Similarity/Dissimilarity of DNA Sequences Based on a New Condensed Curve Representation

Qianjun Xiao

***Abstract*—Based on a 3-D graphical representation, Bo Liao *et al.* [B. Liao *et al.*, J. Molec. Struct. (THEOCHEM) 717 (2005) 199] made a comparison for the coding sequences of the first exon of β-globin gene of 11 different species. However, some results in the Tables IV of Liao's were somewhat rational because the main information focus on the cumulative occurrence numbers S_i of base A, G, C, T. In this paper, we propose another 3D graphical representation by converting the S_i into $1\text{-}1/S_i$. Based on the mathematic invariants S^2, the results of comparison for the coding sequences used in Liao's are improved greatly and the examination of similarities among the full coding sequences shows our graphical representation method is more effective to the comparative study of DNA sequences. Furthermore, our graphical curves are compact and the complexities of computation are very small especially for long sequences.**

***Index Terms*—DNA Sequences, graphical representation, numerical characterization, S^2, similarity.**

I. Introduction

In paper [1], Bo Liao *et al.* made a comparison for the first exon of $\beta-$globin genes sequences belonging to 11 different species (see Table I) based on a 3D condensed curve representation called ***L***-curve. The ***L***-curve was obtained by the ways as follows. Firstly assigned four vectors (-1, 0, 0), (1, 0, 0), (0, -1, 0) and (0, 1, 0) to the four nucleic acid bases. Although the four bases A, G, C and T can be assigned in 4!=24 ways, however based on the classifications of bases of DNA [purine (A, G)/pyrimidine (C, T), amino (A, C)/keto (G, T) and week-H bond (A, T)/strong-H band (G, C)], only three characteristic curves corresponding to the three classifications can be obtained. In other words, Bo Liao *et al.* called the curve obtained by assigning one nucleic base as follows: (-1, 0, 0)→A, (1, 0, 0)→G, (0, -1, 0)→T, (0, 1, 0)→C as *ATGC*–curve. If assigned one nucleic base as follows: (-1, 0, 0)→A, (0, 1, 0)→G, (0, -1, 0)→T, (1, 0, 0)→C, the corresponding curve was called as *ATGC*–curve. If assigned one nucleic base as follows: (-1, 0, 0)→A, (0, -1, 0)→G, (1, 0, 0)→T, (0, 1, 0)→C, the corresponding curve was called as *AGTC*–curve. Each characteristic curve may be regarded as a coarse grained description of the DNA primary sequence. Then Bo Liao *et al.* construct a 3-component vector consisting of the leading eigenvalues the *L*/*L* matrices. The similarities (see Table IV in Ref. [1]) were computed by calculating the Euclidean distance between the end point of the vectors. Compared with other's results in recent papers [2]-[34], we find the result in Liao's work was imperfect because the conclusion that Opossum and Gallus were dissimilar from other species couldn't be derived. On the contrary, Goat-Gallus, Lemur-Gallus Rabbit-Gallus, Opossum-chimpanzee had smaller entries which didn't coincide with the real evolutionary relationship.

In this paper, we propose another 3D graphical representation. Based on the invariant S^2 which are sensitive to the characteristic curves, we construct an 9D vector, then make a comparison for the first exon of $\beta-$globin genes sequences belonging to 11 different species which were most used [1]-[34] (see Table I) . We find our result coincide with the result of similarity reported in others' works [2]-[34]. That's to say, our method avoid the deficiency which appear in Liao's work. Moreover, compared with others' work [1]-[5], [8]-[14], [19]-[28], our graphical curves are compact, that's to say, our graphical curves do not take up much room, and the complexities of computation is very small. Furthermore, the examination of similarities among the full coding sequences of $\beta-$globin gene of different species (see Table I) which were studied by Li *et al.* [15] and Guo *et al.* [29] and Li *et al.* [30] and Qi *et al.* [31] and Huang *et al.* [32] and Guo *et al.*[33] and Qi *et al.* [34] shows our graphical representation method is more effective to the comparative study of DNA sequences.

