MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **59** (2008) 625-634

ISSN 0340 - 6253

A New Approach to Molecular Phylogeny of Primate Mitochondrial DNA

Yusen Zhang [*], Wei Chen
Department of Mathematics, Shandong University at Weihai
Weihai 264209, China

(Received July 19, 2007)

Abstract. *In this paper a novel method for phylogenetic analysis of DNA sequence data has been proposed. At first we provide a new distance matrix of DNA sequence based on the 3DD-Curves and construct new similarities/dissimilarities matrix by using this distance matrix of DNA sequence. As an application, we constructed a phylogenetic tree for 11 species of primates, The phylogeny obtained is generally consistent with evolutionary trees constructed in previous studies.*

1 Introduction

In recent years several authors have presented various graphical representations of DNA sequences [1-15]. In order to numerically characterize DNA sequences, Many methods based on these graphical representations have been proposed and several applications have been made using these techniques. The main idea is to transform the graphic representation into another mathematical object. The selected mathematical object is described by various types of matrices that record distances among DNA sequences. The invariants of matrices, such as the leading eigenvalues, can be used as mathematical descriptors of the DNA sequences to quantitatively compare the sequences and determine similarities and dissimilarities between them [5][12][14][15]. Comparison of two DNA sequences is now transformed into a comparison of the corresponding sequences of mathematical descriptors of DNA. Hence it is very important to determine a suitable matrix to express the graphical representation. A widely used matrices of graphical representation are the distance/distance matrices (D/D matrices or E/G matrices) and the length/length matrices (L/L matrices E/P matrices) which are proposed by Randic [5]. These matrices are constructed by Euclidean distance (ED) between vertices and other distance between the same vertices [4-17]. In order to avoid discrepancies from the two matrices, most of authors use the leading eigenvalues of the distance/distance matrices, not both, to compare the sequences of the first exon of the DNA sequence of the beta globin gene from different species for their similarities and dissimilarities as application.

[*] Corresponding author:zhangys@sdu.edu.cn

However the result obtain differ in different graphical representation. In order to determine the phylogenetic relationships and divergence times between species, It is necessary to choose a suitable mathematical object. Based on this idea, a new mathematical object is proposed in this paper.

This study relies on our previous work of the weighted graphical representation of DNA sequence. It was shown that any DNA sequence can be described by a unique 3-dimensional space curve corresponding to it, the 3DD-Curve, which is the characteristic cure of the weighted graphical representation. It was demonstrated that there exists a one-to-one correspondence between the DNA sequence and the 3DD-Curve. Therefore, the 3DD-Curve is the representative of the DNA sequence. The 3DD-Curve contains all the information that the DNA sequence contains, or vice versa.

In this paper we (1) propose another distance matrix of DNA sequence based on the 3DD-Curves [14]; (2) construct the R/G matrix by using this RD matrix based on 3DD-Curves; (3) discuss the relative properties; (4) reconstruct phylogenetic tree of the primate mitochondrial DNA sequences for 11 different species.

2 Format of 3DD-Curve

Consider a DNA sequence read from the 5'- to the 3'-end with N bases. Inspect the sequence one base at a time. Let the number of steps be denoted by i, i.e. $i = 1, 2, ..., N$. In the i-th step, count the cumulative numbers of the bases A, C, G and T, denoted by the four positive integers A_i, C_i, G_i and T_i, respectively, occurring in the subsequence from the first to the i-th base in the DNA sequence inspected. The 3DD-Curve consists of a series of nodes $P_i(i = 1, 2, ..., N)$, whose coordinates are denoted by x_i,y_i and z_i. Owing to bases of DNA can be classified into groups, purine(A, G)/pyrimidine(C, T), amino(A, C)/keto(G, T) and week-bond(A, T)/strong-H band(G, C). Here we use the three 3DD-Curves corresponding to the three classifications are as follows:

1. The 3DD-Curve of DNA sequences based on pattern GCT is

$$\begin{cases} x_i = \sqrt{u}A_i - \sqrt{v}G_i \\ y_i = \sqrt{u}A_i + \sqrt{v}C_i \\ z_i = \sqrt{u}A_i - \sqrt{v}T_i \end{cases} \tag{1}$$

2. The 3DD-Curve of DNA sequences based on pattern CGT is

$$\begin{cases} x_i = \sqrt{u}A_i - \sqrt{v}C_i \\ y_i = \sqrt{u}A_i + \sqrt{v}G_i \\ z_i = \sqrt{u}A_i - \sqrt{v}T_i \end{cases} \tag{2}$$

