Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 3262813, 12 pages
http://dx.doi.org/10.1155/2016/3262813

Research Article

Circular Helix-Like Curve: An Effective Tool of Biological Sequence Analysis and Comparison

Yushuang Li and Wenli Xiao

College of Science, Yanshan University, Qinhuangdao 066004, China

Correspondence should be addressed to Wenli Xiao; xiaowenli0117@sina.com

Received 18 February 2016; Accepted 19 April 2016

Academic Editor: Nadia A. Chuzhanova

This paper constructed a novel injection from a DNA sequence to a 3D graph, named circular helix-like curve (CHC). The presented graphical representation is available for visualizing characterizations of a single DNA sequence and identifying similarities and differences among several DNAs. A 12-dimensional vector extracted from CHC, as a numerical characterization of CHC, was applied to analyze phylogenetic relationships of 11 species, 74 ribosomal RNAs, 48 Hepatitis E viruses, and 18 eutherian mammals, respectively. Successful experiments illustrated that CHC is an effective tool of biological sequence analysis and comparison.

1. Introduction

Complex biological sequence analysis and comparison have been playing important roles in molecular studies. It is significant to find an effective tool for seeking a better understanding of the ever-increasing biological sequences. Graphical representation is one of such tools, which assists researchers in studying genomes in a perceivable form.

The original contributions in DNA graphical representations were compact H-curves created over 30 years ago by Hamori and cooperators [1–4]. The shape of the H-curve was a path in 3D space by mapping four nucleotides (adenine A, cytosine C, guanine G, and thymine T) to four unit vectors by four directions (NW, NE, SE, and SW). The basic rule for constructing H-curve was to move one unit in the corresponding direction and one for each unit in the z-direction. H-curve not only characterized complex genetic messages but also embodied important parameters concerning the distribution of nucleotides. Hamori's achievements encouraged many researchers to study graphical bioinformatics. Here we focus on 3D graphical representations. Among all the representations following Hamori's idea, the most worthy to mention is that in the year 2000, Randić and cooperators [5] constructed a new model of DNA sequences which was also based on a path in 3D space. The difference from H-curve is that four vectors corresponding to four bases are located along tetrahedral directions. Moreover, Randić et al. described a particular scheme transforming the spatial model into a numerical matrix representation. This is an important breakthrough that led to the expansion of graphical technique from a visual discipline to a qualitative discipline. One of successful models is the Z-curve, created by Zhang et al. [6]. The construction of Z-curve combined with three classifications of the DNA bases, purines/pyrimidines (A, G)/(C, T), amino/keto groups (A, C)/(G, T), and strong/weak hydrogen bonds (A, T)/(G, C), and assigned A$\rightarrow$(1, 1, 1), T$\rightarrow$(−1, −1, 1), C$\rightarrow$(−1, 1, −1), and G$\rightarrow$(1, −1, −1), respectively. Z-curve is famous for its extensive applications in comparative genomics, gene prediction, computation of G + C content with a windowless technique, prediction of replication origins, and terminations of bacterial and archaeal genome. But to our regret, there are crosses and overlaps of the spatial curve in the representation in [5], and the Z-curve might cause a loop if the frequencies of the four bases present in the sequence are the same as pointed out by Tang et al. [7]. To overcome the degeneration appearing in the above representations, other various improvements or transformations were created [8–14]. Recently, Pesek and Zerovnik [15] presented a modified Hamori's curve by using analogous embedding into the strong product of graphs, $K_4 \otimes P_n$ (K_4 is a 4-order complete graph and P_n is an n-order

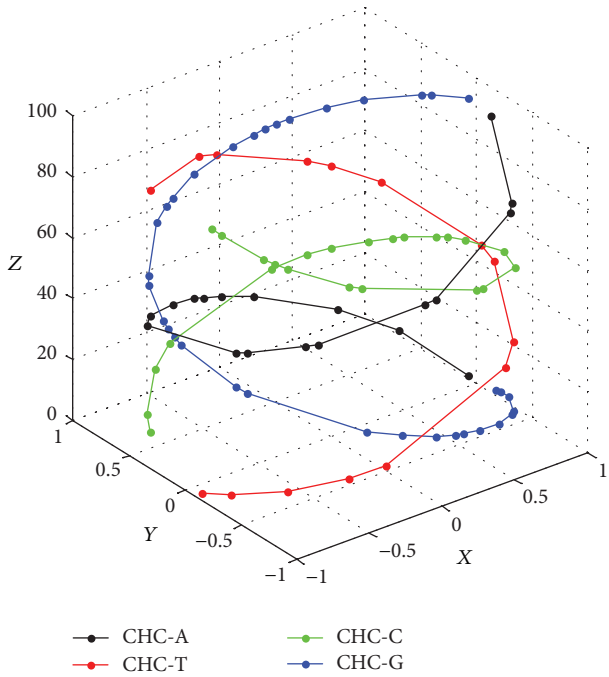

Figure 1: CHC of Gallus.

path), with weighted edges. Xie and Mo [16] also considered three classifications of the DNA bases, assigned three types of vectors to the four bases, respectively, and derived three 3D graphical representations.

The above models were all based on individual nucleotides such that it was easy to inspect compositions and distributions of four bases directly, but difficult to dinucleotides or trinucleotides in DNA sequences. Some researchers solved this problem by assigning different vectors to each dinucleotide or to each trinucleotide in 3D space. For example, Qi and Fan [17] in the year 2007 assigned 16 vectors to 16 dinucleotides and then defined a map from a DNA sequence to a characteristic plot set, while the corresponding curves extended along z axes. Subsequently, based on similar research object Qi et al. [18] presented another 3D graphical representation. Two papers were highly dissimilar in the following aspects: the methods and contents of research, the map used to construct graphical representation, the graphical curve, and numerical invariants characterizing DNA sequences. Other 3D models [19, 20] based on dinucleotides have been also proposed. Yu et al. [21] in the year 2009 presented a novel 3D graphical representation based on trinucleotides, TN-curve, which is the first model that can display the information of trinucleotides within 3D space. Recently, Jafarzadeh and Iranmanesh [22] proposed a 3D model, C-curve, also based on trinucleotides.

All works mentioned above almost involved sequence comparison. The most popular tools for comparing sequences are alignment methods including the alignment-based and the alignment-free. In general, most alignment-free methods take less computational time than alignment-based ones. Moreover, they are more sensitive against short or partial sequences [23] and more efficient in comparing gene regulatory regions [24]. In this paper, we introduce a new 3D graphical representation of DNA sequences, namely, circular helix-like curve (CHC), which is highly different from techniques referred to above. It is composed of four characteristic curves (CHC-A, CHC-C, CHC-G, and CHC-T) which just correspond to four bases (A, C, G, and T) in DNA. The novel injection from a DNA sequence to a point set in 3D space ensures CHC without loss of information. A 12-dimensional vector extracted from CHC, as its numerical characterization, provides effective conditions for alignment-free sequence comparison.

The paper is organized as follows: in Section 2, we describe the construction of the CHC, its several properties, and its numerical characterization; in Section 3, we exhibit applications of CHC by analyzing phylogenetic relationships of 11 species, 74 ribosomal RNAs, 48 Hepatitis E viruses, and 18 eutherian mammals, respectively. Finally, a conclusion ends the paper.