II. Materials and Method

A. Materials

It is well-known that the biological sequences have some chemical and structural properties, thus these attributes are usually used to study biological sequences. For DNA sequences, the four elements (A, T, G, C) can be classified into three groups: (1) purine group R=(A, G)/pyrimidine group Y=(C, T); (2) ketone group M=(A, C)/amino group K=(G, T); and (3) weak hydrogen bonds group W(A, T)/strong hydrogen bonds group S=(C, G). From the view of the above characteristics, the different condensed curve will be derived based on different assignment according to chemical and structural properties of nucleotide bases.

In this paper, in order to verify the proposed method, the most used DNA sequences are reused [1]-[34], which are the coding sequences of the first exon and the full coding sequences of $\beta-$globin gene of 11 different species shown in the Table I.

Manuscript received May 9, 2014; revised June 26, 2014. This work was supported in part by the Scientific Research Fund of Hunan Provincial Education Department (10C0185).

Qianjun Xiao is now with Hunan Vocational Institute of Technology, Xiangtan 411104, China (tel.: +86-731-52720616; e-mail: xqjxt@126.com, qjxiao1978@126.com).

DOI: 10.7763/IJMLC.2014.V4.469

TABLE I: β -GLOBIN GENES OF 11 SPECIES

Species	DB	ID	Location	Length(bp)	Location of each exon
Human	NCBI	U01317	62187-63610	1424	62187…62278, 62409…62631, 63482…63610
Chimpanzee	NCBI	X02345	4189-5532	1344	4189…4293, 4412…4633, 5484…5532
Gorilla	NCBI	X61109	4538-5881	1344	4538…4630, 4716…4982, 5833…5881
Lemur	NCBI	M15734	154-1595	1442	154…245, 376…598, 1467…1595
Rat	NCBI	X06701	310-1505	1196	310…401, 517…739, 1377…1505
Mouse	NCBI	V00722	275-1462	1188	275…367, 484…705, 1334…1462
Goat	NCBI	M15387	279-1749	1471	279…364, 493…715, 1621…1749
Bovine	NCBI	X00376	278-1741	1464	278…363, 492…714, 1613…1741
Rabbit	NCBI	V00882	277-1419	1143	277…368, 495…717, 1291…1419
Opossum	NCBI	J03643	467-2488	2022	467…558, 672…894, 2360…2488
Gallus	NCBI	V00409	465-1810	1346	465…556, 649…871, 1682…1810

B. Condensed Curve Representation of DNA Sequences

First, assign each nucleic base as follow:(-1, 0, 0)→A; (+1, 0, 0)→G; (0, -1, 0)→T; (0, +1, 0)→C. Then, the corresponding curve extends along with *z* axes. In detail, let $S = s_1 s_2 \cdots s_n$ be an arbitrary DNA sequence with length *n*. Then a map Φ which maps *S* into a plot set is defined as follows.

Explicitly, $\Phi(S) = \Phi_j(s_1)\Phi_j(s_2)\cdots\Phi_j(s_n)$, $j = 1,2,3$, where

$$\Phi(s_i) = \begin{cases} (-1,0,1-1/A_i) & \text{if} \quad s_i = A \\ (0,-1,1-1/G_i) & \text{if} \quad s_i = G \\ (+1,0,1-1/T_i) & \text{if} \quad s_i = T \\ (0,+1,1-1/C_i) & \text{if} \quad s_i = C \end{cases}$$

$$\Phi(s_i) = \begin{cases} (-1,0,1-1/A_i) & \text{if} \quad s_i = A \\ (+1,0,1-1/G_i) & \text{if} \quad s_i = G \\ (0,-1,1-1/T_i) & \text{if} \quad s_i = T \\ (0,+1,1-1/C_i) & \text{if} \quad s_i = C \end{cases}$$

$$\Phi(s_i) = \begin{cases} (-1,0,1-1/A_i) & \text{if} \quad s_i = A \\ (0,+1,1-1/G_i) & \text{if} \quad s_i = G \\ (0,-1,1-1/T_i) & \text{if} \quad s_i = T \\ (+1,0,1-1/C_i) & \text{if} \quad s_i = C \end{cases}$$

where A_i is are the cumulative occurrence numbers of A in the subsequence from the first base to the *i*-th base in the sequence, and it is similarly to G_i, C_i and T_i.

