Zou et al. EURASIP Journal on Bioinformatics and Systems Biology 2014, 2014:1
http://bsb.eurasipjournals.com/content/2014/1/1

EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

A 2D graphical representation of the sequences of DNA based on triplets and its application

Sai Zou, Lei Wang* and Junfeng Wang

Abstract

In this paper, we first present a new concept of 'weight' for 64 triplets and define a different weight for each kind of triplet. Then, we give a novel 2D graphical representation for DNA sequences, which can transform a DNA sequence into a plot set to facilitate quantitative comparisons of DNA sequences. Thereafter, associating with a newly designed measure of similarity, we introduce a novel approach to make similarities/dissimilarities analysis of DNA sequences. Finally, the applications in similarities/dissimilarities analysis of the complete coding sequences of β-globin genes of 11 species illustrate the utilities of our newly proposed method.

Keywords: Graphical representation; Similarities/dissimilarities analysis; Triplet; DNA sequence

1. Introduction

In the recent years, an exponential growth of sequence data in DNA databases has been observed by biologists; the importance of understanding genetic sequences coupled with the difficulty of working with such immense volumes of DNA sequence data underscores the urgent need for supportive visual tools. Recently, graphical representation is well regarded which can offer visual inspection of data and provide a simple way to facilitate the similarity analysis and comparison of DNA sequences [1-5]. Because of its convenience and excellent maneuverability, currently, all kinds of methods based on graphical representation have been extensively applied in relevant realms of bioinformatics.

Until now, there are many different graphical representation methods having been proposed to numerically characterize DNA sequences on the basis of different multiple-dimension spaces. For example, Liao et al. [6-9], Randic et al. [10-13], Guo et al. [14,15], Qi et al. [16], Dai et al. [17,18], and Dorota et al. [19] proposed different 2D graphical representation methods of DNA sequences, respectively. Liao et al. [20-23], Randic et al. [24,25], Qi et al. [26], Yu et al. [27], and Aram et al. [28] proposed different 3D graphical representation methods of DNA sequences, respectively. Liao et al. [29], Tang et al. [30], and Chi et al. [31] proposed different 4D graphical representation methods of DNA sequences, respectively. In addition, Liao et al. [32] also proposed a kind of 5D representation method of DNA sequences and so on.

In these approaches mentioned above, most of them adopt the leading eigenvalues of some matrices, such as L/L matrices, M/M matrices, E matrices, covariance matrices, and D/D matrices, to weigh the similarities/dissimilarities among the complete coding sequences of β-globin genes of different species. Because the matrix computation is needed to obtain the leading eigenvalues, these methods are usually computationally expensive for long DNA sequences. Furthermore, in some of these approaches, their results of similarities/dissimilarities analysis are not quite reasonable, and there are some results that do not accord with the fact [7,9].

To degrade the computational complexity and obtain more reasonable results of similarities/dissimilarities analysis of DNA sequences, in this article, we propose a new 2D graphical representation of DNA sequences based on triplets, in which, we present a new concept of 'weight' for 64 triplets and a new concept of 'weight deviation' to weigh the similarities/dissimilarities among the complete coding sequences of β-globin genes of different species. Compared with some existing graphical representations of the DNA sequences, our new scheme has the following advantages: (1) no matrix computation is needed, and (2) it can characterize the graphical representations for DNA sequences exactly and obtain reasonable results of similarities/dissimilarities analysis of DNA sequences.

* Correspondence: phd.leiwang@gmail.com
School of Software Engineering, Chongqing College of Electronic Engineering, Chongqing 401331, People's Republic of China

2. Proposed 2D graphical representation of DNA sequence

Codon is a specific sequence of three adjacent nucleotides on the mRNA that specifies the genetic code information for synthesizing a particular amino acid. As illustrated in Table 1, there are total 20 amino acids and 64 codons in the natural world, and each of these codons has a specific meaning in protein synthesis: 64 codons represent amino acids and the other 3 codons cause the termination of protein synthesis.

For the 64 codons illustrated in Table 1, their corresponding triplets of DNA are illustrated in Table 2.