2. Circular Helix-Like Curve

2.1. Construction of Circular Helix-Like Curve. Given a DNA sequence $G = g_1 g_2, \ldots, g_n$ with length n, define the map φ as follows: for $i = 1, 2, \ldots, n$, if n is even, then

$$\varphi(g_i) = \begin{cases} \left(\cos\frac{2\pi i}{n+1}, \sin\frac{2\pi i}{n+1}, i\right) & \text{if } g_i = \text{A} \\ \left(-\cos\frac{2\pi i}{n+1}, \sin\frac{2\pi i}{n+1}, i\right) & \text{if } g_i = \text{C} \\ \left(\cos\frac{2\pi i}{n+1}, -\sin\frac{2\pi i}{n+1}, i\right) & \text{if } g_i = \text{G} \\ \left(-\cos\frac{2\pi i}{n+1}, -\sin\frac{2\pi i}{n+1}, i\right) & \text{if } g_i = \text{T}. \end{cases} \quad (1)$$

If n is odd, then

$$\varphi(g_i) = \begin{cases} \left(\cos\frac{2\pi i}{n+2}, \sin\frac{2\pi i}{n+2}, i\right) & \text{if } g_i = \text{A} \\ \left(-\cos\frac{2\pi i}{n+2}, \sin\frac{2\pi i}{n+2}, i\right) & \text{if } g_i = \text{C} \\ \left(\cos\frac{2\pi i}{n+2}, -\sin\frac{2\pi i}{n+2}, i\right) & \text{if } g_i = \text{G} \\ \left(-\cos\frac{2\pi i}{n+2}, -\sin\frac{2\pi i}{n+2}, i\right) & \text{if } g_i = \text{T}. \end{cases} \quad (2)$$

The function φ maps each nucleotide g_i in the sequence G to one point (x_i, y_i, z_i) in 3D space. Let $\varphi_{\text{A}} = \{\varphi(g_i) \mid g_i = \text{A} \text{ and } g_i \in G\}$. Similarly define φ_{C}, φ_{G}, and φ_{T}. Connect the adjacent points in φ_{A} by lines and then obtain a circular helix-like curve in 3D space representing the trail of base A in the sequence, namely, circular helix-like curve-A (CHC-A) for convenience. In the same way, we can obtain CHC-C, the symmetric curve about yoz plane of a circular helix-like curve; CHC-G, the symmetric curve about xoz plane of a circular helix-like curve; CHC-T, the symmetric curve about z-axis of a circular helix-like curve. Clearly, projective points on xoy plane of points in four curves are all assigned over the circumference of a unit circle. Figure 1 shows the circular

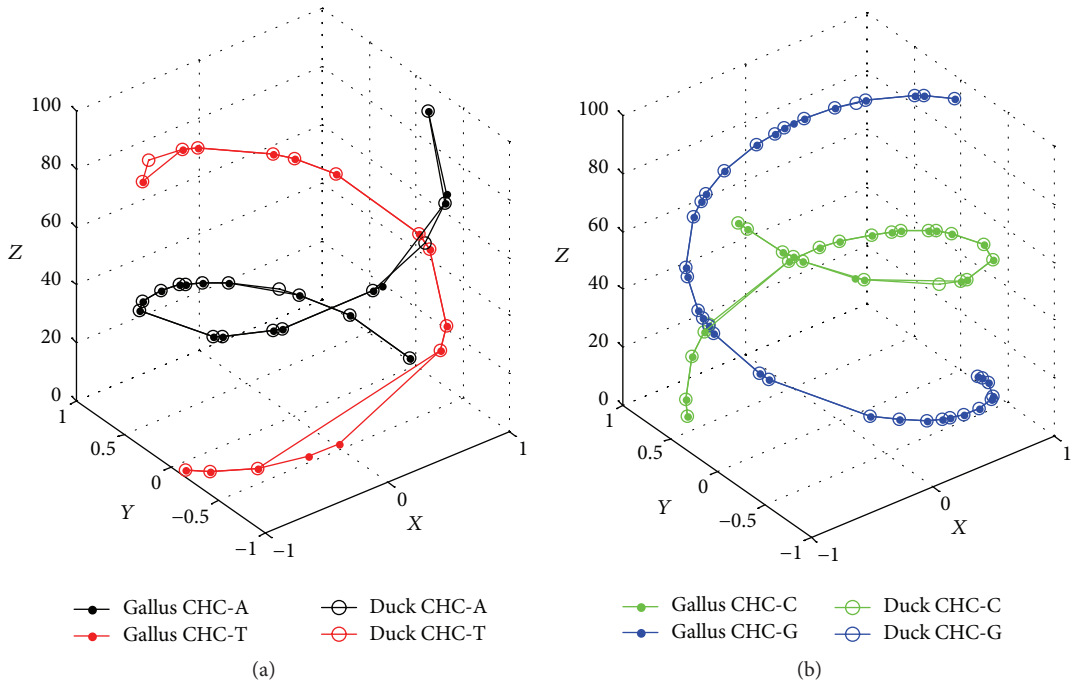

Figure 2: (a) AT-comparison of Gallus and Duck. (b) CG-comparison of Gallus and Duck.

helix-like curve (CHC) of the first exon of beta-globin gene of Gallus:

ATGGTGCACTGGACTGCTGAGGAGAAGCAG-CTCATCACCGGCCTCTGGGGCAAGGTCAAT-GTGGCCGAATGTGGGGCCGAAGCCCTGGCC-AG.

2.2. Properties of CHC

Property 1. The map φ defined in Section 2.1 is an injection; thus no information of DNA sequence is lost.

It is sufficient to prove that $\varphi(g_1) \neq \varphi(g_2)$ if $g_1 \neq g_2$. In fact, it is to prove that for all $i = 1, 2, \ldots, n$, when n is even, $2\pi i/(n+1) \neq \pi/2, \pi, 3\pi/2$, and when n is odd, $2\pi i/(n+2) \neq \pi/2, \pi, 3\pi/2$. That is, $i \neq (n+1)/4, (n+1)/2, 3(n+1)/4$ if n is even and $i \neq (n+2)/4, (n+2)/2, 3(n+2)/4$ if n is odd. It is clear because $n+1$ is odd when n is even, and $n+2$ is also odd when n is odd. Complete the proof.

Property 2. CHC could reflect base composition and distribution of a DNA sequence.

On one hand, the base composition is easily determined by the point density of the corresponding CHC. Take Figure 1 as an example, CHC-G has high point density which implies that the first exon of beta-globin gene of Gallus has high G-content. Oppositely, from CHC-T one could derive that Gallus has low T-content. Also from CHC-A and CHC-C it is clear that Gallus has similar contents of bases A and C. On the other hand, the base distribution could be identified by the arrangement of points on their CHC, respectively. As is shown in Figure 1, we are able to find special regions of curves, such as the thickset regions and the sparse regions. Also, one could easily catch sight of spacing distances of each kind of base.

Property 3. CHC is an effective tool of identifying dissimilarities (similarities) among equal length sequences.

Take the first exons of β-globin genes of Gallus and Duck as instance (their lengths are both 92).

Gallus: ATGGTGCACTGGACTGCTGAGGAGAAGCAG-CTCATCACCGGCCTCTGGGGCAAGGTCAAT-GTGGCCGAATGTGGGGCCGAAGCCCTGGCC-AG.

Duck: ATGGTGCACTGGACAGCCGAGGAGAAGCAG-CTCATCACCGGCCTCTGGGGCAAGGTCAAT-GTGGCCGACTGTGGAGCTGAGGCCCTGGCC-AG.

Two sequences only have six mismatches in the sequence level (shown in red), and both have similar composition and distribution of nucleic bases as Figure 2 goes. Especially for CG-comparison (see Figure 2(b)), their respective CHCs are nearly coincident. Certainly, their differentiations from Figure 2(a) could not be concealed, because CHC-Ts and CHC-As both have obvious deviations.

