Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Similarity studies of DNA sequences based on a new 2D graphical representation

Guohua Huang [a,b,*], Bo Liao [a], Yongfan Li [c], Yougui Yu [d]

[a] School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China
[b] Department of Mathematics, Shaoyang University, Shaoyang, Hunan 422000, China
[c] Hunan First Normal College, Changsha, Hunan 410002, China
[d] Department of Biology and Chemistry Engineering, Shaoyang University, Hunan 42200, China

ARTICLE INFO

Article history:
Received 22 November 2008
Received in revised form 25 March 2009
Accepted 26 March 2009
Available online 8 April 2009

Keywords:
DNA sequences
Graphical representation
Sequences comparison
Quantitative measure

ABSTRACT

We denoted the four nucleotides, A, T, G and C, as four two-component vectors, and illustrated a nucleotide sequence as a curve in the 2D space by concatenating the vectors representing the nucleotides in the sequence. We studied the similarities among multiple nucleotide sequences by comparing their corresponding curves, with the beta-globin genes from 7 species as an example.

1. Introduction

Huge quantities of DNA sequences are available with development of current sequencing techniques. How to deal with the large volume genomic DNA sequences is one of challenges with which bio-scientists are presently confronted. Graphical representation provides a simple way of viewing, sorting and comparing various gene sequences with their intuitive pictures and pattern. Since Hamori [1] first proposed a 3D graphical representation for DNA sequences, some different graphical approaches representing DNA sequences have been reported by several authors. The original plot of a DNA sequence as a random walk in a 2D-space using four orthogonal directions to represent the four bases has been presented by Gates [2] and then has been rediscovered independently by Nandy [3] and Leong and Morgenthaler [4]. The idea is to read a DNA sequence base by base and plot succeeding points on the graph with four orthogonal unit vectors representing four kinds of bases. According to the prescription by Nandy [3], a point was plotted by moving one step in the negative *x*-direction if the base was an adenine (A) and in the opposite direction if it was a guanine (G) and a walk of one step in the positive *y*-direction if the base was a cytosine (C) and in the opposite direction if the base was a thymine (T). However, the graphical representations based on the rectangular walk often overlap with themselves and raise degeneracy. To eliminate the degeneracy, Guo et al. [5] have designed four vectors that are at a small angle to the four axial directions to represent four kinds of bases. This representation has reduced degeneracy, but has not completely avoided the degeneracy. Later, Yau et al. [6] have denoted four vectors within the first quadrant and the fourth quadrant in the two dimension Cartesian coordinate system, e.g., assigned T and C to the first quadrant and A and G to the fourth quadrant. The purine–pyrimidine graph by Yau is associated with a DNA sequence in a one-to-one manner and completely meets requirement of non-degeneracy. More representations involving random walk have been reviewed in a review by Nandy et al. [7]. However, these vectors designed are equal in length and their modules generally are one unit. Hence, the graph doesn't benefit comparing the local difference among various DNA sequences.

Chao Game Representation (CGR) is an important and scale-independent representation for genomic sequences by H.J. Jeffrey [8] in 1990, which has been widely used for visual representation of genome sequences patterns as well as alignment-free comparison of sequences based on oligonucleotide frequencies [9]. A CGR of a DNA sequence is plotted in a unit square, the vertices of which are labelled by the nucleotide A-(0,0), C-(0,−1), G-(1,1), T-(1,0). The plotting procedure can be described by the following steps: the first base of sequence is plotted halfway between the center of the square and the vertex, successive bases in the sequence are plotted halfway between the previous plotted point and the vertex representing the nucleotides being plotted. Since CGR reveals some interesting feature relevant to the DNA sequences organization, it has received widespread and further research. Some examples are recommended in the Refs. [9–16].

The representation based on dinucleotide has been proposed as an important part of graphical techniques by several authors. For example, Randic [17] has proposed a condense representation of DNA sequences based on pairs of nucleotide which allows quantitative comparisons among various DNA sequences. Recently, Liu [18], Qi

* Corresponding author. Present address: School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China. Fax: +86 731 8821715.
E-mail address: guohuahhn@163.com (G. Huang).