In other word, a DNA sequence is reduced into characteristic plot set $P_1, P_2, \cdots, P_n$. The curve connecting all plots of the characteristic plot set in turn is called characteristic curve. Based on classification of four bases of DNA sequences mentioned in Section 2.1, only three characteristic curves for one DNA sequence corresponding to the three classification of nucleotide bases can be outlined, which named with *ATGC*-curve, *ATCG*-curve, *AGTC*-curve obtained by map Φ_1, Φ_2 and Φ_3 based on three pattern *ATGC, ATCG* and *AGTC* respectively. In Fig. 1, one characteristic curve representing the sequence ATGGTGCACC based on *ATGC* pattern is given.

C. Numerical Characterization of DNA Sequence

In this section, the numerical characterization of graphical curve which facilitates quantitative comparisons of sequences is introduced. That's to say, a mathematical invariants is used to characterize graphical curve. According to the graphical representation of DNA sequences, each DNA sequence can be represented by a set of vector points in Cartesian coordinates. This invariant sensitive to the graphical curve, S^2 is defined as follows.

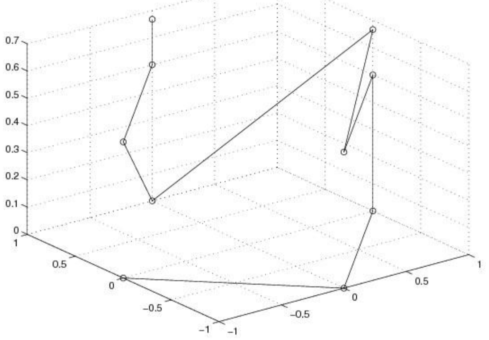

Fig. 1. Characteristic curves based the pattern ATGC for the sequence ATGGTGCACC.

For any sequence, there is a set of points $(x_i, y_i, z_i), i = 1, 2, \cdots, n$ where *n* is the length of the sequence. Similar to [7], $\bar{x}, \bar{y}, \bar{z}$ may be calculated as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

Then the mathematical formula for the computing S^2 is as follows:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$
$$S_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

III. RESULT

Comparison of similarities/dissimilarities is the essential motivation of graphical representation, which is reflected in recently published papers [1]-[34]. Here we also illustrate the use of our quantitative characterization of the DNA sequences with an examination of similarities/dissimilarities among 11 species in the Table I. A direct comparison of these sequences using computer codes is somewhat less straightforward due to the fact that these sequences have different lengths. In this paper, we represent the sequences with corresponding 9-component vectors then different lengths of the sequences do not cause difficulties.

The analysis of similarity among DNA sequences represented by the 9-component vectors is based on the assumption that two DNA sequences are similar if the corresponding 9-component vectors point to a similar direction in the 9D space and have similar magnitudes. The similarity/dissimilarity between these two vectors can be measured by calculating the correlation angle of two vectors, too. That is to say, the smaller are the correlation angles, the more similar are the two DNA sequences.