Discussion. Property 3 shows that CHC is convenient to visually compare sequences with equal length; but for unequal length sequences, there may appear some puzzles. For example, the lengths of the first exons of globin genes of Human and Gorilla are 92 and 93, respectively, and their

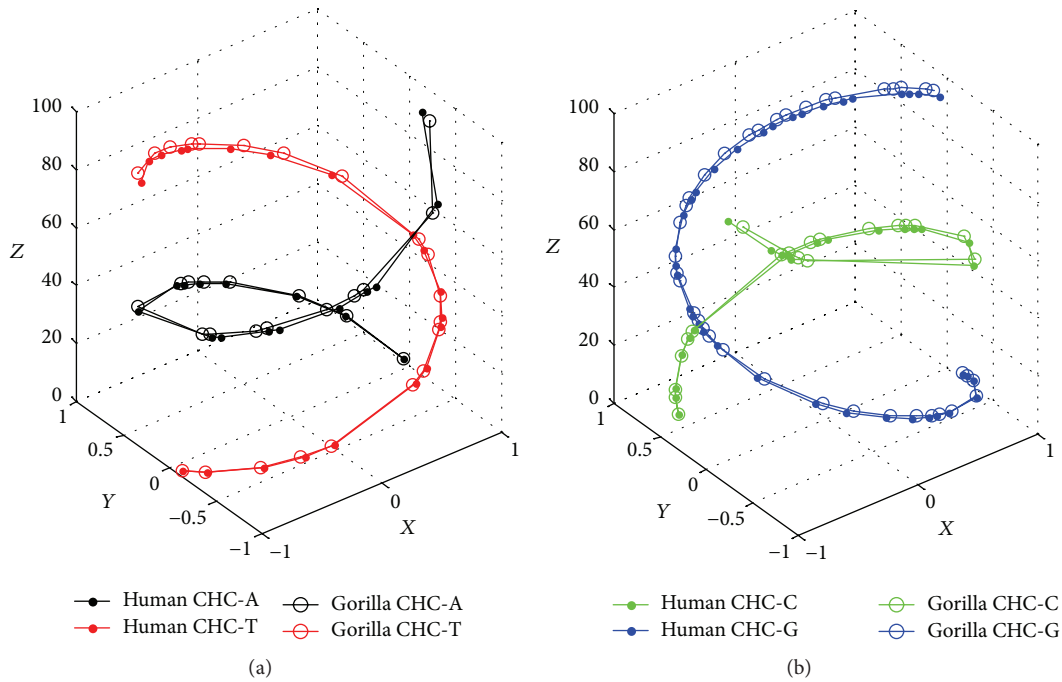

FIGURE 3: (a) Original AT-comparison of Human and Gorilla. (b) Original CG-comparison of Human and Gorilla.

corresponding bases are completely matched from 1 to 92 except the last base in gorilla, but their CHCs deviate from each other (see Figures 3(a) and 3(b)) due to different lengths. This phenomenon is unacceptable. How to solve this trouble? One could change the function φ defined in Section 2.1 slightly for two sequences to avoid the influence of different lengths. Without loss of generality, suppose sequence Seq1 has l bases, sequence Seq2 has m bases, and $m = l + t$ $(t \geq 1)$.

(i) If l and m are both even, that is, t is even, assign $n+1$ in (1) to $l+t+1$ for Seq1 and to $m+1$ for Seq2, respectively; thus $l + t + 1 = m + 1$ are both odd.

(ii) If l and m are both odd, that is, t is even, assign $n+2$ in (2) to $l+t+2$ for Seq1 and to $m+2$ for Seq2, respectively; thus $l + t + 2 = m + 2$ are both odd.

(iii) If l is even and m is odd, that is, t is odd, assign $n + 1$ in (1) to $l+t+2$ for Seq1 and assign $n+2$ in (2) to $m+2$ for Seq2, respectively; thus $l + t + 2 = m + 2$ are both odd.

(iv) If l is odd and m is even, that is, t is odd, assign $n + 2$ in (2) to $l+t+1$ for Seq1 and assign $n+1$ in (1) to $m+1$ for Seq2, respectively; thus $l + t + 1 = m + 1$ are both odd.

It is not difficult to conclude that the above four modifications still keep Property 1. Execute measure (iii) to compare the first exons of globin genes of Human and Gorilla; that is, assign $n + 1$ in (1) to $92 + 1 + 2$ for Human and assign $n+2$ in (2) to $93 + 2$ for Gorilla. Optimal CHC comparison of Human and Gorilla appears (see Figures 4(a) and 4(b)). Their corresponding bases are matched very well from 1 to 92, and the last base G in Gorilla is quite striking.

Above programs solve the CHC comparison of DNA sequences from front to back. In fact, we are also able to dispose the CHC comparison of DNA sequences from back to front by taking the similar method. Besides making lengths of compared sequences "equal" by changing the assignment of $n + 1$ in (1) or $n + 2$ in (2), one needs to adjust "the start location" of comparison by changing the assignment of i in (1) or (2) for the shorter sequence, such that two compared sequences have the same "end locations." For example,

G_1: ATTTGGCACCTAAAACGTCGTATATAAAGG-GGTCTCA.

G_2: GGCACCTAAAACGTCGTATATAAAGGGGTC-TCA.

The lengths of G_1 and G_2 are 37 and 33, respectively. G_2 just matches the fragment of G_1 from position 5 to position 37. Modify the function φ as (3). Figures 5(a) and 5(b) show the CHC comparison of two sequences:

$$\varphi(g_i) = \begin{cases} \left(\cos\dfrac{2\pi i}{37+2}, \sin\dfrac{2\pi i}{37+2}, i\right) & \text{if } g_i = \text{A} \\ \left(-\cos\dfrac{2\pi i}{37+2}, \sin\dfrac{2\pi i}{37+2}, i\right) & \text{if } g_i = \text{C} \\ \left(\cos\dfrac{2\pi i}{37+2}, -\sin\dfrac{2\pi i}{37+2}, i\right) & \text{if } g_i = \text{G} \\ \left(-\cos\dfrac{2\pi i}{37+2}, -\sin\dfrac{2\pi i}{37+2}, i\right) & \text{if } g_i = \text{T} \end{cases}$$

$$g_i \in G_1, \quad i = 1, 2, \ldots, 37,$$

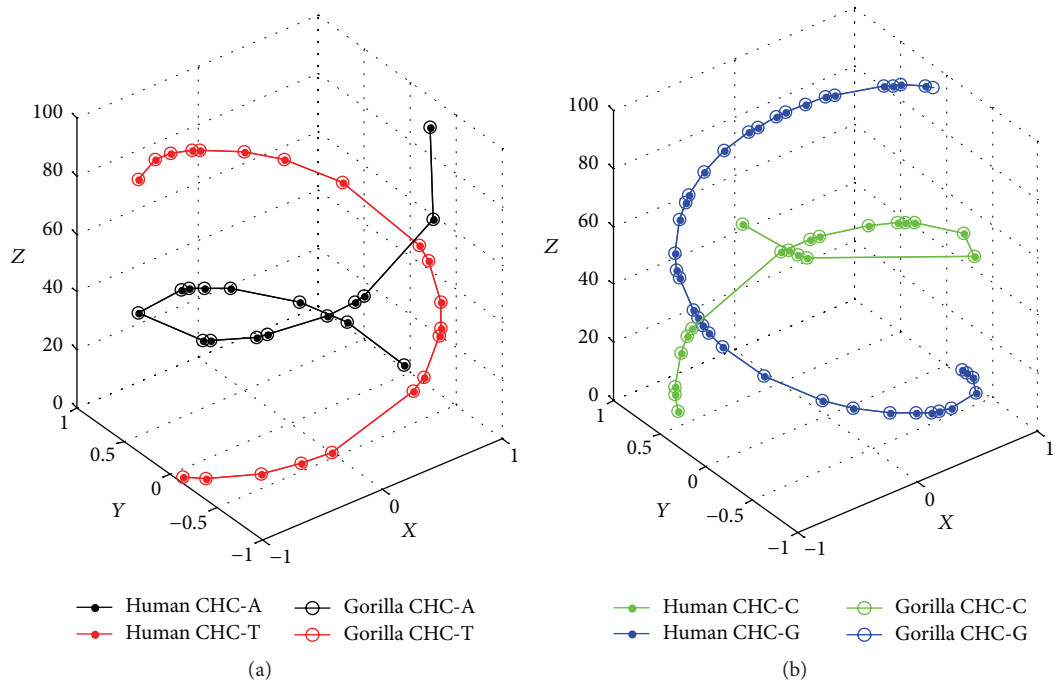

FIGURE 4: (a) Optimal AT-comparison of Human and Gorilla. (b) Optimal CG-comparison of Human and Gorilla.