0301-4622/$ – see front matter. Crown Copyright © 2009 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2009.03.013

[19,20] and Qi [21] et al. have presented their DNA sequence analysis schemes in terms of the novel graphical representation based on neighboring bases of DNA sequence. These representations have revealed some biological information hidden between nucleotides that other methods have not revealed. Besides, some illustrations of other representations for DNA sequences and their application should be recommended in the Refs. [22–26].

In this paper, we designed four novel two-component vectors representing four kinds of nucleotides, where the first elements are constant (equal to 1) and the second elements are different from each other and proposed a new and universal 2D graphical representation of DNA sequences. On the basis of the presented representation, we outlined a new measure of similarity and dissimilarity among various DNA sequences. Finally, we illustrated the use of the measure with the examination of similarities and dissimilarities among the complete coding sequences of beta-globin gene of different 7 species.

2. Data and theory

2.1. Data

For sake of clarity, this report considered the complete coding sequence of beta-globin genes from 7 different species, which are very relatively conservative and were studied by Qi et al. [21]. They are respectively Human [GenBank: U01317], Opossum[GenBank: J03643], Gallus [GenBank: V00409], Lemur [GenBank: M15734], Mouse [GenBank: V00722], Rabbit [GenBank: V00882] and Rat [GenBank: X06701]. Their detailed information was listed in Table 1.

2.2. A new 2D graphical representation of DNA sequences

As shown in Fig. 1, we denoted four novel two-component vectors representing four kinds of nucleotide A, G, T, and C. These vectors are equal in their first elements and angles between them and the x-axis are respectively alpha, beta, gamma and rho. That is to say, the angle between the vector $\overrightarrow{v}_T$ for base T and the x-axis is alpha, between the vector $\overrightarrow{v}_C$ for base C and the x-axis is beta, between the vector $\overrightarrow{v}_G$ for base G and the x-axis is gamma, and between the vector $\overrightarrow{v}_A$ for base A and the x-axis is rho. In order to construct purine–pyrimidine graph, We defined these angles as follows: $0 \prec \alpha \prec \frac{\pi}{2}, 0 \prec \beta \prec \frac{\pi}{2}, -\frac{\pi}{2} \prec \gamma \prec 0$ and $-\frac{\pi}{2} \prec \rho \prec 0$. To avoid the degeneracy, $\alpha \neq \beta$ and $\gamma \neq \rho$. If $\alpha = -\gamma$ and $\beta = -\rho$, the assignment for four kinds of base to vectors is identical with that in the Ref. [22]. Obviously, the generalized model is the development of the previous work [22]. Moreover, the graphical representations are constructed more changeable, based on the new assignment than the previous assignment. According to the new assignment, we reduced a DNA sequence considered into a set of vectors that are called as characteristic vectors. For example, a DNA sequence is TACTGACTGCAG, and then the corresponding set of characteristic vectors is $\{\overrightarrow{v}_T \overrightarrow{v}_A \overrightarrow{v}_C \overrightarrow{v}_T \overrightarrow{v}_G \overrightarrow{v}_A \overrightarrow{v}_C \overrightarrow{v}_T \overrightarrow{v}_G \overrightarrow{v}_C \overrightarrow{v}_A \overrightarrow{v}_G\}$. These vectors orderly were connected head and tail using walk strategy introduced in the Refs. [2–7], and then put in shape a curve shown in

Table 1
The complete coding sequences of β-globin genes of seven species.