First, in Table II, based on method in Section II-B, we list the values of S^2 of the coding sequence of the first exon of $\beta-$globin gene of 11 species in Table I corresponding to patterns ATGC, ATCG and AGTC respectively. Then we construct a 9-component vector consisting S^2 to represent a DNA sequence. In Table III, we list the similarity/dissimilarity matrix for the coding sequence of the first exon of $\beta-$globin gene of 11 species in Table I based on the correlation angle of two 9-component vectors. Observing from Table III, we find Gallus (the only nonmammal among them) is very dissimilar to others among the 11 species because the corresponding rows have larger entries and Opossum (the most remote species from the remaining mammals) is very dissimilar to other 8 species among the 11 species because the corresponding nodes in corresponding rows have large entries. On the other hand, Human-Chimpanzee, Gorilla-Human, Human-Gorilla, and Bovine-goat have smaller entries, so they are more similar species pairs. This is not an accident, but shows they have close evolutionary relationship. Furthermore, our approach also can be applied to long sequences. In Table IV, we give the similarity/dissimilarity matrix for the full coding sequences of $\beta-$globin genes belonging to 11 different species in Table I based on the correlation angle of two 9-component vectors. Form Table IV, we can find the same result mentioned above. On the basis of these findings we conclude that the presented 3D graphical representation of DNA have apparently captured important features of the DNA sequences considered.

TABLE II: THREE S^2 OF 3D CURVES OF THE CODING SEQUENCE OF THE FIRST EXON OF β -GLOBIN GENE OF 11 SPECIES IN TABLE I

Species	ATGC(S_x^2,S_y^2,S_z^2)	ATCG(S_x^2,S_y^2,S_z^2)	AGTC(S_x^2,S_y^2,S_z^2)
Bovine	0.567442, 0.399453, 0.046445	0.388098, 0.583995, 0.046445	0.411628, 0.550616, 0.046445
Chimpanzee	0.546154, 0.421612, 0.040340	0.384615, 0.598535, 0.040340	0.421612, 0.546154, 0.040340
Gallus	0.555542, 0.418896, 0.044243	0.469541, 0.495342, 0.044243	0.371715, 0.625418, 0.044243
Goat	0.567442, 0.400000, 0.046436	0.400000, 0.567442, 0.046436	0.400000, 0.567442, 0.046436
Gorilla	0.540206, 0.423796, 0.043966	0.390837, 0.585788, 0.043966	0.401122, 0.570827, 0.043966
Human	0.532728, 0.439083, 0.044202	0.395127, 0.591973, 0.044202	0.415671, 0.562828, 0.044202
Lemur	0.562828, 0.409938, 0.044279	0.372725, 0.620162, 0.044279	0.459627, 0.501672, 0.044279
Mouse	0.515328, 0.461336, 0.043481	0.396820, 0.599062, 0.043481	0.425989, 0.558225, 0.043481
Opossum	0.541806, 0.461061, 0.043895	0.450430, 0.554587, 0.043895	0.472408, 0.528786, 0.043895
Rabbit	0.556804, 0.402497, 0.045076	0.370662, 0.604370, 0.045076	0.414607, 0.540449, 0.045076
Rat	0.562231, 0.427496, 0.044077	0.417105, 0.576206, 0.044077	0.450430, 0.533564, 0.044077

TABLE III: SIMILARITY/DISSIMILARITY MATRIX FOR THE CODING SEQUENCE OF THE FIRST EXON OF β -GLOBIN GENE OF 11 SPECIES IN TABLE I BASED ON THE CORRELATION ANGLE OF TWO 9-COMPONENT VECTORS

Species	Chimpanzee	Gallus	Goat	Gorilla	Human	Lemur	Mouse	Opossum	Rabbit	Rat
Bovine	0.0308	0.1226	0.0240	0.0359	0.0449	0.0657	0.0682	0.0917	0.0256	0.0464
Chimpanzee		0.1347	0.0464	0.0299	0.0255	0.0553	0.0429	0.0823	0.0234	0.0425
Gallus			0.0995	0.1118	0.1210	0.1809	0.1345	0.1291	0.1449	0.1270
Goat				0.0347	0.0490	0.0886	0.0737	0.0934	0.0476	0.0553
Gorilla					0.0196	0.0842	0.0438	0.0899	0.0419	0.0573
Human						0.0771	0.0252	0.0776	0.0452	0.0501
Lemur							0.0831	0.0968	0.0504	0.0598
Mouse								0.0747	0.0636	0.0605
Opossum									0.1002	0.0482
Rabbit										0.0550