$$\varphi(g_i) = \begin{cases} \left(\cos\frac{2\pi(i+4)}{33+4+2}, \sin\frac{2\pi(i+4)}{33+4+2}, i+4\right) & \text{if } g_i = \text{A} \\ \left(-\cos\frac{2\pi(i+4)}{33+4+2}, \sin\frac{2\pi(i+4)}{33+4+2}, i+4\right) & \text{if } g_i = \text{C} \\ \left(\cos\frac{2\pi(i+4)}{33+4+2}, -\sin\frac{2\pi(i+4)}{33+4+2}, i+4\right) & \text{if } g_i = \text{G} \\ \left(-\cos\frac{2\pi(i+4)}{33+4+2}, -\sin\frac{2\pi(i+4)}{33+4+2}, i+4\right) & \text{if } g_i = \text{T} \end{cases}$$

$$g_i \in G_2, \quad i = 1, 2, \ldots, 33. \tag{3}$$

2.3. Numerical Characterization of CHC. As we have seen, CHC appears pleasing to the eyes about identifying single DNA sequence and comparing DNA sequences. The more important is the numerical characterization derived from CHC, a 12-dimensional vector. It not only captures the essence of the base composition and distribution in DNA sequence, but also allows one to estimate similarity or dissimilarity between different DNAs quantitatively. Given a sequence $G = g_1 g_2, \ldots, g_n$ with length n, let

$$X_{\text{A}} = \frac{\sum_{i=1}^{n} x_i^{\text{A}}}{n},$$

$$Y_{\text{A}} = \frac{\sum_{i=1}^{n} y_i^{\text{A}}}{n},$$

$$Z_{\text{A}} = \frac{2\sum_{i=1}^{n} z_i^{\text{A}}}{n(n+1)},$$

$$X_{\text{C}} = \frac{\sum_{i=1}^{n} x_i^{\text{C}}}{n},$$

$$Y_{\text{C}} = \frac{\sum_{i=1}^{n} y_i^{\text{C}}}{n},$$

$$Z_{\text{C}} = \frac{2\sum_{i=1}^{n} z_i^{\text{C}}}{n(n+1)},$$

$$X_{\text{G}} = \frac{\sum_{i=1}^{n} x_i^{\text{G}}}{n},$$

$$Y_{\text{G}} = \frac{\sum_{i=1}^{n} y_i^{\text{G}}}{n},$$

$$Z_{\text{G}} = \frac{2\sum_{i=1}^{n} z_i^{\text{G}}}{n(n+1)},$$

$$X_{\text{T}} = \frac{\sum_{i=1}^{n} x_i^{\text{T}}}{n},$$

$$Y_{\text{T}} = \frac{\sum_{i=1}^{n} y_i^{\text{T}}}{n},$$

$$Z_{\text{T}} = \frac{2\sum_{i=1}^{n} z_i^{\text{T}}}{n(n+1)}, \tag{4}$$

where $(x_i^{\text{A}}, y_i^{\text{A}}, z_i^{\text{A}})$ is the coordinate of the ith base A in the sequence and others are the same. Define the 12-dimensional

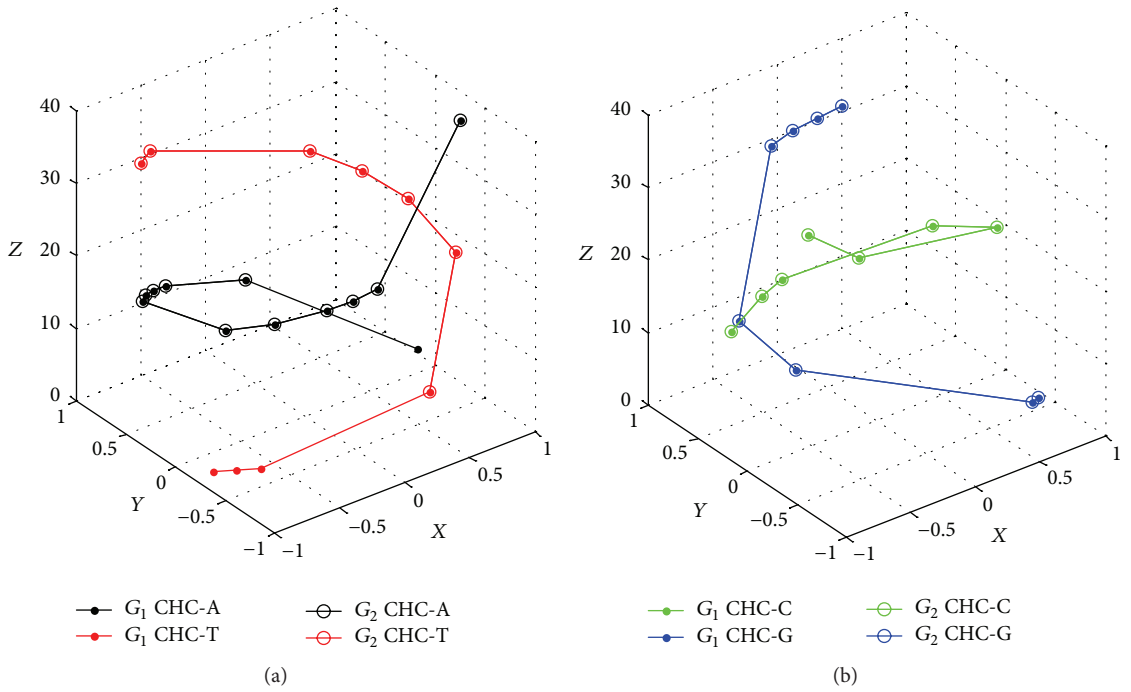

Figure 5: (a) AT-comparison of G_1 and G_2 from back to front. (b) CG-comparison of G_1 and G_2 from back to front.

vector $(X_A, X_C, X_G, X_T, Y_A, Y_C, Y_G, Y_T, Z_A, Z_C, Z_G, Z_T)$ as a numerical characterization of CHC. Note that $Z_A + Z_C + Z_G + Z_T = 1$ because

$$\begin{aligned} & Z_A + Z_C + Z_G + Z_T \\ &= \frac{2}{n(n+1)} \left(\sum_{i=1}^{n} \left(Z_i^A + Z_i^C + Z_i^G + Z_i^T \right) \right) \\ &= \frac{2}{n(n+1)} \left(\sum_{g_i=A} i + \sum_{g_i=C} i + \sum_{g_i=G} i + \sum_{g_i=T} i \right) \\ &= \frac{2}{n(n+1)} (1 + 2 + \cdots + n) = 1. \end{aligned} \tag{5}$$

3. Applications of CHC

In this section we use numerical characterization of CHC to compare and analyze complete coding sequences of β-globin genes of 11 species, 74 sequences from 16S ribosomal RNA, 48 Hepatitis E viruses, and whole mitochondrial genomes of 18 eutherian mammals. The average lengths of sequences from four experiments are 444, 1471, 7214, and 16572, respectively (see Tables 1, 2, 3, and 4 in Supplementary Materials available online at http://dx.doi.org/10.1155/2016/3262813). Here we choose Euclidean distance as the measure tool. The basis of sequence comparison is that the smaller the Euclidean distance of two numerical characterizations is, the more similar the two corresponding sequences are. We first calculate the similarity/dissimilarity matrix of sequences by computing their Euclidean distances and then utilize the similarity/dissimilarity matrix to construct phylogenetic tree by Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method in the Molecular Evolutionary Genetics Analysis (MEGA) software package. Comparisons with existing results confirm that the presented method is an effective classification tool of DNA sequences.