Based on the above 64 triplets of DNA illustrated in Table 2, we define a new mapping Ψ to map each of these triplets into a different weight. Obviously, the mapping Ψ shall satisfy the following rule: for any two pairs of triplets (X_1, Y_1) and (X_2, Y_2), where X_1, Y_1, X_2, and Y_2 are all triplets, if the corresponding codons of X_1 and Y_1 code the same amino acid but the corresponding codons of X_2 and Y_2 code two different amino acids, then there shall be $|\Psi(X_1) - \Psi(Y_1)| < |\Psi(X_2) - \Psi(Y_2)|$. So, according to the above rule and for the sake of convenience, weights consist of amino acid and codon. Amino acid is the integer part of weight, and codon is the fractional part of weight. Alanine is defined as 1, arginine is defined as 2, and the rest can be done in the same manner. Codons of every amino acid are reordered, so the first codon of alanine's (GCT) weight value is 1.1. We design the detailed mapping rules of Ψ as illustrated in Table 3.

Table 1 Relationship between 20 different kinds of most common amino acids and 64 different kinds of mRNA codons

Codons	Amino acid	Codons	Amino acid
GCU, GCC, GCA, GCG	Alanine	CUU, CUC, CUA, CUG, UUA, UUG	Leucine
CGU, CGC, CGA, CGG, AGA, AGG	Arginine	AAA, AAG	Lysine
GAU, GAC	Aspartic acid	AUG	Methionine
AAU, AAC	Asparagine	UUU, UUC	Phenylalanine
UGU, UGC	Cysteine	CCU, CCC, CCA, CCG	Proline
GAA, GAG	Glutamic acid	UCU, UCC, UCA, UCG, AGU, AGC	Serine
CAA, CAG	Glutamine	ACU, ACC, ACA, ACG	Threonine
GGU, GGC, GGA, GGG	Glycine	UGG	Tryptophan
CAU, CAC	Histidine	UAU, UAC	Tyrosine
AUU, AUC, AUA	Isoleucine	GUU, GUC, GUA, GUG	Valine
UAA, UAG, UGA			

Table 2 The corresponding triplets of 64 codons

Codons	Corresponding triplets	Codons	Corresponding triplets
GCU, GCC, GCA, GCG	GCT, GCC, GCA, GCG	CUU, CUC, CUA, CUG, UUA, UUG	CTT, CTC, CTA, CTG, TTA, TTG
CGU, CGC, CGA, 0020CGG, AGA, AGG	CGT, CGC, CGA, CGG, AGA, AGG	AAA, AAG	AAA, AAG
GAU, GAC	GAT, GAC	AUG	ATG
AAU, AAC	AAT, AAC	UUU, UUC	TTT, TTC
UGU, UGC	TGT, TGC	CCU, CCC, CCA, CCG	CCT, CCC, CCA, CCG
GAA, GAG	GAA, GAG	UCU, UCC, UCA, UCG, AGU, AGC	TCT, TCC, TCA, TCG, AGT, AGC
CAA, CAG	CAA, CAG	ACU, ACC, ACA, ACG	ACT, ACC, ACA, ACG
GGU, GGC, GGA, GGG	GGT, GGC, GGA, GGG	UGG	TGG
CAU, CAC	CAT, CAC	UAU, UAC	TAT, TAC
AUU, AUC, AUA	ATT, ATC, ATA	GUU, GUC, GUA, GUG	GTT, GTC, GTA, GTG
UAA, UAG, UGA	TAA, TAG, TGA		

For example, from Table 3, we will have Ψ (GCT) = 1.1, Ψ (GCC) = 1.2, Ψ (ATG) = 20.1, etc., and in addition, we can propose a novel 2D graphical representation of DNA sequences as follows:

Let $G = g_1, g_2, g_3 \ldots g_N$ be an arbitrary DNA primary sequence, where $g_i \in \{A, T, G, C\}$ for any $i \in \{1, 2, \ldots, N\}$, and then, we can transform G into a sequence of triplets such as $G = t_1, t_2, t_3 \ldots t_M$, where $M = [N/3]$ and t_i is a triplet of DNA for any $i \in \{1, 2, \ldots, M\}$. Thereafter, we can define a new mapping Θ to map G into a plot set as illustrated in the formula (1).

$$\Theta(G) = \{(1, \Psi(t_1)), (2, \Psi(t_2)), \ldots, (M, \Psi(t_M))\} \quad (1)$$

As for the complete coding sequences of β-globin genes of 11 species illustrated in the Table 4, each of them can be mapped into a plot set by using the new given mapping Θ, and the 2D graphical representations corresponding to the complete coding sequences of β-globin genes of human, chimpanzee, and opossum are shown in Figures 1, 2, and 3, respectively.