3. The 3DD-Curve of DNA sequences based on pattern TGC is

$$\begin{cases} x_i = \sqrt{u}A_i - \sqrt{v}G_i \\ y_i = \sqrt{u}A_i + \sqrt{v}T_i \\ z_i = \sqrt{u}A_i - \sqrt{v}C_i \end{cases} \tag{3}$$

where u, v is different positive real numbers, but not perfect square number. We define $A_0 = C_0 = G_0 = T_0 = 0$ and thus $x_0 = y_0 = z_0 = 0$. By this way, we can reduce a DNA sequence into a series of nodes $P_0, P_1, P_2, ..., P_N$, whose coordinates x_i, y_i, z_i (i = 0, 1, 2, . . ., N, where N is the length of the DNA sequence.

By proceeding as we do in proof of property 1 [14], we can easily get

Property 1: For a given DNA sequence there is a unique 3DD-Curve corresponding to it.

Property 2: There is no circuit or degeneracy in 3DD-Curve.

We call the corresponding plot set be characteristic plot set. The curve connecting all plots of the characteristic plot set in turn is called 3DD-Curve (3D Curve of DNA).

Based on the 3DD-Curve, any DNA sequence can be uniquely described by three independent distributions, i.e., x_i, y_i and z_i. Therefore, the 3DD-Curve contains all the information that the corresponding DNA sequence carries. A DNA sequence can be analyzed by studying the corresponding 3DD-Curve. One of the advantages of the 3DD-Curve is its intuitiveness; the entire 3DD-Curve of a genome can be viewed on a computer screen, regardless of genome length, thus allowing both global and local compositional features of genomes to be easily grasped. By combining use of the 3DD-curve with statistical analysis, better results may be obtained.

From the construction of the 3DD-Curve, we can see that it can provide more information than existing 3D graphic representation by choosing the appropriate parameters and we can choose the system most appropriate to the problem at hand.

It is easy to see that different parameters can result in different visual clues to DNA sequence. So we should choose the 3DD-Curve most appropriate to the problem under consideration.

3 Distance matrices

For any sequence, we propose three 3DD-Curves base on different patterns to represent it. The points of the 3DD-Curve is denoted as (x_i, y_i, z_i), $i = 0, 1, 2, 3, ..., N$, where N is the length of the sequence, if we define:

$$\begin{cases} x_{0i} = (\sqrt{u} - \sqrt{v})i \\ y_{0i} = (\sqrt{u} + \sqrt{v})i \\ z_{0i} = (\sqrt{u} - \sqrt{v})i \end{cases} \tag{4}$$

where $i = 0, 1, 2, 3, ..., N$, then the line connecting all points (x_{0i}, y_{0i}, z_{0i}) in turn is similar for all three 3DD-Curves base on patterns CGT, GCT and TGC. We call it reference line.

Noting that

$$A_i + C_i + T_i + G_i = i,$$

where $i = 0, 1, 2, 3, ..., N$, we know that the reference line has deeply relations with the three 3DD-Curves. It reflects the tendency of bases distribution of DNA sequence.

Let $\rho = \sqrt{x_{0i}^2 + y_{0i}^2 + z_{0i}^2}$, then we have

$$\cos\alpha = \frac{x_{0i}}{\rho}, \cos\beta = \frac{y_{0i}}{\rho}, \cos\gamma = \frac{z_{0i}}{\rho}$$

.

then we have the following properties:

Property 3: The direct cosine of the reference line is

$$(\cos\alpha, \cos\beta, \cos\gamma) = (\frac{\sqrt{u}-\sqrt{v}}{\sqrt{3u+3v-2\sqrt{uv}}}, \frac{\sqrt{u}+\sqrt{v}}{\sqrt{3u+3v-2\sqrt{uv}}}, \frac{\sqrt{u}-\sqrt{v}}{\sqrt{3u+3v-2\sqrt{uv}}})$$

.

Let

$$X_i = x_i - x_{0i}, Y_i = y_i - y_{0i}, Z_i = z_i - z_{0i}, (i = 0, 1, 2, ..., N).$$

If the curve connecting all points (X_i, Y_i, Z_i) in turn is called relative curve of 3DD-Curve of DNA sequence, then we have:

Property 4: For a given DNA sequence, there is a unique relative curve with parameters u, v corresponding to it.