Species	Complete coding sequence
Human	ACCESSION U01317; REGION: join(62187 … 62278; 62409 … 62631; 63482 … 63610) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTG CTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGG CAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCT GCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCT GCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
Opossum	ACCESSION J03643; REGION: join(467 … 558; 672 … 894; 2360 … 2488) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAGACAC AGCAAACACAACGACCCATATAGACATTGATGTGAAATTGTCTATTGTCAATTTATGGGAAAACAAGTATGTACTTTTTCTACTAAGCCATTGAAACAG GAATAACAGAACAAGATTGAAAGAATACATTTTCCGAAATTACTTGAGTATTATACAAAGACAAGCACGTGGACCTGGGAGGAGGGTTATTGTCCAT GACTGGTGTGTGGAGACAAATGCTGTTTGCTAGTATTTTTTGTTTAACTGCAATCATTCTTGCTGCAGGTGAAAACTAGTGTTCTGTACTTTATGCCCA TTCATCTTTAACTGTAATAATAAAAATAACTGACATTTATTGAAGGCTATCAG
Gallus	ACCESSION V00409; REGION: join(465..556,649..871,1682..1810) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAGGCTG CTGATCGTCTACCCCTGGACCCAGAGGTTCTTTGCGTCCTTTGGGAACCTCTCCAGCCCCACTGCCATCCTTGGCAACCCCATGGTCCGCGCCCACGG CAAGAAAGTGCTCACCTCCTTTGGGGATGCTGTGAAGAACCTGGACAACATCAAGAACACCTTCTCCCAACTGTCCGAACTGCATTGTGACAAGCT GCATGTGGACCCCGAGAACTTCAGGCTCCTGGGTGACATCCTCATCATTGTCCTGGCCGCCCACTTCAGCAAGGACTTCACTCCTGAATGCCAGGCT GCCTGGCAGAAGCTGGTCCGCGTGGTGGCCCATGCCCTGGCTCGCAAGTACCACTAA
Lemur	ACCESSION M15734; REGION: join(154 … 245; 376 … 598; 1467 … 1595) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAGGCTG CTGGTCGTCTACCCATGGACCCAGAGGTTCTTCGAGTCCTTTGGGGACCTGTCCTCTCCTTCTGCTGTTATGGGGAACCCTAAGGTGAAGGCCCATGG CAAGAAGGTGCTGAGTGCCTTTAGTGAAGGTCTGCATCACCTGGACAACCTCAAGGGCACCTTTGCTCAACTGAGTGAGCTGCACTGTGACAAGTT GCACGTGGATCCTCAGAACTTCACTCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTGAACACTTTGGCAATGCATTCAGCCCGGCGGTGCAGGCT GCCTTTCAGAAGGTGGTGGCTGGTGTGGCCAATGCTCTGGCTCACAAGTACCACTGA
Mouse	ACCESSION V00722; REGION: join(275 … 367; 484 … 705; 1334 … 1462) Exon1 1 … 93; Exon2 94 … 315; Exon3 316 … 444; ATGGTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTG CTGGTTGTCTACCCTTGGACCCAGCGGTACTTTGATAGCTTTGGAGACCTATCCTCTGCCTCTGCTATCATGGGTAATCCCAAGGTGAAGGCCCATGGC AAAAAGGTGATAACTGCCTTTAACGAGGGCCTGAAAAACCTGGACAACCTCAAGGGCACCTTTGCCAGCCTCAGTGAGCTCCACTGTGACAAGCTG CATGTGGATCCTGAGAACTTCAGGCTCCTAGGCAATGCGATCGTGATTGTGCTGGGCCACCACCTGGGCAAGGATTTCACCCCTGCTGCACAGGCTG CCTTCCAGAAGGTGGTGGCTGGAGTGGCCACTGCCCTGGCTCACAAGTACCACTAA
Rabbit	ACCESSION V00882; REGION: join(277 … 368; 495 … 717; 1291 … 1419) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAGGCTG CTGGTTGTCTACCCATGGACCCAGAGGTTCTTCGAGTCCTTTGGGGACCTGTCCTCTGCAAATGCTGTTATGAACAATCCTAAGGTGAAGGCTCATGG CAAGAAGGTGCTGGCTGCCTTCAGTGAGGGTCTGAGTCACCTGGACAACCTCAAAGGCACCTTTGCTAAGCTGAGTGAACTGCACTGTGACAAGCT GCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTCACTCCTCAGGTGCAGGCTG CCTATCAGAAGGTGGTGGCTGGTGTGGCCAATGCCCTGGCTCACAAATACCACTGA
Rat	ACCESSION X06701; REGION: join(310 … 401; 517 … 739; 1377 …1505) Exon1 1 … 92; Exon2 93 … 315; Exon3 316 … 444; ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAGGCTGC TGGTTGTCTACCCTTGGACCCAGAGGTACTTTTCTAAATTTGGGGACCTGTCCTCTGCCTCTGCTATTATGGGTAACCCCCAGGTGAAGGCCCATGGC AAGAAGGTGATAAATGCCTTCAATGATGGCCTGAAACACTTGGACAACCTCAAGGGCACCTTTGCTCATCTGAGTGAACTCCACTGTGACAAGCTGC ATGTGGATCCTGAGAACTTCAGGCTCCTGGGCAATATGATTGTGATTGTGTTGGGCCACCACCTGGGCAAGGAATTCACCCCGTCTGCACAGGCTGC CTTCCAGAAGGTGGTAGCTGGAGTGGCCAGTGCCCTTGCTCACAAGTACCACTAA