TABLE IV: SIMILARITY/DISSIMILARITY MATRIX FOR THE CODING SEQUENCES OF β -GLOBIN GENES BELONG TO 11 DIFFERENT SPECIES IN TABLE I BASED ON THE CORRELATION ANGLE OF TWO 9-COMPONENT VECTORS

Species	Chimpanzee	Gallus	Goat	Gorilla	Human	Lemur	Mouse	Opossum	Rabbit	Rat
Bovine	0.0723	0.1451	0.0075	0.0744	0.0633	0.0537	0.0475	0.0642	0.0466	0.0362
Chimpanzee		0.2167	0.0755	0.0063	0.0120	0.0388	0.0628	0.0391	0.0466	0.0545
Gallus			0.1417	0.2191	0.2081	0.1897	0.1718	0.2042	0.1810	0.1704
Goat				0.0781	0.0669	0.0537	0.0442	0.0682	0.0451	0.0348
Gorilla					0.0128	0.0440	0.0675	0.0380	0.0518	0.0584
Human						0.0390	0.0566	0.0287	0.0422	0.0038
Lemur							0.0433	0.0607	0.0241	0.0598
Mouse								0.0582	0.0196	0.0150
Opossum									0.0527	0.0478
Rabbit										0.0174

IV. Discussions

In Table V, in order to prove the utility of our approach, we list the recently reported results of the examinations of the degree of similarity/dissimilarity of the coding sequences of the first exon of β-globin gene of several species with the coding sequence of the first exon of the human β-globin gene by means of approaches using alternative DNA sequence descriptors [1]-[3], [7], [13], [16]-[19]. The alternative descriptors used are: the 3-component vectors of the leading eigenvalues of the L/L matrices [1], the 12-component vector whose components are made up of the normalized leading eigenvalues of the L/L matrices associated with DNA [2], the 15-component vectors consisting of the average bandwidths of a DNA sequence [3], the 9-component vectors of the geometrical centers [7], the 6-component vectors of the normalized leading eigenvalues of the L/L matrices [13], the 16-component vector whose components are made up of frequency of occurrence of all possible ordered pairs of adjacent bases [16], the 12-component vector made by using the leading eigenvalues of 12(4×4) matrices obtained from the cubic (4×4×4) matrix with elements denoting the frequencies of occurrence of all 64 possible triplets in a DNA sequence [17], the 6-component vector whose components are composed of the leading eigenvalues of six (2×2) condensed matrices associated with a DNA sequence[18], the 4-vectors of the normalized leading eigenvalues of the L/L matrices [19]. In order to visualize Table V directly, we transform Table V to Fig. 2. As one can see from Fig. 2, there exists an overall qualitative agreement among similarities based on different descriptors except Liao's work [1], despite some variations among them. As for the reason, we think it was main information focus on the cumulative occurrence numbers of base A, G, C, T, because the value of z were more bigger than the value of x and y in Liao's work, that's to say, their 3D graphical representation of DNA sequences lose some biological information. This fact illustrates the utility of our approach.

TABLE V: The Comparison Similarity for Coding Sequences of the First Exon of β- Globin Gene between Human and the Other Six Species Based on Our Method and Others' Method

Species	A^a	B^b	C^c	D^d	E^e
Goat	0.04904	0.7441	0.061	0.10836	0.08696
Gallus	0.12093	0.1792	0.109	0.06044	0.17658
Opossum	0.0776	1.4439	0.148	0.13045	0.15976
Lemur	0.07706	0.2504	0.087	0.13417	0.10657
Rabbit	0.04515	0.4943	0.042	0.06773	0.08573
Rat	0.05011	0.598	0.043	0.06643	0.08135

Species	F^f	G^g	H^h	I^i	J^j
Goat	0.1356	6.928	4.996	4.375	0.0128
Gallus	0.1727	9.592	5.015	7.8	0.0368
Opossum	0.1408	8.602	4.491	7.183	0.0289
Lemur	0.0774	7.483	2.97	4.804	0.0155
Rabbit	0.0742	6.083	3.171	4.644	0.0133
Rat	0.0965	6.325	4.857	3.85	0.0138