3. Similarity analysis of DNA sequence

Let $G = g_1, g_2, g_3 \ldots g_N$ be an arbitrary complete coding sequence, where $g_i \in \{A, T, G, C\}$ for any $i \in \{1, 2, \ldots, N\}$, and $G = t_1, t_2, t_3 \ldots t_M$ be its corresponding sequence of triplets, where $M = [N/3]$ and t_i is a triplet of DNA for any $i \in \{1, 2, \ldots, M\}$. Then, we define a function δ and let $\delta(t_i)$ represent the total number of times that the triplet

Table 3 The mapping rules of Ψ

Triplet	Corresponding weight	Triplet	Corresponding weight
GCT	1.1	CTT	11.1
GCC	1.2	CTC	11.2
GCA	1.3	CTA	11.3
GCG	1.4	CTG	11.4
		TTA	11.5
		TTG	11.6
CGT	2.1	AAA	12.3
CGC	2.2	AAG	12.4
CGA	2.3		
CGG	2.4		
AGA	2.5		
AGG	2.6		
GAT	3.3	TTT	13.1
GAC	3.4	TTC	13.2
AAT	4.1	CCT	14.1
AAC	4.2	CCC	14.2
		CCA	14.3
		CCG	14.4
TGT	5.1	TCT	15.1
TGC	5.2	TCC	15.2
		TCA	15.3
		TCG	15.4
		AGT	15.5
		AGC	15.6
GAA	6.1	ACT	16.3
GAG	6.2	ACC	16.4
		ACA	16.5
		ACG	16.6
CAA	7.1	TGG	17.3
CAG	7.2		
GGT	8.1	TAT	18.1
GGC	8.2	TAC	18.2
GGA	8.3		
GGG	8.4		
CAT	9.1	GTT	19.1
CAC	9.2	GTC	19.2
		GTA	19.3
		GTG	19.4
ATT	10.1	ATG	20.1
ATC	10.2		
ATA	10.3		
TAA	21.1		
TAG	21.2		
TGA	21.3		

t_i repeats in the sequence of triplets $G = t_1, t_2, t_3 \ldots t_M$ for any $i \in \{1, 2, \ldots, M\}$.

Let T_1 = GCT, T_2 = GCC, T_3 = GCA, T_4 = GCG, T_5 = CGT, T_6 = CGC, T_7 = CGA, T_8 = CGG, T_9 = AGA, T_{10} = AGG, T_{11} = GAT, T_{12} = GAC, T_{13} = AAT, T_{14} = AAC, T_{15} = TGT, T_{16} = TGC, T_{17} = GAA, T_{18} = GAG, T_{19} = CAA, T_{20} = CAG, T_{21} = GGT, T_{22} = GGC, T_{23} = GGA, T_{24} = GGG, T_{25} = CAT, T_{26} = CAC, T_{27} = ATT, T_{28} = ATC, T_{29} = ATA, T_{30} = CTT T_{31} = CTC, T_{32} = CTA, T_{33} = CTG, T_{34} = TTA, T_{35} = TTG, T_{36} = AAA, T_{37} = AAG, T_{38} = TTT, T_{39} = TTC, T_{40} = CCT, T_{41} = CCC, T_{42} = CCA, T_{43} = CCG, T_{44} = TCT, T_{45} = TCC, T_{46} = TCA, T_{47} = TCG, T_{48} = AGT, T_{49} = AGC, T_{50} = ACT, T_{51} = ACC, T_{52} = ACA, T_{53} = ACG, T_{54} = TGG, T_{55} = TAT, T_{56} = TAC, T_{57} = GTT, T_{58} = GTC, T_{59} = GTA, T_{60} = GTG, T_{61} = ATG, T_{62} = TAA, T_{63} = TAG, and T_{64} = TGA.

Thereafter, according to Table 2, since there are a total of 64 triplets of DNA, then we can construct a set of 64 vectors $\{<T_1, \delta(T_1)>, <T_2, \delta(T_2)>, \ldots, <T_{64}, \delta(T_{64})>\}$ for the given sequence of triplets $G = t_1, t_2, t_3 \ldots t_M$ as follows: if $T_i = t_j \in \{t_1, t_2, t_3, \ldots t_M\}$, then $\delta(T_i) = \delta(t_j)$, else $\delta(T_i) = 0$, for any $i \in \{1, 2, \ldots, 64\}$ and $j \in \{1, 2, \ldots, M\}$.

For convenience, we call $\{<T_1, \delta(T_1)>, <T_2, \delta(T_2)>, \ldots, <T_{64}, \delta(T_{64})>\}$ as the triplet-repeat model set of G.