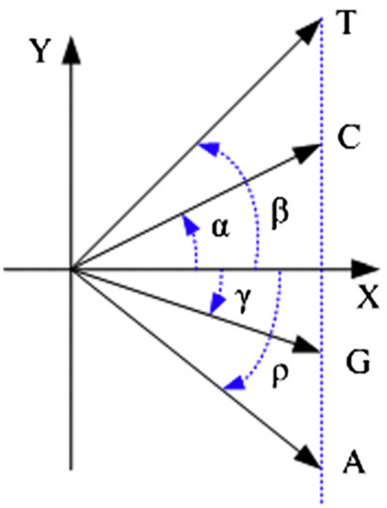

Fig. 1. The assignment for four kinds of bases to four vectors respectively.

Fig. 2. In Fig. 3, we depicted the graphical representations of the complete coding sequences of beta-globin gene of 7 different species listed in Table 1. Observing this figure, we found that there are some interesting characters which reveal the biological relationship among 7 DNA sequences. First, the figure shows intuitively that, compared with the other curves, the graph for the species Opossum is first diverted at position about 80. This indicates there should be numerous mutations over a period of regional DNA segment after position about 80; Second, from about the 120th position to about the 150th position, the patterns of the corresponding curves for the species Human, Lemur and Rat seem very highly similar. Furthermore, it wonderfully takes place over the region from position 80 to 120 in the sequence for the species Mouse. The similar patterns imply that these species Human, Lemur, Rat and Mouse have the local homology that conduces to analysis of the evolutional relation among species; Third, the curve for the species Human suddenly turned down on the bottom, and is very different from those for the other species. Due to change of these bases, Human should gradually be evolved from primitive organism and parted from them. In addition, the patterns of curves for Mouse and Rat look similar. The graphical representation of DNA sequences helps in recognizing the major difference among similar DNA sequences.

Compared with other graphical representations, our method possesses some advantages as follows:

(1) With DNA sequence extending in the order from 5′ to 3′, its graphical curve extends toward the positive direction of the x-axis, and then no longer intersects and overlaps itself. Therefore, the degeneracy is avoided completely.

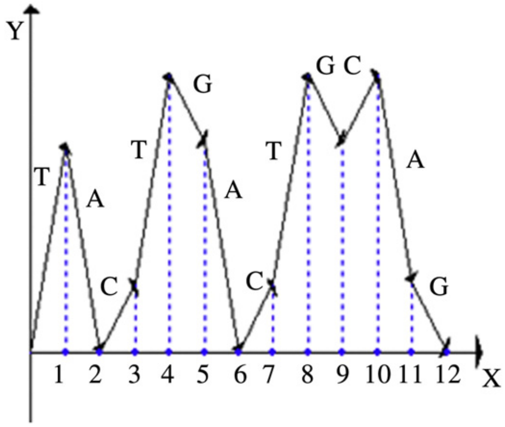

Fig. 2. The graphical representation for the example model DNA sequence TACTGACTGCAG.