3.1. Similarity/Dissimilarity Analysis of the Complete Coding Sequences of β-Globin Genes of 11 Species. Table 1 exhibits the similarity/dissimilarity matrix of the complete coding sequences of β-globin genes of 11 species (see Table 1 in Supplementary Materials) based on Euclidean distance. Conclusions are in agreement with known facts of evolution. The most similar species pairs are Gorilla-Chimpanzee; species pairs Human-Chimpanzee, Human-Gorilla, Goat-Bovine, and Rat-Mouse are closely related to each other, while Opossum and Gallus tend to be significantly different from others. Our phylogenetic tree (see Figure 6) is in good consistency with the common accepted structure except Lemur which is primitive quadrumana but it is not in one branch together with (Human, Gorilla, and Chimpanzee). This phenomenon may be possible, because one gene may have begun to differentiate before the variation of its corresponding species happens, and thus, the differentiation time of gene may be earlier than that of the species.

To check the validity of the presented technique, we compared results in [20, 25–27] with ours (they all applied the same test data). Since different methods generate different magnitudes of the indexes, all indexes normalized to Human-Gallus number individually. Table 2 shows the comparisons of similarity/dissimilarity indexes for 11 species, and Figure 7

Table 1: The similarity/dissimilarity matrix of 11 species.

Species	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Goat	Bovine	Rabbit	Opossum	Gallus
Human	0.0316	0.0286	0.0432	0.0319	0.0439	0.0373	0.0546	0.0385	0.0594	0.0862
Chimpanzee		0.0121	0.0341	0.0519	0.0656	0.0338	0.0475	0.0457	0.0649	0.1072
Gorilla			0.0852	0.0487	0.0653	0.0352	0.0446	0.0385	0.0573	0.1058
Lemur				0.0477	0.0653	0.0292	0.0366	0.0391	0.0605	0.1050
Rat					0.0329	0.0343	0.0485	0.0341	0.0557	0.0822
Mouse						0.0531	0.0747	0.0610	0.0801	0.0667
Goat							0.0271	0.0327	0.0534	0.1041
Bovine								0.0306	0.0429	0.1234
Rabbit									0.0423	0.1049
Opossum										0.1186

Table 2: Comparisons of similarity/dissimilarity indexes for 11 species.

Number	Species	Ours	Ref [25]	Ref [20]	Ref [26]	Ref [27]
1	Human-Chimpanzee	0.3666	0.3142	0.1987	0.1965	0.1204
2	Human-Gorilla	0.3318	0.2595	0.1934	0.2562	0.0473
3	Human-Lemur	0.5012	0.5140	0.4506	1.1974	0.1657
4	Human-Rat	0.3701	0.6385	0.5911	0.6023	0.2742
5	Human-Mouse	0.5093	0.5807	0.4296	0.6085	0.1025
6	Human-Goat	0.4327	0.4920	0.4455	0.8131	0.0434
7	Human-Bovine	0.6334	0.4873	0.4354	0.6860	0.3471
8	Human-Rabbit	0.4466	0.4684	0.5049	0.6119	0.1914
9	Human-Opossum	0.6891	0.9340	0.6912	1.2931	0.8817
10	Human-Gallus	1.0000	1.0000	1.0000	1.0000	1.0000

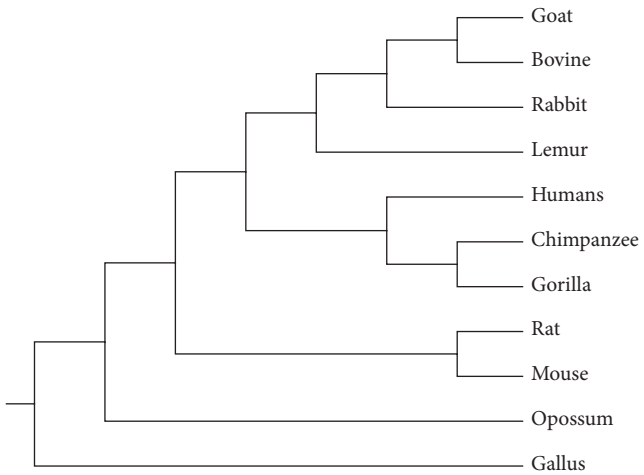

Figure 6: The phylogenetic tree of 11 species.

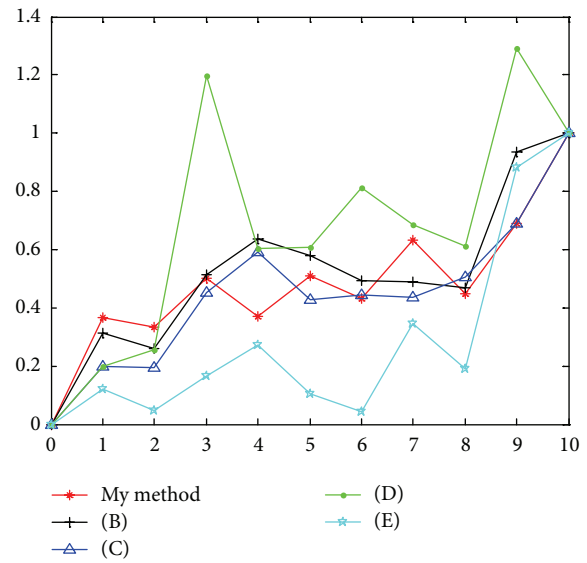

Figure 7: Line chart of Table 2.

is the line chart of Table 2. Obviously [20, 25] and ours get the right trend in the rough, but [26, 27] both show some unreasonable or contrary results. For example, the number of Human-Lemur in [26] is 1.1974 and Human-Opossum is 1.2931, which are both bigger than 1, the number of Human-Gallus. There are two digits in [27], Human-Mouse 0.1025 and Human-Goat 0.0434, both less than Human-Chimpanzee 0.1204.