A^a:	This work, Table 3.	B^b:	From Ref. 1, Table 4.
C^c:	From Ref. 2, Table 2.	D^d:	From Ref. 3, Table 7.
E^e:	From Ref. 7, Table 4.	F^f:	From Ref.13, Table 3.
G^g:	From Ref.16, Table 6.	H^h:	From Ref.17, Table 12.
I^i:	From Ref.18, Table 11.	J^j:	From Ref.19, Table 10.

Furthermore, from Table IV, we surprisingly find that Rabbit-Mouse, Rabbit-Rat and Rat-Mouse have smaller entries which didn't appear in [15], [29]-[34]. But this result coincided with result of phylogenetic tree [35]-[39] (Cao *et al.*, 1997, 1998; Li *et al.*, 2001; Madsen *et al.*, 2001; Murphy *et al.*, 2001; Reyes *et al.*, 2000; Otu *et al.*, 2003). So we think the graphical representation method can capture the main information on the interspecies similarity of β-globin gene contained in the first exon. But if we want to derive more information on the interspecies similarity of β-globin gene, we ought to consider the full coding sequences β-globin. Therefore, we think our graphical representation method is more effective to the comparative study of DNA sequences.

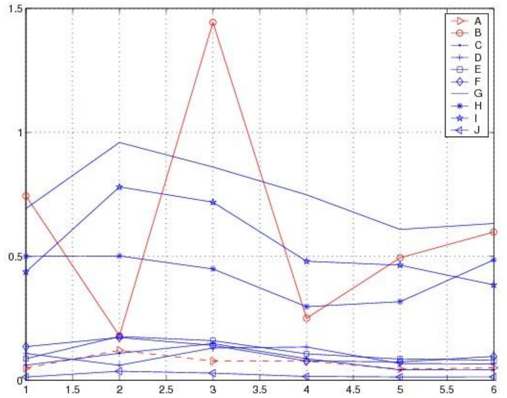

Fig. 2. The comparison degree of similarity between Human and the other six species. The value of *y*-coordinate denotes the relative distance. *i* of *x*-coordinate denotes the species in Table V (*x*-coord 1: Goat, *x*-coord 2: Gallus, *x*-coord 3: Opossum, *x*-coord 4: Lemur, *x*-coord 5: rabbit, and *x*-coord 6: rat). A denotes this paper work, B, C, D, E, F, G, H, I and J denotes the Ref.' work mentioned in Table V.

V. Conclusions

In this paper, we outline another 3D graphical representation of DNA sequences without degeneracy according to chemical structures of the bases, and presented a similarity measure among DNA sequences. The representation considers not only strings structure but also chemical structure for DNA sequences, reflects distribution of the base pairs, reduces the loss of information in the transfer of data from DNA sequence to its graphical representation and allows numerical characterization. Based on this representation, we propose a numerical characterization approach by constructing a 9-component vector whose components are mathematic invariants, S^2, associated with the DNA sequences rather than strings sequence themselves. We make a comparison for the first exon β-globin genes sequences belong to 11 different species based on our method. We find our result coincide with the results of similarity analysis reported in other's works [2]-[34]. The examination of similarities among the coding sequences of β-globin gene of different species illustrates the utility of the approach. In a word, our approach has some advantages:

1) The graphical representation we introduced is simple and direct; it can uniquely represent a DNA sequence, helping in recognizing major similarities among different DNA sequences without sequence alignment.
2) By the comparison with Liao's work, our method avoids the deficiency which appears in Liao's work, that's to say,

our method is more rational.

3) A mathematic invariant, S^2, is proposed, then we can construct a 9-component vector to represent a DNA sequence rather than strings sequence themselves.
4) Our graphical curves are compact and our approach also can be applied to long sequences such as complete sequences and their complexities of computation very small.