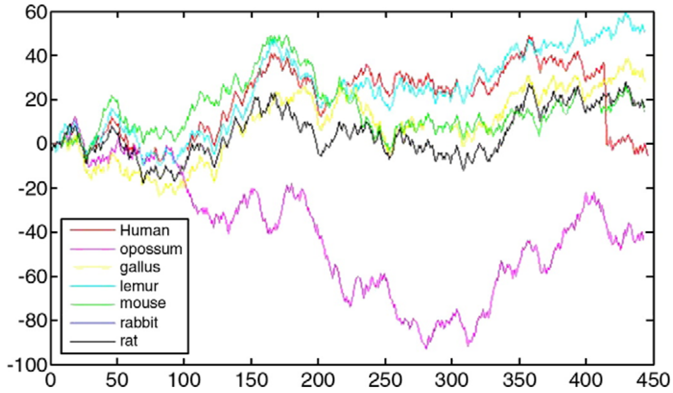

Fig. 3. The graphical representations of the complete coding sequences of β-globin gene of different 7 species.

(2) For a given array of angles, there is a one-to-one correspondence between DNA sequences and their corresponding graphical representation respectively. That is, a DNA sequence is mapped into uniquely one curve in 2D space, and on the contrary, a 2-dimensional curve allows reconstruction of only one DNA sequence.

(3) Most significantly, our method produces a simple and intuitive curve of a gene sequence which displays both local and global patterns, and any visible pattern in the curve corresponds to some pattern in the sequence of base. Therefore, it allows the identification of local homology between genomic sequences which is particularly important for functional genomic studies. Moreover, the identification is independent on positions of bases in sequences.

(4) Our 2D graphical representation is more sensitive. As shown in Fig. 3, as long as any nucleotides were changed in gene sequences, the corresponding curves would be changed and the changes in graphs are differentiated visually.

2.3. Quantitative measure in sequence comparison

Graphical representations provide many visual clues to underlying patterns in DNA sequences and can be useful in highlighting similarity and difference among various DNA sequences. However, these are generally qualitative in nature, as shown in Fig. 3, so efforts should be made to devise quantitative measures of these similarities and differences for more precise comparison among different sequences [27]. In the past two decades, the majority of work on quantitative measures among various DNA sequences has taken place, with most reports published in the past five years. Two categories of methods have been proposed—methods based on sequence alignment which has already been studied well, and methods that do not require alignment. For the first type, their computational load escalates as a power function of the length of the sequences (exponent 2 for un-gapped alignment and somewhat higher for the best gapped algorithms) making its use for searching the large databases unfeasible. Therefore, sequence comparison by alignment has both fundamental and computational limitations. Various quantitative measures for alignment-free sequence comparison have been proposed in the past, as summarized in the excellent review by Vinga et al. [28]and in the literatures [29,30]. The simplest similarity score is the Euclidian distance between the two 4^k-dimensional vectors of k-word counts [31]. Information theoretic measures like the ‘Kullback–Leibler distance’ (also called ‘relative entropy’) [32], geometric measures such as the cosine of the angle between the count vectors [33], and statistical measures such as the correlation coefficient [34] have been investigated by different authors in the context of alignment-free sequence comparison. There are also measures that do not treat every k-word's count equally, in view of the fact that different k-words' counts have different probability

Table 2
The lower triangles of the distance matrix among the complete coding sequences of beta-globin gene of different 7 species with angles $\alpha = \frac{\pi}{3}, \beta = \frac{5\pi}{18}, \gamma = -\frac{\pi}{3}, \rho = -\frac{4\pi}{9}$.