For any two given complete coding sequences A and B, suppose that their triplet-repeat model sets are $\{<T_1, X_1>, <T_2, X_2>, \ldots, <T_{64}, X_{64}>\}$ and $\{<T_1, Y_1>, <T_2, Y_2>, \ldots, <T_{64}, Y_{64}>\}$, respectively. Then, on the basis of the 2D graphical representation given in the previous Section 2, we can define the weight deviation between the two DNA sequences A and B as the following formula (2) to measure the similarity between A and B.

$$\mathrm{WD}(A,B) = \frac{\sum_{i=1}^{64} |X_i - Y_i| * \Psi(T_i)}{64} \qquad (2)$$

Obviously, the above formula (2) satisfies the fact that the smaller the weight deviation between the two DNA sequences A and B, the higher the degree of similarity of A and B. According to formula (2), the detailed similarity/dissimilarity matrix obtained for the coding sequences listed in Table 4 is illustrated in Table 5. Basing on the similarity matrix (Table 5) constructs a phylogenetic tree, which is shown in Figure 4.

Observing Table 5, it is easy to find out that human, gorilla, and chimpanzee are most similar to each other, and the pairs like gorilla-chimpanzee (with weight deviation of 1.1266), human-gorilla (with weight deviation of 4.3359), and human-chimpanzee (with weight deviation of 5.2500) are the most similar species pairs, but *Gallus* and opossum are the most dissimilar to the others (with weight deviation bigger than 11). It is consistent with the fact that *Gallus* is not a mammal, whereas the others

Table 4 The complete coding sequences of β-globin genes of 11 species

Species	Complete coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAG TCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACT GAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAA GTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGCTGCTGGTGGTCTACCCTTGGACCCAG AGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCAC CTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGT CCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTG AGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAATGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAG
Black lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAGGCTGCTGGTCGTCTACCCATGGACCCAGAGGTTCTTCGAGT CCTTTGGGGACCTGTCCTCTCCTTCTGCTGTTATGGGGAACCCTAAGGTGAAGGCCCATGGCAAGAAGGTGCTGAGTGCCTTTAGTGAAGGTCTGCATCACCTGGACAACCTCAAGGGCACCTTTGCTCAACTG AGTGAGCTGCACTGTGACAAGTTGCACGTGGATCCTCAGAACTTCACTCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTGAACACTTTGGCAATGCATTCAGCCCGGCGGTGCAGGCTGCCTTTCAGAAGG TGGTGGCTGGTGTGGCCAATGCTCTGGCTCACAAGTACCACTGA
Norway rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAATGCTGATAATGTTGGCGCTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCTTGGACCCAGAGGTACTTTTCTA AATTTGGGGACCTGTCCTCTGCCTCTGCTATCATGGGTAACCCCCAGGTGAAGGCCCATGGCAAGAAGGTGATAAATGCCTTCAATGATGGCCTGAAACACTTGGACAACCTCAAGGGCACCTTTGCTCATCT GAGTGAACTCCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGCTCCTGGGCAATATGATTGTGATTGTGTTGGGCCACCACCTGGGCAAGGAATTCACCCCCTGTGCACAGGCTGCCTTCCAGAA GGTGGTGGCTGGAGTGGCCAGTGCCCTGGCTCACAAGTACCACTAA
House mouse	ATGGTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCTTGGACCCAGCGGTACTTTGATAG CTTTGGAGACCTATCCTCTGCCTCTGCTATCATGGGTAATCCCAAGGTGAAGGCCCATGGCAAAAAGGTGATAACTGCCTTTAACGAGGGCCTGAAAAACCTGGACAACCTCAAGGGCACCTTTGCCAGCCTC AGTGAGCTCCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGCTCCTAGGCAATGCGATCGTGATTGTGCTGGGCCACCACCTGGGCAAGGATTTCACCCCTGCTGCACAGGCTGCCTTCCAGAAGG TGGTGGCTGGAGTGGCCACTGCCCTGGCTCACAAGTACCACTAA
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCCTGGACTCAGAGGTTCTTTGAGCACTTT GGGGACTTGTCCTCTGCTGATGCTGTTATGAACAATGCTAAGGTGAAGGCCCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCTCAGCTGAGT GAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTG GTGGCTGGTGTTGCCAATGCCCTGGCCCACAGATATCACTAA
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCCTGGACTCAGAGGTTCTTTGAGTCCTTTG GGGACTTGTCCACTGCTGATGCTGTTATGAACAACCCTAAGGTGAAGGCCCATGGCAAGAAGGTGCTAGATTCCTTTAGTAATGGCATGAAGCATCTCGATGACCTCAAGGGCACCTTTGCTGCGCTGAGTGA GCTGCACTGTGATAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTAGTGGTTGTGCTGGCTCGCAATTTTGGCAAGGAATTCACCCCGGTGCTGCAGGCTGACTTTCAGAAGGTGGTG GCTGGTGTGGCCAATGCCCTGGCCCACAGATATCATTAA
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTTGTCTACCCATGGACCCAGAGGTTCTTCGAG TCCTTTGGGGACCTGTCCTCTGCAAATGCTGTTATGAACAATCCTAAGGTGAAGGCTCATGGCAAGAAGGTGCTGGCTGCCTTCAGTGAGGGTCTGAGTCACCTGGACAACCTCAAAGGCACCTTTGCTAAGCT GAGTGAACTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTCACTCCTCAGGTGCAGGCTGCCTATCAGAAG GTGGTGGCTGGTGTGGCCAATGCCCTGGCTCACAAATACCACTGA
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAGGATGCTCGTTGTCTACCCCTGGACCACCAGGTTTTTTGGGA GCTTTGGTGATCTGTCCTCTCCTGGCGCTGTCATGTCAAATTCTAAGGTTCAAGCCCATGGTGCTAAGGTGTTGACCTCCTTCGGTGAAGCAGTCAAGCATTTGGACAACCTGAAGGGTACTTATGCCAAGTTG AGTGAGCTCCACTGTGACAAGCTGCATGTGGACCCTGAGAACTTCAAGATGCTGGGGAATATCATTGTGATCTGCCTGGCTGAGCACTTTGGCAAGGATTTTACTCCTGAATGTCAGGTTGCTTGGCAGAAGCT CGTGGCTGGAGTTGCCCATGCCCTGGCCCACAAGTACCACTAA
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAGGCTGCTGATCGTCTACCCCTGGACCCAGAGGTTCTTTGCG TCCTTTGGGAACCTCTCCAGCCCCACTGCCATCCTTGGCAACCCCATGGTCCGCGCCCACGGCAAGAAAGTGCTCACCTCCTTTGGGGATGCTGTGAAGAACCTGGACAACATCAAGAACACCTTCTCCCAAC TGTCCGAACTGCATTGTGACAAGCTGCATGTGGACCCCGAGAACTTCAGGCTCCTGGGTGACATCCTCATCATTGTCCTGGCCGCCCACTTCAGCAAGGACTTCACTCCTGAATGCCAGGCTGCCTGGCAGAA GCTGGTCCGCGTGGTGGCCCATGCCCTGGCTCGCAAGTACCACTAA