References

[1] B. Liao, Y. S. Zhang, K. Q. Ding, and T. M. Wang, "Analysis of similarity/ dissimilarity of DNA sequences based on a condensed curve representation," *Journal of Molecular Structure: THEOCHEM*, vol. 717, pp. 199-203, Jan. 2005.

[2] M. Randić, M. Vračko, N. Lerš, and D. Plavšić, "Analysis of similarity/ dissimilarity of DNA sequences based on novel 2-D graphical representation," *Chemical Physical Letter*, vol. 371, pp. 202-207, Jan. 2003.

[3] B. Liao and T. M. Wang, "Analysis of similarity of DNA sequences based on 3-D graphical representation," *Chemical Physical Letter*, vol. 388, pp. 195-200, March 2004.

[4] Y. H. Yao and T. M. Wang, "A class of new 2-D graphical representation of DNA sequences and their application," *Chemical Physical Letter*, vol. 398, pp. 318-323, Oct. 2004.

[5] R. Chi and K. Q. Ding, "Novel 4D numerical representation of DNA sequences," *Chemical Physical Letter*, vol. 407, pp. 63-67, April 2005.

[6] Q. Dai, X. Q. Liu, and T. M. Wang, "A novel 2D graphical representation of DNA sequences and its application," *Journal of Molecular Graphics and Modelling*, vol. 25, pp. 340-344, March 2006.

[7] B. Liao and K. Q. Ding, "A 3D graphical representation of DNA sequences and its application," *Theoretical Computer Science*, vol. 358, pp. 56-64, Jan. 2006.

[8] X. Q. Liu, Q. Dai, Z. L. Xiu, and T. M. Wang, "PNN-curve: A new 2D graphical representation of DNA sequences and its application," *Journal of Theoretical Biology*, vol. 243, pp. 555-561, July 2006.

[9] J. Wang and Y. Zhang, "Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation," *Chemical Physical Letter*, vol. 423, pp. 50-53, March 2006.

[10] Y. H. Yao, X. Y. Nan, and T. M. Wang, "A new 2D graphical representation Classification curve and the analysis of similarity/dissimilarity of DNA sequences," *Journal of Molecular Structure: THEOCHEM*, vol. 764, pp. 101-108, April 2006.

[11] Y. S. Zhang and W. Chen, "Invariants of DNA sequence based on 2DD-curves," *Journal of Theoretical Biology*, vol. 242, pp. 382-388, May 2006.

[12] Z. H. Qi and X. Q. Qi, "Novel 2D graphical representation of DNA sequence based on dual nucleotides," *Chemical Physical Letter*, vol. 440, pp. 139-144, March 2007.

[13] J. Song, "Analysis of similarity/dissimilarity of DNA sequences by a new 3D graphical representation," *Journal of Biological Systems*, vol. 15, No. 3, pp. 287-297, Jan. 2007.

[14] Y. S. Zhang and B. Liao, "On the Similarity of DNA sequences Based on 3-D Graphical Representation," *Journal of Biomathematic*, vol. 22, no. 4, pp. 583-590, March 2007.

[15] C. Li, X. Q. Yu, and N. Helal, "Similarity analysis of DNA sequences based on codon usage," *Chemical Physical Letter*, vol. 459, pp. 172-174, May 2008.

[16] M. Randić, "Condensed representation of DNA primary sequences," *Journal of chemical information and computer sciences*, vol. 40, pp. 50-56, Jan. 2000.

[17] M. Randić, X. F. Guo, and S. C. Basak, "On the characterization of DNA primary sequence by triplet of nucleic acid bases," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 619-626, March 2001.

[18] P. A. He and J. Wang, "Characteristic sequences for DNA primary sequence," *Journal of chemical information and computer sciences*, vol. 42, pp. 1080-1085, July 2002.

[19] 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 Physical Letter*, vol. 411, pp. 248-255, July 2005.