Human	0						
Opossum	0.13225	0					
Gallus	0.037966	0.14306	0				
Lemur	0.022185	0.14984	0.045987	0			
Mouse	0.030033	0.14496	0.047197	0.034129	0		
Rabbit	0.014027	0.14178	0.03963	0.022185	0.028058	0	
Rat	0.031375	0.15269	0.050054	0.035247	0.015545	0.032544	0

distributions. Thus, the 'Standardized Euclidian distance' and the ' Mahalanobis' distance [32,35] account for the variances of *k*-words in computing the Euclidian distance between the count vectors. Recently, in order to facilitate the quantitative comparison, a novel measure based on various mathematical descriptors of DNA sequences has been proposed. Those mathematical descriptors representing biological characterization of gene sequences in numerical value are often a multidimensional vector. The analysis of similarity and dissimilarity among DNA sequences is based on the assumption that two sequences are similar if their corresponding mathematical descriptors point to a similar direction and have similar magnitude. Obviously, similarity between two vectors can be measured by calculating the Euclidean distance between their end points. The smaller the Euclidean distance is, the more similar two sequences are. Some examples are described in detail in the literatures [7,17–20,23–25]. However, the computation for these mathematical descriptors representing numerical characterization of DNA sequences is complicated. Especially, for very long DNA sequences, it would take computers much computing time and memory space to finish the calculation. In the following, we proposed a new measure of the similarity and difference among various DNA sequences whose calculation is simple. Let S1 and S2 be two arbitrary DNA sequences, and $\{\vec{V}_1^1, \vec{V}_2^1, \vec{V}_3^1, \cdots, \vec{V}_n^1\}$ and $\{\vec{V}_1^2, \vec{V}_2^2, \vec{V}_3^2, \cdots, \vec{V}_n^2\}$ were respectively the corresponding sets of characteristic vectors, where *n* was the length for the DNA sequences. We defined a 'Similar Factor' between any sequences as

$$SF = \sum_{i=1}^{n}\left(1-\frac{\left|f\left(\vec{V}_i^1\right)-f\left(\vec{V}_i^2\right)\right|}{\pi}\right) \tag{1}$$

where the function $f(\vec{V}_i^k)$ $(k=1,2)$ represents angles between the *i*th characteristic vector and the *x*-axis. The larger Similar Factor is, the more similar two DNA sequences are. If S1 and S2 were identical sequences, SF = 1. Otherwise, $0 \prec SF \prec 1$. For any two sequences, SF is not equal to zero for ever. This is because all biology arose from a single molecule and they have more or less evolutional relationship each other. For N DNA sequences S1, S2, S3,…, SN, to compare their similarity and difference, we defined a similarity matrix by M and its elements m_{ij} are computed as follows

$$m_{ij} = \sum_{k=1}^{n}\left(1-\frac{\left|f\left(\vec{V}_k^i\right)-f\left(\vec{V}_k^j\right)\right|}{\pi}\right) \tag{2}$$

where $\vec{V}_k^i$ and $\vec{V}_k^j$ represent respectively the *k*th characteristic vectors for the *i*th DNA sequence and the *j*th DNA sequence. m_{ij} describes the Similar Factor between the *i*th DNA sequence and the *j*th DNA sequence. Obviously, the main diagonal elements are 1, and $m_{ij} = m_{ji}$, so similarity matrixes are symmetrical. Distance matrix representing directly the evolutionary or structural relationship is denoted by D and its elements d_{ij} are calculated as follows:

$$d_{ij} = \log_{10}\left(m_{ij}\right) \tag{3}$$

distance matrix is also symmetrical.

3. Results and discussions

We illustrated the use of the measure with the examination of similarities and dissimilarities among the complete coding sequences of beta-globin gene of different 7 species listed in Table 1. As an example, we assumed an array of angles: $\alpha = \frac{\pi}{3}, \beta = \frac{5\pi}{18}, \gamma = -\frac{\pi}{3}, \rho = -\frac{4\pi}{9}$, and then computed the distance matrix listed in Table 2. Observing Table 2, we found that the second column has the greatest entries in the distance matrix, so Opossum is most dissimilar to the others among the 7 species. This has further proven the result obtained directly by observing the graphical representations. The fourth column, the fifth column and the sixth column have smaller entries, which mean that species Lemur, Mouse, Rabbit and Rat would have been closely relative to each other. On the other hand, the most similar species pairs are Mouse–Rat and Human–Rabbit. The more similar species pairs are Human–Lemur, Rabbit–Lemur and Rabbit–Mouse. The result is in an agreement with that in the literature [21]. This conveniently discovers their evolutionary relationship among various DNA sequences.