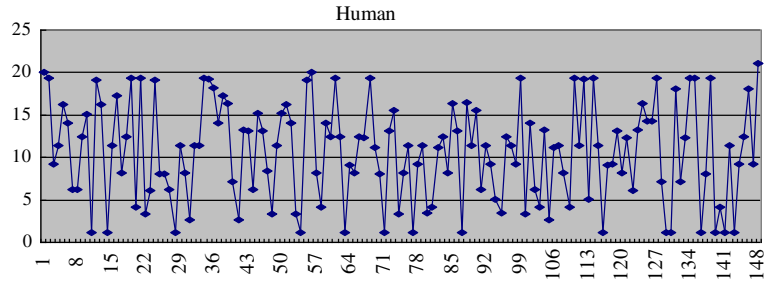

Figure 1 The 2D graphical representations of the complete coding sequences of β-globin genes of human.

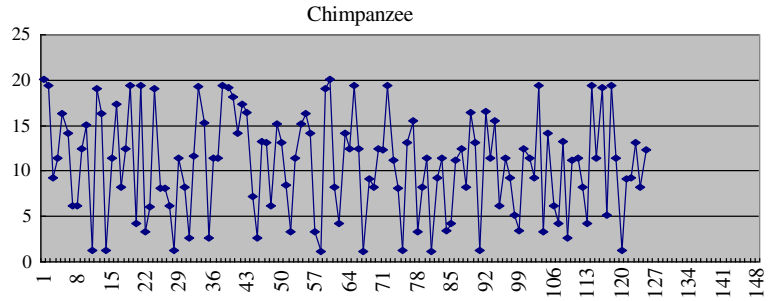

Figure 2 The 2D graphical representations of the complete coding sequences of β-globin genes of chimpanzee.

are mammals, and opossum is the most remote species from the remaining mammals. Similar results have been obtained in other papers by different approaches [2,5,7,9,33].