3.2. Similarity/Dissimilarity Analysis of 74 Sequences from 16S Ribosomal RNA. 16S ribosomal RNA is a DNA sequence

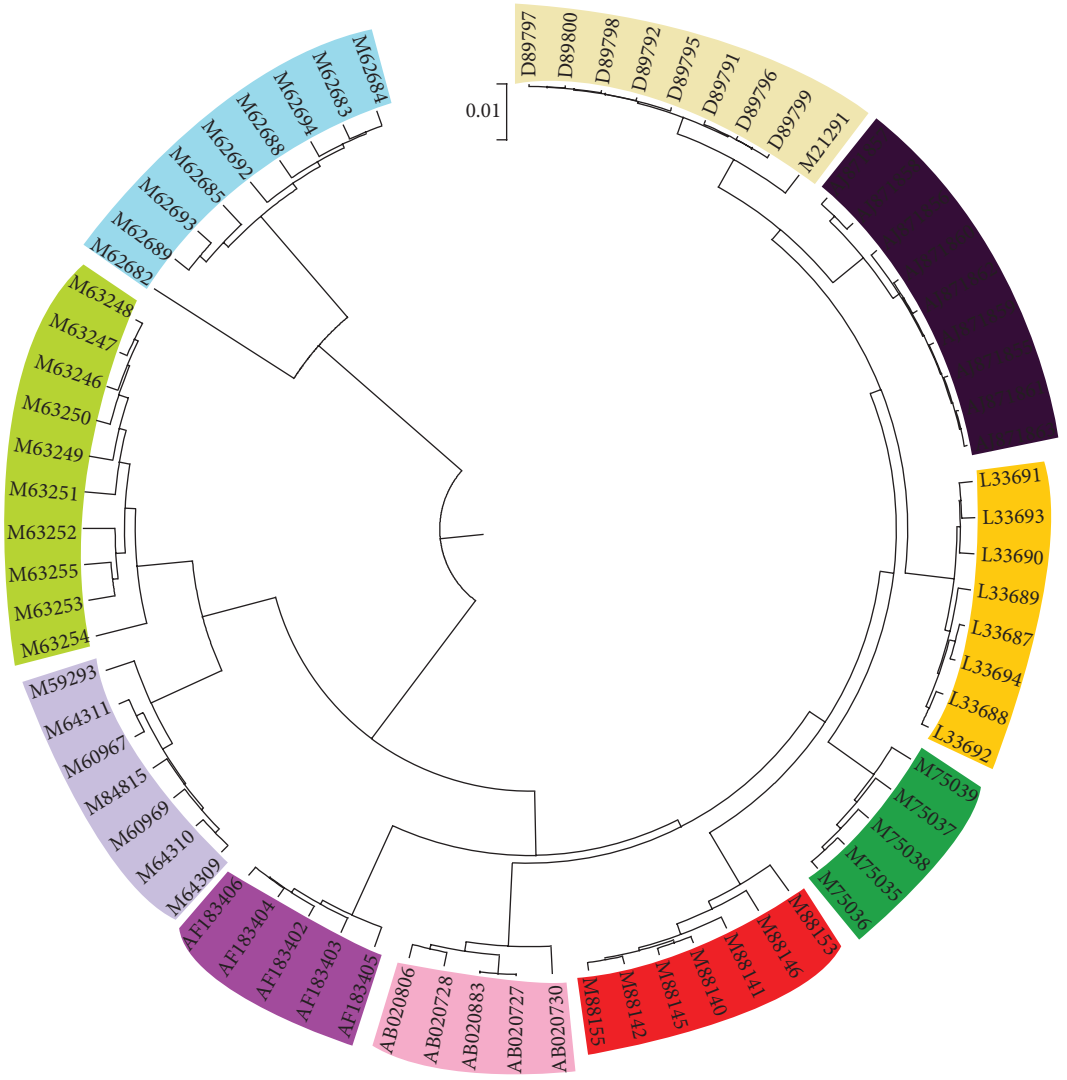

Figure 8: The phylogenetic tree of 74 sequences of 16S ribosomal RNA.

corresponding to encoding rRNA in bacteria and has high conservation and specificity. In this subsection we analyze 74 sequences from 16S ribosomal RNA. The data set consists of 10 *Buchnera aphidicola*, 9 *Coxiella burnetii*, 9 *Fibrobacter succinogenes*, 9 *Klebsiella oxytoca*, 8 *Azoarcus tolulyticus*, 7 *Borrelia burgdorferi*, 7 *Helicobacter* sp., 5 *Aggregatibacter actinomycetemcomitans*, 5 *Alloprevotella tannerae*, and 5 *Clostridium scindens*. Detail information is described in Table 2 in Supplementary Materials. Utilizing similarity/dissimilarity matrix of 74 sequences (see Table 5 in Supplementary Materials) we construct the phylogenetic tree (see Figure 8) which is consistent with the result in [28]. Ten genotypes just correspond to ten branches of the tree as anticipation.

3.3. Similarity/Dissimilarity Analysis of 48 Hepatitis E Viruses. Hepatitis E viruses (HEV) are nonenveloped, positive-sense, and single-stranded RNA viruses and belong to Herpesvirus genus [29]. Hepatitis E is considered as a public health problem and caused much concern. Until now several classifications of HEV have been proposed; the most accepted one is the classification of four major genotypes [28–32]. Genotypes I–IV are represented by the Burmese isolates, the Mexican isolate, the US isolates, and the new Chinese isolates, respectively. Here we construct the phylogenetic tree (see Figure 9) of 48 Hepatitis E viruses (see Table 3 in Supplementary Materials) based on the similarity/dissimilarity matrix (see Table 6 in Supplementary Materials), which is basically in agreement with the results presented in [28–32]. 48 HEVs are divided into four genotypes distinctly: 16 HEVs are included in genotype I, 17 HEVs in genotype III, and 14 HEVs in genotype IV; M1 is only contained in genotype II and far away from genotype I which is consistent with the structure in [31]. Moreover, some divergences in subtype classification with the result [29] keep high consistency with the result [32]: T1, which is of subtype IVc in [29], is more close to subtype IVa in [32] and ours. Also, subtype IIIc is more close to subtype IIIa.

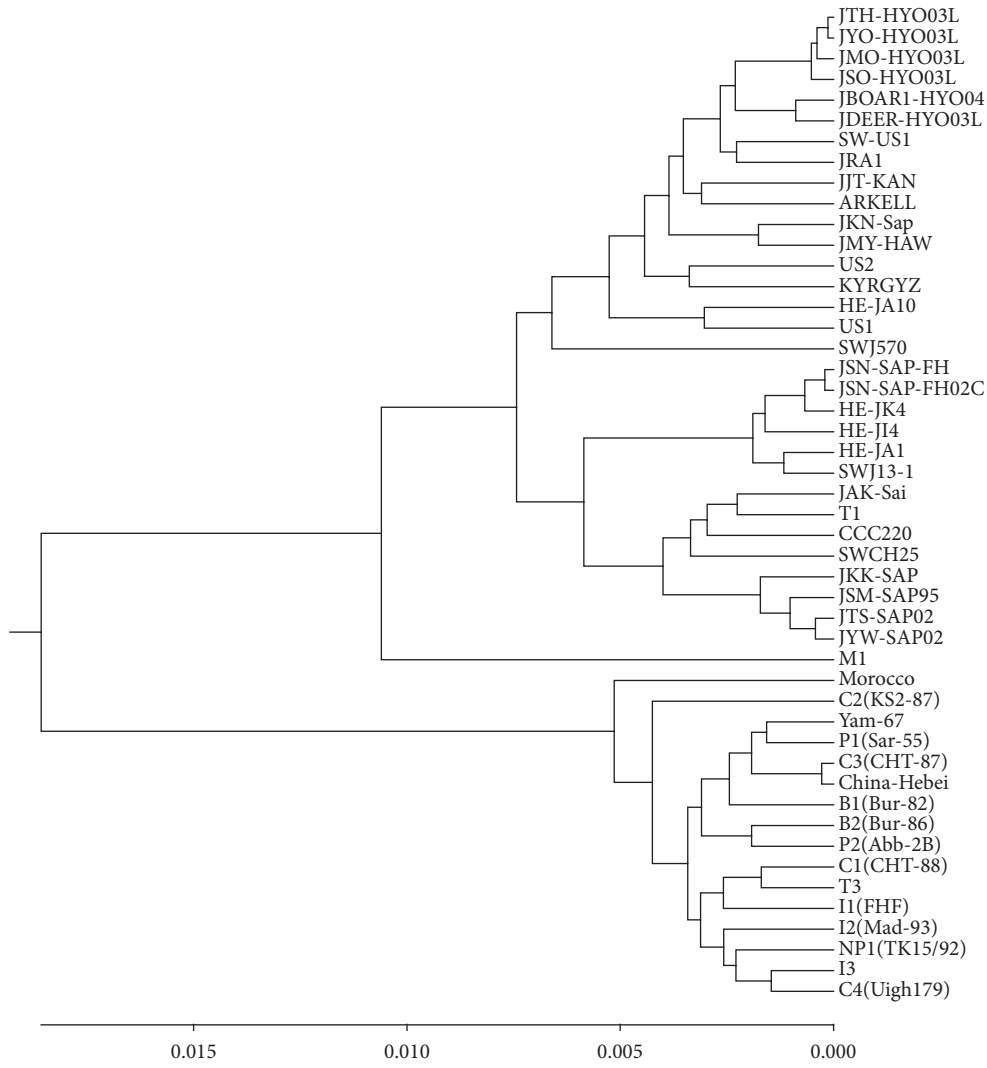

Figure 9: The phylogenetic tree of 48 Hepatitis E viruses.

3.4. Similarity/Dissimilarity Analysis of Whole Mitochondrial Genomes of 18 Eutherian Mammals. We choose a complete DNA sequence of 18 eutherian mammals as a long sequence set, which had been studied in [28, 32–34]; the longest and the shortest lengths of sequences are 17019 and 16295, respectively. Table 4 in Supplementary Materials gives the detailed information. 18 eutherian mammals could be divided into two classes: placental mammals and nonplacental mammal. Placental mammals could also be divided into three groups: Primates, Ferungulates, and Rodents. We construct the phylogenetic tree (see Figure 10) of 18 eutherian mammals based on the similarity/dissimilarity matrix (see Table 7 in Supplementary Materials). In our phylogenetic tree, Platypus, the only nonplacental mammal, is significantly different from others and in the outside of the tree. Rodents first cluster with Ferungulates, and then they cluster with Primates; that is, our result supports the topology of Primates, Rodents, Ferungulates, which is consistent with the structures in [28, 33, 34] and is slightly different from the result in [32].