[20] M. Randić, M. Vračko, J. Zupan, and M. Novič, "Compact 2-D graphical representation of DNA," *Chemical Physical Letter*, vol. 373, pp. 558-562, April 2003.

[21] M. Randić, M. Vračko, N. Lerš, and D.Plavšić, "Novel 2-D graphical representation of DNA sequences and their numerical characterization," *Chemical Physical Letter*, vol. 368, pp. 1-6, Oct. 2003.

[22] M. Randić, J. Zupan, D. Vikić-Topić, and D. Plavšić, "A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences," *Chemical Physical Letter*, vol.; 431, pp. 375-379, Sep. 2006.

[23] M. Randić, "Another look at the chaos-game representation of DNA," *Chemical Physical Letter*, vol. 456, pp. 84-88, March 2008.

[24] B. Liao, M. S. Tan, and K. Q. Ding, "Application of 2-D graphical representation of DNA sequence," *Chemical Physical Letter*, vol. 414, pp. 296-300, Sep. 2005.

[25] B. Liao and T. M. Wang, "3-D graphical representation of DNA sequences and their numerical characterization," *Journal of Molecular Structure: THEOCHEM*, vol. 681, pp. 209-212, May 2004.

[26] 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, March 2010.

[27] G. S. Xie and Z. X. 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, PP. 123–130, March 2011.

[28] Y. S. Li, Y. F. Qin, X. Q. Zheng, and Y. Zhang, "Three-unit semicircles curve: A compact 3D graphical representation of DNA sequences based on classifications of nucleotides," *International Journal of Quantum Chemistry*, vol. 112, pp. 2330-2335, Feb. 2012.

[29] Y. Guo and T. M. Wang, "A new method to analyze the similarity of the DNA sequences," *Journal of Molecular Structure: THEOCHEM*, vol. 853, pp. 62-67, Dec. 2008.

[30] C. Li and J. Wang. "Similarity/dissimilarity of DNA sequences based on the generalized LZ complexity of (0, 1)-sequence," *Journal Mathematical Chemistry*, vol. 43, no. 1, pp. 26-31, Jan. 2008.

[31] 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, pp. 681-690, Sep. 2007.

[32] G. H. Huang, B. Liao, Y. F. Li, and Y. G. Yu, "Similarity studies of DNA sequences based on new 2D graphical representation," *Biophysical Chemistry*, vol. 143, pp. 55-59, April 2009.

[33] Y. Guo, Y. F. Wang, and S. L. Zhang, "A Novel Way to Numerically Characterize DNA Sequences and Its Application," *International Journal of Quantum Chemistry*, vol. 111, pp. 3971–3979, Nov. 2011.

[34] Z. H. Qi, L. Li, and X. Q. Qi, "Using Huffman coding method to visualize and analyze DNA sequences," *Journal of Computational Chemistry*, vol. 32, pp. 3233-3240, Aug. 2011.

[35] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Y. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149-154, Feb. 2001.

[36] O. Madsen, M. Scally, C. J. Douady, D . J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer, "Parallel adaptive radiations in two major clades of placental mammals," *Nature*, vol. 409, pp. 610-618, Feb. 2001.

[37] W. J. Murphy, E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer, "Resolution of the early placental mammal radiation using Bayesian phylogenetics," *Science*, vol. 294, no. 5550, pp. 2348-2351, Dec. 2001.

[38] A. Reyes, C. Gissi, G. Pesole, F. M. Catzeflis, and C. Saccone, "Where do rodents fit? Evidence from the compete mitochondrial genome of Sciurus vulgaris," *Molecular Biology and Evolution*, vol. 17, no. 6, pp. 979-983, June 2000.

[39] H. H. Out and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122-2130, Nov. 2003.

Qianjun Xiao received master degree in applied mathematics from School of Mathematics and Computational Science, Xiangtan University (XTU), Xiangtan, P.R. China, in July, 2008. He has been positioned as a lecturer since 2009 in Hunan Vocational Institute of Technology. And his current research interest areas include bioinformatics and teaching in mathematics.