For testing the validity of our method, the existing results of the examination of the degree of similarity/dissimilarity of the coding sequences of β-globin genes of several species with the coding sequence of the human β-globin gene by means of approaches using alternative DNA sequence descriptors [2,5,7,9] are listed in Table 6 for comparison.

From Table 6, we can find that the pairs like human-gorilla and human-chimpanzee are the two most similar species pairs when adopting (A) the method of our work, (B) the method of [2], (C) the method of [5], and (D) the method of [7], which is in accordance with the fact that gorilla and chimpanzee are the two most closest species of human, but when adopting (E) the method of [9], the most similar species pair is human-goat, which is obviously not correct. In addition, the pairs like human-*Gallus* and human-opossum are the two most dissimilar species pairs when adopting (A) the method of our work, (C) the method of [5], and (E) the method of [9], which is in accordance with the fact that *Gallus* is not a mammal, whereas the others are mammals, and opossum is the most remote species from the remaining mammals. However, when adopting (D) the method of [7], the two most dissimilar species pairs are human-opossum and human-lemur, which is obviously not reasonable also.

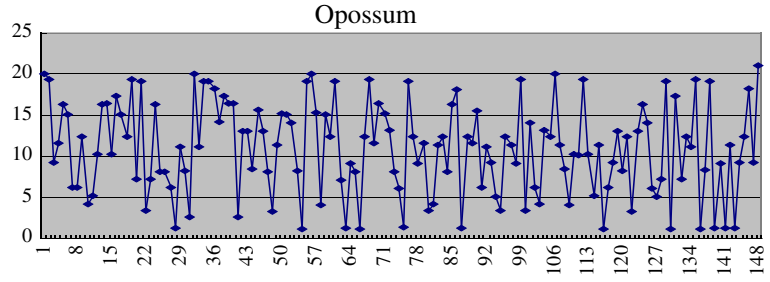

Figure 3 The 2D graphical representations of the complete coding sequences of β-globin genes of opossum.

Table 5 The similarity/dissimilarity matrix for the coding sequences of Table 1 based on the weight deviation

	Human	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	*Gallus*
Human	0	5.2500	4.3359	8.5891	10.670	9.7047	8.2219	8.1438	7.8281	15.6078	16.7109
Chimpanzee		0	1.1266	8.0297	10.645	9.6016	8.4375	9.3219	9.6000	14.2578	15.8734
Gorilla			0	7.8688	9.9625	8.6063	7.6734	8.5578	8.5547	13.9719	14.8781
Lemur				0	8.7219	9.5500	7.1328	9.3891	5.6891	12.9281	15.2000
Rat					0	6.0750	7.0484	9.3641	9.6578	13.5906	14.1219
Mouse						0	9.4953	9.2641	10.7984	12.3406	12.3688
Goat							0	5.2625	8.7219	11.9703	14.5359
Bovine								0	9.2906	12.5922	15.0234
Rabbit									0	14.8984	15.6953
Opossum										0	14.2750
Gallus											0