For comparison with the multiple sequences alignment, as shown in Fig. 4, two phylogenetic trees were drawn respectively based on the distance matrix in this report and the multiple sequences alignment, by using the program Neighboring Joining in Phylip package. The program employed to make multiple sequences alignment is version 1.8 of the program Clustalx. As we saw, there was over variation tendency over both trees based on both different methods respectively, despite some variation among them. Fig. 4(a) by our method indicated that the species Rabbit and Lemur would be more closely relative, while Fig. 4(b) by multiple sequences alignment implied that both species Lemur and Human would be more closely relative. Obviously, our result seems more rational than that based on the alignment method. Compared with methods involving multiple sequences alignment and approaches associated with the computations of D/D, L/L and leading eigenvalue, our presented measure doesn't need a great deal of running time and memory space for computer, and furthermore obtain better results.

In this report, the SF measure is relatively dependent on values of angles alpha, beta, gamma and rho, chosen to represent each of the four nucleotide. With angles shifting, the distance matrix calculated out would be completely different. As shown in Eq. (1), each algebraic expression brings an impact on the values of Similar Factor, and in fact, describes the similar degree that reflects difference and homology of their structure and function between both bases. Difference for four types of nucleotide in their structure and function would determine values of angles alpha, beta, gamma and rho. If value of angles were chosen appropriately, our method works well over a large range in genome size. Next, we summarized several key points as follows:

(1) The SF measure is particularly useful for comparing whole genomes or genomic region that has high homology, and its computation is effective.

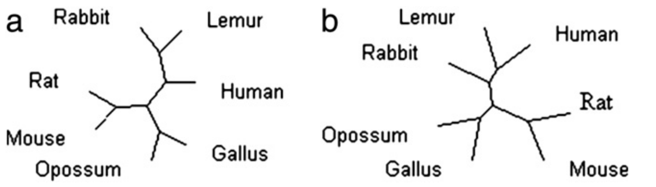

Fig. 4. Two phylogenetic trees among the complete coding sequences of seven species. a. Using our method in this report; b. Using multiple sequences alignment algorithm.

(2) The SF measure is largely dependent on values of angles alpha, beta, gamma and rho. Thus, the significance of value difference should be carefully considered. In general, we should choose values of angles alpha, beta, gamma and rho to meet the condition: $|\alpha-\beta|<|\alpha-\gamma|$, $|\alpha-\beta|<|\alpha-\rho|$, $|\alpha-\beta|<|\beta-\gamma|$, $|\alpha-\beta|<|\beta-\rho|$, $|\gamma-\rho|<|\alpha-\gamma|$, $|\gamma-\rho|<|\beta-\rho|$, $|\gamma-\rho|<|\alpha-\rho|$ and $|\gamma-\rho|<|\beta-\rho|$.

4. Conclusion

Graphical representation technique is a power and visual tool to analyze similarity and dissimilarity among various DNA sequences that has been developed in recent years. In the letter, we have denoted as four kinds of basic nucleotide A, T, G and C four different two-component vectors respectively, where the first elements are constant (equal to 1) and the second elements are different from each other and have proposed a new and universal 2D graphical method for representing DNA sequences. It permits the representation and investigation of patterns in DNA sequence, visually comparing the global and local homology among various genome sequences. On the basis of the graphical representation, we have presented a new quantitative measure of similarity and dissimilarity among various DNA sequences. In comparison with previously published method, our approach is not involved with multiple sequences alignment and computation of mathematical descriptors of DNA sequences such as eigenvalues, leading eigenvalues, average bandwidths, and so on. This provides a quick and efficient way to analyze the similarity and dissimilarity among various DNA sequences for both computational scientists and molecular biologists.

Acknowledgements

This work is supported in part by the National Nature Science Foundation of China (Grant 10571019), the National Nature Science Foundation of Hunan province (Grant 07JJ5080) and Scientific Research Fund of Hunan Provincial Education Department. The author thanks both the anonymous referees for many valuable that improved the paper.