3.5. Discussion. Generally speaking, CHC of DNA sequence depends on the map order of four bases on the graph. By changing the map order we will obtain different graphical representations for the same DNA sequence. Even so, applying each graphical representation to compare DNA sequences, we will draw the same analysis conclusion.

Proposition 1. *Both geometric center vectors and Euclidean distances ensure together that similarities between DNA sequences are independent of the map order of four bases.*

Proof. Take two DNA sequences G_1 and G_2 as an example. Without loss of generality, suppose their lengths are n_1 and n_2, respectively, and both even (other cases are similar). Their corresponding geometric center vectors are $\mathbf{V}_1 = [X_A X_G X_C X_T Y_A Y_G Y_C Y_T Z_A Z_G Z_C Z_T]$ and $\mathbf{V}_2 = [X'_A X'_G X'_C X'_T Y'_A Y'_G Y'_C Y'_T Z'_A Z'_G Z'_C Z'_T]$, respectively. Then the Euclidean distance between $\mathbf{V}_1$ and $\mathbf{V}_2$ is

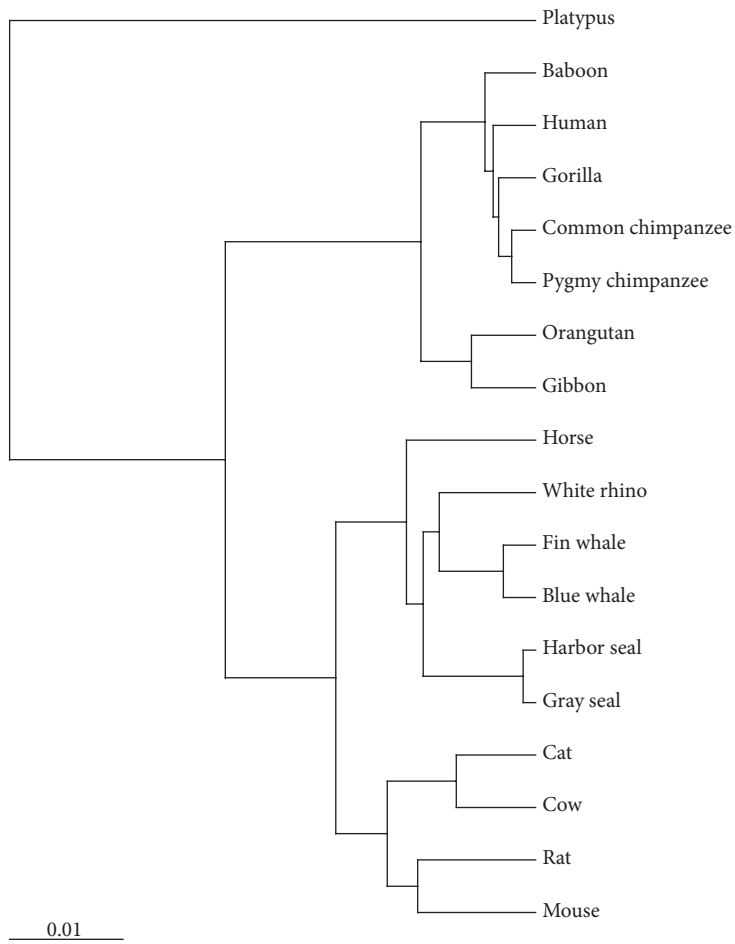

Figure 10: The phylogenetic tree of 18 eutherian mammals.

$$\left|\mathbf{V}_1 - \mathbf{V}_2\right| = \sqrt{\left(X_A - X'_A\right)^2 + \cdots + \left(X_T - X'_T\right)^2 + \left(Y_A - Y'_A\right)^2 + \cdots + \left(Y_T - Y'_T\right)^2 + \cdots + \left(Z_A - Z'_A\right)^2 + \cdots + \left(Z_T - Z'_T\right)^2}. \quad (6)$$

From (1),

$$\begin{aligned}
&\left(X_A - X'_A\right)^2 \\
&= \frac{\cos^2\left(2\pi i_1/\left(n_1+1\right)\right) + \cdots + \cos^2\left(2\pi i_s/\left(n_1+1\right)\right)}{{n_1}^2} \\
&+ \frac{2\sum_{p,q\in I, p\neq q}\cos\left(2\pi p/\left(n_1+1\right)\right)\cos\left(2\pi q/\left(n_1+1\right)\right)}{{n_1}^2} \\
&+ \frac{\cos^2\left(2\pi k_1/\left(n_2+1\right)\right) + \cdots + \cos^2\left(2\pi k_t/\left(n_2+1\right)\right)}{{n_2}^2} \\
&+ \frac{2\sum_{p',q'\in K, p'\neq q'}\cos\left(2\pi p'/\left(n_2+1\right)\right)\cos\left(2\pi q'/\left(n_2+1\right)\right)}{{n_2}^2} \\
&- \frac{2\sum_{l\in I, l'\in K}\cos\left(2\pi l/\left(n_1+1\right)\right)\cos\left(2\pi l'/\left(n_2+1\right)\right)}{n_1 n_2}, \\
&\left(Y_A - Y'_A\right)^2 \\
&= \frac{\sin^2\left(2\pi i_1/\left(n_1+1\right)\right) + \cdots + \sin^2\left(2\pi i_s/\left(n_1+1\right)\right)}{{n_1}^2} \\
&+ \frac{2\sum_{p,q\in I, p\neq q}\sin\left(2\pi p/\left(n_1+1\right)\right)\sin\left(2\pi q/\left(n_1+1\right)\right)}{{n_1}^2} \\
&+ \frac{\sin^2\left(2\pi k_1/\left(n_2+1\right)\right) + \cdots + \sin^2\left(2\pi k_t/\left(n_2+1\right)\right)}{{n_2}^2} \\
&+ \frac{2\sum_{p',q'\in K, p'\neq q'}\sin\left(2\pi p'/\left(n_2+1\right)\right)\sin\left(2\pi q'/\left(n_2+1\right)\right)}{{n_2}^2} \\
&- \frac{2\sum_{l\in I, l'\in K}\sin\left(2\pi l/\left(n_1+1\right)\right)\sin\left(2\pi l'/\left(n_2+1\right)\right)}{n_1 n_2}, \\
&\left(Z_A - Z'_A\right)^2 = \left[\frac{i_1 + \cdots + i_s}{n_1\left(n_1+1\right)}\right]^2 + \left[\frac{k_1 + \cdots + k_t}{n_2\left(n_2+1\right)}\right]^2 \\
&- 2\left[\frac{i_1 + \cdots + i_s}{n_1\left(n_1+1\right)}\right]\left[\frac{k_1 + \cdots + k_t}{n_2\left(n_2+1\right)}\right].
\end{aligned} \quad (7)$$

Then

$$\begin{aligned}
&\left(X_{\mathrm{A}}-X'_{\mathrm{A}}\right)^2+\left(Y_{\mathrm{A}}-Y'_{\mathrm{A}}\right)^2+\left(Z_{\mathrm{A}}-Z'_{\mathrm{A}}\right)^2\\
&\quad=\frac{s+2\sum_{p,q\in I,p\neq q}\cos\left(2\pi\left((p-q)/(n_1+1)\right)\right)}{{n_1}^2}\\
&\quad+\frac{t+2\sum_{p',q'\in K,p'\neq q'}\cos\left(2\pi\left(\left(p'-q'\right)/(n_2+1)\right)\right)}{{n_2}^2}\\
&\quad-\frac{2\sum_{l\in I,l'\in K}\cos\left(2\pi\left(l/(n_1+1)-l'/(n_2+1)\right)\right)}{n_1n_2}\\
&\quad+\left[\frac{i_1+\cdots+i_s}{n_1(n_1+1)}\right]^2+\left[\frac{k_1+\cdots+k_t}{n_2(n_2+1)}\right]^2\\
&\quad-2\left[\frac{i_1+\cdots+i_s}{n_1(n_1+1)}\right]\left[\frac{k_1+\cdots+k_t}{n_2(n_2+1)}\right].
\end{aligned}\tag{8}$$