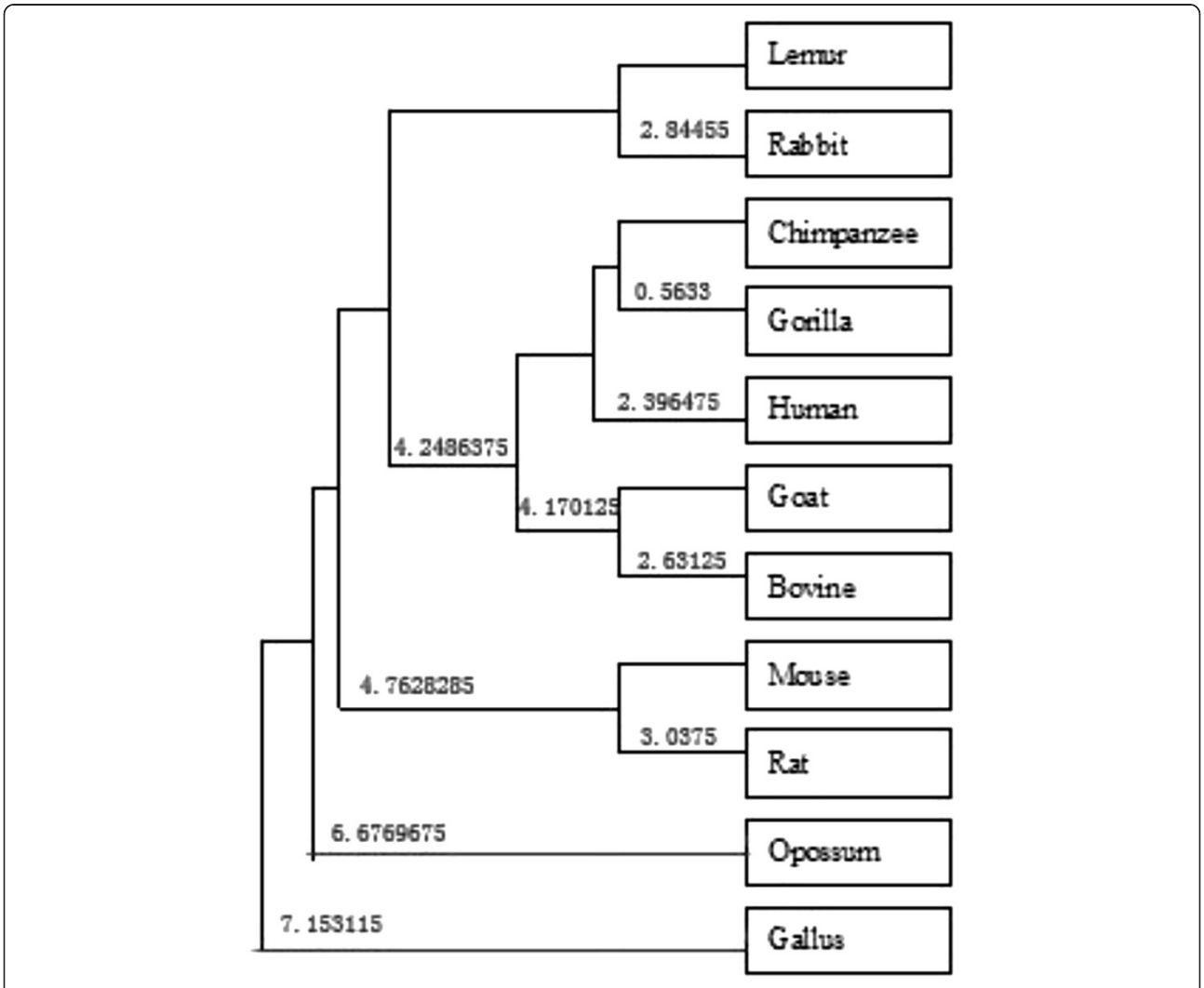

Figure 4 Phylogenetic tree based on the similarity matrix (Table 5).

Table 6 The similarity/dissimilarity of the coding sequences

Species	A	B	C	D	E
Chimpanzee	5.2500	0.0144	14.00	0.005069	0.863
Gorilla	4.3359	0.0125	13.63	0.006611	0.339
Lemur	8.5891	-	31.75	0.030894	1.188
Rat	10.670	0.1377	41.65	0.015539	1.966
Mouse	9.7047	0.1427	30.27	0.015700	0.735
Goat	8.2219	0.1161	31.39	0.020980	0.311
Bovine	8.1438	0.0773	30.68	0.017700	2.489
Rabbit	7.8281	0.1332	35.575	0.015788	1.372
Opossum	15.6078	-	48.701	0.033363	6.322
Gallus	16.7109	-	70.46	0.025801	7.170

4. Conclusion

In this paper, we propose a new 2D graphical representation for DNA sequences based on triplets, and associating with a newly introduced concept of weight of triplets and a newly designed measure of similarity named weight deviation, we propose a new method to make similarity analysis of DNA sequences, in which no matrix computation is needed and reasonable and useful approaches for both computational scientists and molecular biologists to effectively analyze DNA sequences can be provided at the same time.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by the Chongqing Education Science Project of China in 2014, Chongqing "Twelfth Five Year plan" educational programming projects of China (2013-ZJ-077), program for university youth backbone teachers of Chongqing in 2014.