References

[1] E. Hamori, J. Ruskin, H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (1983) 1318.
[2] M.A.J. Gates, A simple way to look at DNA, J. Theor. Biol. 119 (1986) 319.
[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin genes, Curr. Sci. 66 (1994) 309.
[4] P.M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, Comput. Appl. Biosci. 11 (1995) 503.
[5] X. Guo, M. Randic, S.C. Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy, Chem. Phys. Lett. 350 (2001) 106.
[6] S.S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.-K. Ho, DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (2003) 3078.
[7] A. Nandy, M. Harle, S.C. Basak, Mathematical descriptors of DNA sequences: development and applications, ARKIVOC (2006) 211.
[8] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (1990) 2163.
[9] J. Joseph, R. Sasikumar, Chaos game representation for comparison of whole genomes, BMC Bioinformatics 7 (2006) 243.
[10] S. Basu, A. Pan, C. Dutta, J. Das, Mathematical characterization of chaos game representation. New algorithms for nucleotide sequence analysis, J. Mol. Biol. 228 (1992) 715.
[11] K.A. Hill, N.J. Schisler, S.M. Singh, Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species, J. Mol. Evol. 35 (1992) 261.
[12] J.L. Oliver, P. Bernaola-Galvan, J. Guerrero-Garcia, R. Roman-Roldan, Entropic profiles of DNA sequences through chaos-game-derived images, J. Theor. Biol. 160 (1993) 457.
[13] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, Mol. Biol. Evol. 16 (1999) 1391.
[14] N. Goldman, Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences, Nucleic Acids Res. 21 (1993) 2487.
[15] J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, Bioinformatics 17 (2001) 429.
[16] Y. Wang, K. Hill, S. Singh, L. Kari, The spectrum of genomic signatures: from dinucleotides to chaos game representation, Gene 346 (2005) 173.
[17] M. Randić, On characterization of DNA primary sequences by a condensed matrix, Chem. Phys. Lett. 317 (2000) 29.
[18] X.Q. Liu, Q. Dai, Z. Xiu, T. Wang, PNN-curve: a new 2D graphical representation of DNA sequences and its application, J. Theor. Biol. 243 (2006) 555.
[19] Z.H. Qi, T.R. Fan, PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 442 (2007) 434.
[20] Z. Qi, X. Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides, Chem. Phys. Lett. 440 (2007) 139.
[21] X.Q. Qi, J. Wen, Z.H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, J. Theor. Biol. 249 (2007) 681.
[22] G. Huang, B. Liao, Y. Li, Z. Liu, H.-L. Curve, A novel 2D graphical representation for DNA sequences, Chem. Phys. Lett. 462 (2008) 129.
[23] Y. Yao, T. Wang, A class of new 2-D graphical representation of DNA sequences and their application, Chem. Phys. Lett. 398 (2004) 318.
[24] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1.
[25] D. Bielinska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. 442 (2007) 140.
[26] B. Liao, A 2D graphical representation of DNA sequence, Chem. Phys. Lett. 401 (2005) 196.
[27] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences—a review, J. Biosci. 23 (1998) 55.
[28] S. Vinga, J.S. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (2003) 513.
[29] M.R. Kantorovi, G.E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, Bioinformacis 23 (2007) 249.
[30] S. Vinga, J.S. Almeida, Local Renyi entropic profiles of DNA sequences, BMC Bioinformatics 8 (2007) 393.
[31] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Natl Acad. Sci. U.S.A. 83 (1986) 5155.
[32] T.J. Wu, Y.C. Hsieh, L.A. Li, Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition, Biometrics 57 (2001) 441.
[33] G.W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, Bioinformatics 18 (2002) 100.
[34] G. Fichant, C. Gautier, Statistical method for predicting protein coding regions in nucleic acid sequences, Comput. Appl. Biosci. 3 (1987) 287.
[35] T.J. Wu, J.P. Burke, D.B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, Biometrics 53 (1997) 1431.