By this property, one can characterize a DNA sequence by its relative curve instead of its 3DD-Curve. That means one can associate with the relative curve a matrix and use the matrix invariants of relative curve to facilitate quantitative comparisons of DNA sequences.

Now, we define another quotient matrix R/G instead of the quotient matrix E/G. The (i,j) element $[R/G]_{ij}$ of matrix R/G is defined to be $[RD]_{ij}/|i-j|$. $[RD]_{ij}$ is the Euclidean distance between a pair of vertices of the relative curve, that is

$$[RD]_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2}, \tag{5}$$

where $i, j = 1, 2, ..., N$.

Then, by straight forward calculations, we can obtain,

Property 5: For any DNA sequence,

(1) the (i, j) element of matrix RD of 3DD-Curve based on pattern CGT is:

$$[RD]_{ij} = \sqrt{3uA'^2_{ij} + v(G'^2_{ij} + C'^2_{ij} + T'^2_{ij}) + 2\sqrt{uv}A'_{ij}(C'_{ij} - T'_{ij} - G'_{ij})},$$

(2) the (i, j) element of matrix RD of 3DD-Curve based on pattern GCT is:

$$[RD]_{ij} = \sqrt{3uA'^2_{ij} + v(G'^2_{ij} + C'^2_{ij} + T'^2_{ij}) + 2\sqrt{uv}A'_{ij}(G'_{ij} - T'_{ij} - C'_{ij})},$$

(3) the (i, j) element of matrix RD of 3DD-Curve based on pattern TGC is:

$$[RD]_{ij} = \sqrt{3uA'^2_{ij} + v(G'^2_{ij} + C'^2_{ij} + T'^2_{ij}) + 2\sqrt{uv}A'_{ij}(T'_{ij} - C'_{ij} - G'_{ij})},$$

where $A'_{ij} = A_i - A_j + j - i$, $C'_{ij} = C_i - C_j + j - i$, $G'_{ij} = G_i - G_j + j - i$ and $T'_{ij} = T_i - T_j + j - i$ and $i, j = 1, 2, ..., s$.

In order to make a comparison with previous works, we recall the quotient matrix E/G [13-17]. The (i, j) element $[E/G]_{ij}$ of matrix E/G is defined to be $[ED]_{ij}/|i-j|$, where $[ED]_{ij}$ is the Euclidean distance between a pair of vertices of 3DD-Curves. Then, we can have:

Property 6: For any DNA sequence,

(1) the (i, j) element of matrix RD of 3DD-Curve based on pattern CGT is:

$$[RD]_{ij} = \sqrt{3uA_{ij}^2 + v(G_{ij}^2 + C_{ij}^2 + T_{ij}^2) + 2\sqrt{uv}A_{ij}(C_{ij} - T_{ij} - G_{ij})},$$

(2) the (i, j) element of matrix RD of 3DD-Curve based on pattern GCT is:

$$[RD]_{ij} = \sqrt{3uA_{ij}^2 + v(G_{ij}^2 + C_{ij}^2 + T_{ij}^2) + 2\sqrt{uv}A_{ij}(G_{ij} - T_{ij} - C_{ij})},$$

(3) the (i, j) element of matrix RD of 3DD-Curve based on pattern TGC is:

$$[RD]_{ij} = \sqrt{3uA_{ij}^2 + v(G_{ij}^2 + C_{ij}^2 + T_{ij}^2) + 2\sqrt{uv}A_{ij}(T_{ij} - C_{ij} - G_{ij})},$$

where $A_{ij} = A_i - A_j$, $C_{ij} = C_i - C_j$, $G_{ij} = G_i - G_j$ and $T_{ij} = T_i - T_j$ and $i, j = 1, 2, ..., s$.

We also choose the leading eigenvalues of quotient matrices R/G and E/G as mathematical descriptors of DNA sequence. For a given DNA sequence, let $\lambda_1, \lambda_2, \lambda_3$, be the leading eigenvalue of matrix R/G (or E/G) of DNA sequence based on patterns CGT, GCT and TGC, respectively, we construct a 3-component vector

$$(\theta_1, \theta_2, \theta_3) = (\lambda_1/N, \lambda_2/N, \lambda_3/N),$$

where N is the number of bases making up the corresponding DNA sequence. Then we get a one-to-one correspondence between the DNA sequence and 3-component vectors $(\theta_1, \theta_2, \theta_3)$ of the R/G (or E/G). So $(\theta_1, \theta_2, \theta_3)$ of R/G (or E/G) can characterize the DNA sequences. Comparison between sequences becomes comparison between these 3-component vectors.