Received: 17 August 2013 Accepted: 10 December 2013
Published: 2 January 2014

References

1. W Chen, B Liao, Y Liu, W Zhu, Z Su, A numerical representation of DNA sequences and its applications. MATCH: Commun Math Comput Chem. **60**, 291–300 (2008)
2. N Jafarzadeh, A Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons. MATCH: Commun Math Comput Chem. **68**, 611–620 (2012)
3. B Liao, BY Liao, XM Sun, QG Zeng, A novel method for similarity analysis and protein sub-cellular localization prediction. Bioinformatics **26**, 2678–2683 (2010)
4. XQ Qi, Q Wu, Y Zhang, E Fuller, CQ Zhang, A novel model for DNA sequence similarity analysis based on graph theory. J Evol Bioinform **7**, 149–158 (2011)
5. JF Yu, JH Wang, X Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. MATCH: Commun Math Comput Chem. **63**, 493–512 (2010)
6. Y Li, G Huang, B Liao, Z Liu, H-L curve: a novel 2D graphical representation of protein sequences. MATCH: Commun Math Comput Chem. **61**, 519–532 (2009)
7. B Liao, TM Wang, New 2D graphical representation of DNA sequences. J. Comput. Chem. **25**, 1364–1368 (2004)
8. B Liao, XY Xiang, W Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence. J. Comput. Chem. **27**, 1196–1202 (2006)
9. ZB Liu, B Liao, W Zhu, GH Huang, A 2D graphical representation of DNA sequence based on dual nucleotides and its application. Int. J. Quantum Chem. **109**, 948–958 (2009)
10. M Randic, M Vracko, J Zupan, M Novic, Compact 2D graphical representation of DNA. Chem. Phys. Lett. **373**, 558–562 (2003)
11. M Randic, M Vracko, N Lers, D Plavsic, Analysis of similarity/dissimilarity of 2D graphical representation. Chem. Phys. Lett. **371**, 202–207 (2003)
12. M Randic, M Vracko, N Lers, D Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. **368**, 1–6 (2003)
13. M Randic, Graphical representations of DNA as 2-D map. Chem. Phys. Lett. **386**, 468–471 (2004)
14. XF Guo, M Randic, SC Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem. Phys. Lett. **350**, 106–112 (2001)
15. XF Guo, A Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy. Chem. Phys. Lett. **369**, 361–366 (2003)
16. ZH Qi, XQ Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chem Phys Lett. **440**, 139–144 (2007)
17. Q Dai, ZL Xiu, TM Wang, A novel 2D graphical representation of DNA sequences and its application. J Mol Graph Model. **25**, 340–344 (2006)
18. XQ Liu, Q Dai, ZL Xiu, TM Wang, PNN–curve: a new 2D graphical representation of DNA sequences and its application. J. Theor. Biol. **243**, 555–561 (2006)
19. BW Dorota, C Timothy, W Piotr, 2D-dynamic representation of DNA sequences. Chem. Phys. Lett. **442**, 140–144 (2007)
20. CX Yuan, B Liao, TM Wang, New 3D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. **379**, 412–417 (2003)
21. B Liao, TM Wang, 3-D graphical representation of DNA sequences and their numerical characterization. J. Mol. Struct. (THEOCHEM) **681**, 209–212 (2004)
22. B Liao, TM Wang, A 3D graphical representation of RNA secondary structure. J Biomol Struct Dynam. **21**, 827–832 (2004)
23. Z Cao, B Liao, RF Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides. Int. J. Quantum Chem. **108**, 1485–1490 (2008)
24. M Randic, M Vracko, A Nandy, SC Basak, On 3D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci. **40**, 1235–1244 (2000)
25. M Randic, J Zupan, M Novic, On 3D graphical representation of proteomics maps and their numerical characterization. J. Chem. Inf. Comput. Sci. **41**, 1339–1344 (2001)
26. XQ Qi, TR Fan, PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. **442**, 434–440 (2007)
27. JF Yu, X Sun, JH Wang, TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J. Theor. Biol. **261**, 459–468 (2009)
28. V Aram, A Iranmanesh, 3D-dynamic representation of DNA sequences. MATCH: Commun Math Comput Chem. **67**, 809–816 (2012)
29. B Liao, MS Tan, KQ Ding, A 4D representation of DNA sequences and its application. Chem. Phys. Lett. **402**, 380–383 (2005)
30. XC Tang, PP Zhou, WY Qiu, On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chin. Sci. Bull. **55**, 701–704 (2010)
31. R Chi, KQ Ding, Novel 4D numerical representation of DNA sequences. Chem. Phys. Lett. **407**, 63–67 (2005)
32. B Liao, XY Xiang, RF Li, W Zhu, On the similarity of DNA primary sequences based on 5D representation. J. Math. Chem. **42**, 47–57 (2007)
33. P He, J Wang, Characteristic sequences for DNA primary sequence. J. Chem. Inf. Comput. Sci. **42**, 1080–1085 (2002)

doi:10.1186/1687-4153-2014-1
Cite this article as: Zou *et al.*: **A 2D graphical representation of the sequences of DNA based on triplets and its application.** *EURASIP Journal on Bioinformatics and Systems Biology* 2014 **2014**:1.