Here $I=\{i_1,\ldots,i_s\}$, $K=\{k_1,\ldots,k_t\}$. Note that $(X_{\mathrm{A}}-X'_{\mathrm{A}})^2+(Y_{\mathrm{A}}-Y'_{\mathrm{A}})^2+(Z_{\mathrm{A}}-Z'_{\mathrm{A}})^2$ is only determined by the sets I and K (the sets of positions of base A, resp., in sequences G_1 and G_2) and regardless of the map order of base A. This observation is also valid for other base pairs. In conclusion, the Euclidean distance between $\mathbf{V}_1$ and $\mathbf{V}_2$ is invariable no matter what map order of four bases. □

4. Conclusions

CHC, based on a novel one-to-one mapping from nucleic bases in a DNA sequence to the points in 3D space, characterizes graphically a DNA sequence and reflects base composition and distribution of the sequence. As a consequence, DNA comparison with identical or different lengths intuitively transforms into CHC comparison whether in the normal order or in the reversed order.

The 12-dimensional vector extracted from CHC, as a numerical characterization of CHC, captures the essence of the base composition and distribution in DNA sequence, avoids the trouble of different sequence lengths, and allows quantitative estimates of the degree of similarity or dissimilarity among different DNAs. Reasonable phylogenetic analyses of four experiments illustrate that CHC technique is an effective tool for investigating biological structure and inferring evolutionary relationship. We expect that the presented method could help us explore more information hidden in the biological sequences.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (no. 11201409), Natural Science Foundation of Hebei Province (no. A2013203009), and Young Talents Plan of Higher School in Hebei Province (no. BJ2014060).

References

[1] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.

[2] E. Hamori, "Novel DNA sequence representations," *Nature*, vol. 314, no. 6012, pp. 585–586, 1985.

[3] E. Hamori and G. Varga, "DNA sequence (H) curves of the human immunodeficiency virus 1 and some related viral genomes," *DNA*, vol. 7, no. 5, pp. 371–378, 1988.

[4] E. Hamori, "Graphic representation of long DNA sequences by the method of H curves—current results and future aspects," *BioTechniques*, vol. 7, no. 7, pp. 710–720, 1989.

[5] M. Randić, M. Vračko, 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, no. 5, pp. 1235–1244, 2000.

[6] C.-T. Zhang, R. Zhang, and H.-Y. Ou, "The Z curve database: a graphic representation of genome sequences," *Bioinformatics*, vol. 19, no. 5, pp. 593–599, 2003.

[7] X. C. Tang, P. P. Zhou, and W. Y. Qiu, "On the similarity/dissimilarity of DNA sequences based on 4D graphical representation," *Chinese Science Bulletin*, vol. 55, no. 8, pp. 701–704, 2010.

[8] C. Yuan, B. Liao, and T.-M. Wang, "New 3D graphical representation of DNA sequences and their numerical characterization," *Chemical Physics Letters*, vol. 379, no. 5-6, pp. 412–417, 2003.

[9] C. Li and J. Wang, "On a 3-D representation of DNA primary sequences," *Combinatorial Chemistry and High Throughput Screening*, vol. 7, no. 1, pp. 23–27, 2004.

[10] Y.-H. Yao, X.-Y. Nan, and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation," *Chemical Physics Letters*, vol. 411, no. 1–3, pp. 248–255, 2005.

[11] B. Liao and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation," *Chemical Physics Letters*, vol. 388, no. 1–3, pp. 195–200, 2004.

[12] G. Huang, H. Zhou, Y. Li, and L. Xu, "Alignment-free comparison of genome sequences by a new numerical characterization," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 107–112, 2011.

[13] G. Huang, B. Liao, Y. Li, and Y. Yu, "Similarity studies of DNA sequences based on a new 2D graphical representation," *Biophysical Chemistry*, vol. 143, no. 1-2, pp. 55–59, 2009.

[14] G. Huang, B. Liao, Y. Li, and Z. Liu, "H–L curve: a novel 2D graphical representation for DNA sequences," *Chemical Physics Letters*, vol. 462, no. 1–3, pp. 129–132, 2008.

[15] I. Pesek and J. Zerovnik, "A numerical characterization of modified Hamori curve representation of DNA sequences," *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 60, no. 2, pp. 301–312, 2008.

[16] G. Xie and Z. Mo, "Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications," *Journal of Theoretical Biology*, vol. 269, no. 1, pp. 123–130, 2011.

[17] Z.-H. Qi and T.-R. Fan, "PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization," *Chemical Physics Letters*, vol. 442, no. 4–6, pp. 434–440, 2007.

[18] 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, no. 4, pp. 681–690, 2007.

[19] Z. Cao, B. Liao, and R. Li, "A group of 3D graphical representation of DNA sequences based on dual nucleotides," *International Journal of Quantum Chemistry*, vol. 108, no. 9, pp. 1485–1490, 2008.

[20] J. F. Yu, J. H. Wang, and X. Sun, "Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical presentation," *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 63, pp. 493–512, 2010.

[21] J.-F. Yu, X. Sun, and J.-H. Wang, "TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications," *Journal of Theoretical Biology*, vol. 261, no. 3, pp. 459–468, 2009.

[22] N. Jafarzadeh and A. Iranmanesh, "C-curve: a novel 3D graphical representation of DNA sequence based on codons," *Mathematical Biosciences*, vol. 241, no. 2, pp. 217–224, 2013.

[23] L. W. Parfrey, J. Grant, Y. I. Tekle et al., "Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life," *Systematic Biology*, vol. 59, no. 5, pp. 518–533, 2010.

[24] A. Ivan, M. S. Halfon, and S. Sinha, "Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs," *Genome Biology*, vol. 9, no. 1, article R22, 2008.

[25] S. Zou, L. Wang, and J. Wang, "A 2D graphical representation of the sequences of DNA based on triplets and its application," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2014, no. 1, article 1, 2014.

[26] B. Liao and T.-M. Wang, "New 2D graphical representation of DNA sequences," *Journal of Computational Chemistry*, vol. 25, no. 11, pp. 1364–1368, 2004.

[27] Z. Liu, B. Liao, W. Zhu, and G. Huang, "A 2D graphical representation of DNA sequence based on dual nucleotides and its application," *International Journal of Quantum Chemistry*, vol. 109, no. 5, pp. 948–958, 2009.

[28] X. Yang and T. Wang, "Linear regression model of short *k*-word: a similarity distance suitable for biological sequences with various lengths," *Journal of Theoretical Biology*, vol. 337, pp. 61–70, 2013.

[29] L. Lu, C. Li, and C. H. Hagedorn, "Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis," *Reviews in Medical Virology*, vol. 16, no. 1, pp. 5–36, 2006.

[30] Z. Liu, J. Meng, and X. Sun, "A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping," *Biochemical and Biophysical Research Communications*, vol. 368, no. 2, pp. 223–230, 2008.

[31] G. Chang, H. Wang, and T. Zhang, "A novel alignment-free method for whole genome analysis: application to HIV-1 subtyping and HEV genotyping," *Information Sciences*, vol. 279, pp. 776–784, 2014.

[32] W. Hou, Q. Pan, and M. He, "A novel representation of DNA sequence based on CMI coding," *Physica A: Statistical Mechanics and its Applications*, vol. 409, pp. 87–96, 2014.

[33] X. Zheng, C. Li, and J. Wang, "A complexity-based measure and its application to phylogenetic analysis," *Journal of Mathematical Chemistry*, vol. 46, no. 4, pp. 1149–1157, 2009.

[34] Y. Huang and T. Wang, "Phylogenetic analysis of DNA sequences with a novel characteristic vector," *Journal of Mathematical Chemistry*, vol. 49, no. 8, pp. 1479–1492, 2011.