Let $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3}), i = 1, ..., s$, denote all 3-component vectors of the E/G from 3DD-Curves of s DNA sequences and $\beta_i = (\beta_{i1}, \beta_{i2}, \beta_{i3}), i = 1, ..., s$, denote that of R/G. The analysis of similarity/dissimilarity among these DNA sequences represented by the 3-component vectors is based on the assumption that two DNA sequences are similar if the corresponding 3-component vectors in the 3D-space have similar magnitudes.

The similarities/dissimilarities matrix can be formulated as the symmetric matrix M_e (M_r) whose (i,j) element is defined as:

$$[M_e]_{ij} = \sqrt{(\alpha_{i1} - \alpha_{j1})^2 + (\alpha_{i2} - \alpha_{j2})^2 + (\alpha_{i3} - \alpha_{j3})^2}, \quad (6)$$

$$[M_r]_{ij} = \sqrt{(\beta_{i1} - \beta_{j1})^2 + (\beta_{i2} - \beta_{j2})^2 + (\beta_{i3} - \beta_{j3})^2}, \qquad (7)$$

where $i, j = 1, 2, ..., s$.

4 Construction of the phylogenetic tree

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of DNA sequence analysis is phylogeny reconstruction for understanding evolutionary history of organisms.

Table 1: Database Source

Species	ID/ ACCESSION	Abbreviation	*length*(*bp*)	database
Saimiri sciureus	M22655	S. sci	893	*NCBI*
Hylobates	V00659	Hyl	896	*NCBI*
Lemur catta	M22657	Lemur	895	*NCBI*
Macaca fascicular	M22653	M. fas	896	*NCBI*
Gorilla	V00658	Gorilla	896	*NCBI*
Macaca fuscata	M22651	M. fus	896	*NCBI*
Macaca mulatta	M22650	M. mul	896	*NCBI*
Macaca sylvanus	M22654	M. syl	896	*NCBI*
Chimpanzee	V00672	Chi	896	*NCBI*
Orangutan	V00675	Ora	895	*NCBI*
Tarsius syrichta	M22656	T. syr	895	*NCBI*

Many different methods for phylogenetic analysis of DNA sequence data have been proposed and studied in the literature.

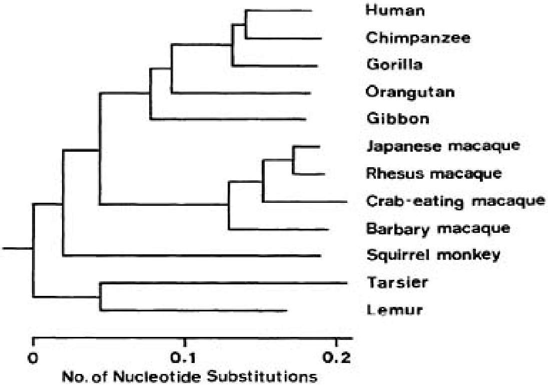

Figure 1: Phylogenetic tree based on No. of nucleotide substitutions

Hayasaka, Gojobori and Horai calculated the number of nucleotide substitutions for a given pair of species by the six-parameter method. Using the calculated numbers, they constructed a phylogenetic tree by the NJ method, the distance Wagner method and unweighed pair grouping method, respectively. the algorithms for constructing phylogenetic trees are different from each other. These three different methods give phylogenetic trees with the same topology, the phylogenetic relationships derived from these mtDNA sequence comparisons appear reliable. In Figure 1, we show the phylogenetic tree which was constructed in [18].

In table 1, we list the sequences of homologous 0.9-kb mtDNA fragments from seven species of primates (four old-world monkeys, a new-world monkey, and two prosimians) and the 0.9-kb mtDNA fragments from four hominoid species (chimpanzee, gorilla, orangutan and Hylobates) which are used by Hayasaka, Gojobori and Horai.

Table 2: The upper triangular part of the similarities/dissimilarities matrix M_r of the R/G of the 3DD-Curves

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0831	0.1645	0.5631	0.1984	0.2528	0.3847	0.1935	0.5759	0.7124	0.2456
Gorilla		0	0.1197	0.5807	0.2271	0.2481	0.4003	0.1174	0.5932	0.7296	0.2541
Hyl			0	0.4915	0.1727	0.1652	0.3196	0.1654	0.4978	0.6255	0.1742
Lemur				0	0.3680	0.3434	0.1861	0.6285	0.1231	0.1551	0.3266
M. Fas					0	0.1202	0.1992	0.3096	0.3783	0.5148	0.0761
M. Fus						0	0.1631	0.2855	0.3767	0.4941	0.0534
M. Syl							0	0.4477	0.2427	0.3405	0.1491
Ora								0	0.6515	0.7774	0.3111
S. Sci									0	0.1670	0.3491
T. Syr										0	0.4764
M. Mul											0

In this section we constructed a phylogenetic tree by comparing these sequences and determine the phylogenetic relationships for 11 different species which are listed in table 1.

