (**1994**), pp. 1-8.

[15] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen and K. C. Chou, “Using cellular automata to generate Image representation for biological sequences”, Amino Acids, vol. 28, (**2005**), pp. 29-35.

[16] X. Xiao, S. H. Shao, Y. S. Ding, Z. D. Huang and K. C. Chou, “Using cellular automata images and pseudo amino acid composition to predict protein subcellular location”, Amino Acids, vol. 30, (**2006**), pp. 49-54.

[17] X. Xiao, S. H. Shao and K. C. Chou, “A probability cellular automaton model for hepatitis B viral infections”, Biochem. Biophys. Res. Comm., vol. 342, (**2006**), pp. 605-610.

[18] M. Wang, J. S. Yao, Z. D. Huang, Z. J. Xu, G. P. Liu, H. Y. Zhao, X. Y. Wang, J. Yang, Y. S. Zhu and K. C. Chou, “A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis”, Medicinal Chemistry, vol. 1, (**2005**), pp.39-47.

[19] A. Nandy, M. Harle and S. C. Basak, “Mathematical descriptors of DNA sequences: development and applications”, Arkivo, ix, vol. 211, (**2006**), pp. 211-238.

[20] D. Bielinska-Waz, T. Clark, P. Waz, W. Nowak and A. Nandy, “2D-dynamic representation of DNA sequences”, Chem. Phys. Lett., vol. 442, (**2007**), pp. 140-144.

[21] A. Nandy, “A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes”, Curr. Sci., vol. 66, no. 10, (**1994**), pp. 309-314.

[22] D. Bielinska-Waz, P. Waz, and T. Clark, “Similarity Studies of DNA Sequences Using Genetic Methods”, Chem. Phys. Lett., vol. 445, (**2007**), pp. 68-73.

[23] X. Q. Qi, J. Wen and Z. H. Qi, “New 3D graphical representation of DNA sequence based on dual nucleotides”, Journal of Theoretical Biology, vol. 249, (**2007**), pp. 681-690.

[24] Z. H. Qi and X. Q. Qi, “Novel 2D graphical representation of DNA sequence based on dual nucleotides”, Chemical Physics Letters, vol. 440, (**2007**), pp. 139-144.

[25] Z. H. Qi, L. Li and X. Q. Qi, “Using Huffman Coding Method to Visualize and Analyze DNA Sequences”, Journal of Computational Chemistry, vol. 32, (**2011**), pp. 3233-3240.

[26] Z. H. Qi, M. H. Du, X. Q. Qi and L. J. Zheng, “Gene comparison based on the repetition of single-nucleotide structure patterns”, Computers in Biology and Medicine, vol. 42, (**2012**), pp. 975-981.

[27] Y. Yang, Y. Y. Zhang, M. D. Jia, C. Li and L. Y. Meng, “Non-Degenerate Graphical Representation of DNA Sequences and its Applications to Phylogenetic Analysis”, Combinatorial Chemistry & High Throughput Screening, vol. 16, (**2013**), pp. 585-589.

[28] Z. J. Zhang, X. X. Zeng, T. Song, Z. H. Chen, X. Wang and Y. M. Ye, “WormStep: An Improved Compact Graphical Representation of DNA Sequences Based on Worm Curve”, Journal of Computational and Theoretical Nanoscience, vol. 10, (**2013**), pp. 189-193.

[29] Y. H. Yao, Q. Dai, X. Y. Nan, P. A. He, Z. M. Nie, S. P. Zhou and Y. Z. Zhang, “Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation”, Journal of Computational Chemistry, vol. 29, (**2008**), pp. 1632-1639.

[30] A. T. Balaban, D. Plavsic and M. Randić, “DNA invariants based on nonoverlapping triplets of nucleotide bases”, Chemical Physics Letters, vol. 379, (**2003**), pp. 147-154.

[31] J. Wang and Y. Zhang, “Characterization and similarity analysis of DNA sequences based on mutually direct-complementary triplets”, Chemical Physics Letters, vol. 425, (**2006**), pp. 324-328.

[32] F. G. Bai and T. M. Wang, “A 2-D graphical representation of protein sequences based on nucleotide triplet codons”, Chemical Physics Letters, vol. 413, pp. 458-462.

[33] M. Randić and M. Vracko, “On the Similarity of DNA Primary Sequences”, Journal of Chemical Information and Computer Sciences, vol. 40, (**2000**), pp. 599-606.

[34] M. Randić, M. Vracko, A. Nandy and S. C. Basak, “On 3-D graphical representation of DNA primary sequences and their numerical characterization”, Journal of Chemical Information and Computer Sciences, vol. 40, (**2000**), pp. 1235-1244.

[35] Y. H. Wu, A. W-C. Liew, H. Yan and M. S. Yang, “DB-Curve: a novel 2D method of DNA sequence visualization and representation”, Chemical Physics Letters, vol. 367, (**2003**), pp. 170-176.

[36] World Health Organization (WHO), Pandemic (H1N1) - update 64, WHO, Available from: http://www.who.int/csr/don /2009_09_04/en/index.html., (**2009**).

[37] World Health Organization (WHO), Pandemic (H1N1) 2009 - update 97, WHO, Available from: http://www.who.int/csr/don/2010_04_23a/en/index.html., (**2010**).

[38] T. Koichiro, S. Glen, P. Daniel, F. Alan and K. Sudhir, "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0", Molecular Biology and Evolution, vol. 30, (**2013**), pp. 2725-2729.

Authors

Yong-Bin Zhao, he received the MS degrees in cryptology from Xidian University, China. Currently, he is an associate professor in College of Information Science and Technology at Shijiazhuang Tiedao University. His research interests in Bioinformatics and stream cipher

Zhao-Hui Qi, he is a professor in College of Information Science and Technology at Shijiazhuang Tiedao University. He received the MS and the PhD in computer science from Tianjin University, China, in 2003 and 2006. His research interests in Bioinformatics and Pattern Recognition.