We take $u = 57$ and $v = 29$ for 3DD-Curve of the primate mitochondrial DNA sequences, then we compute the similarities/dissimilarities matrix M_r of the R/G for these DNA sequences and M_e of E/G.

In Table 2, we give the upper triangular part of the similarities/dissimilarities matrix of R/G of the 3DD-Curves. Observing Table 2, we find that, the smallest entries are associated with the pairs (gorilla, chimpanzee), (Macaca fascicular, Macaca fuscata), (Macaca fascicular, Macaca mulatta) and (Macaca fuscata, Macaca mulatta).

In Table 3, we give the upper triangular part of the similarities/dissimilarities matrix of the E/G of the 3DD-Curves. We can only see some species qualitative agreement among similarities based on E/G of 3DD-Curve and based on R/G of similar 3DD-Curve. In fact we may get different results based on 3DD-curves with different parameters. And we have not find good result that can suit to construct the phylogenetic tree. Hence we only provide

Table 3: The upper triangular part of the similarities/dissimilarities matrix M_e of the E/G of the 3DD-Curves

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0839	0.2383	0.7192	0.2419	0.3299	0.5303	0.2072	0.6731	0.9560	0.3253
Gorilla		0	0.2003	0.7277	0.2555	0.3205	0.5368	0.1361	0.6802	0.9635	0.3249
Hyl			0	0.5578	0.1491	0.1652	0.3775	0.1909	0.5065	0.7870	0.1708
Lemur				0	0.4800	0.4202	0.2036	0.7166	0.1693	0.2405	0.4035
M. Fas					0	0.1333	0.2993	0.2923	0.4343	0.7151	0.0992
M. Fus						0	0.2246	0.2976	0.4081	0.6569	0.0543
M. Syl							0	0.5211	0.2518	0.4429	0.2148
Ora								0	0.6878	0.9505	0.3234
S. Sci									0	0.3191	0.3765
T. Syr										0	0.6403
M. Mul											0

the phylogenetic tree which is constructed by similarities/dissimilarities matrix M_r of the 3DD-Curves.

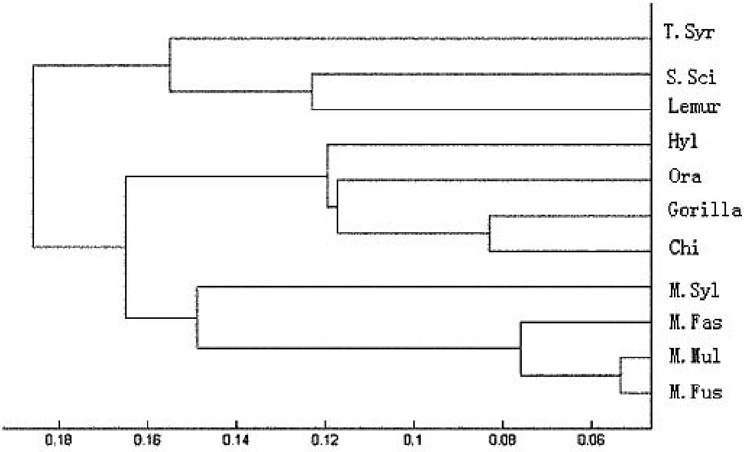

Figure 2: Phylogenetic tree based on 3DD-Curve

In Figure 2, we have presented the phylogenetic tree which is constructed with complete linkage algorithm from Table 4 for the 11 different species. Among the macaque species, macaca sylvanus appears to be distantly related to the other three types of macaques. Moreover, macaca mulatta and Macaca fuscata shown as the most closely related among the macaque species. Chimpanzee and gorilla are shown as the most closely related among the

11 different species. The topology of the tree, except for the positions of the tarsier, is generally in agreement with the widely accepted classification of primates that is based on fossil records and other molecular analysis. That the tarsier is more closely related to the lemur than to anthropoids is in agreement with conclusions of Hayasaka, Gojobori and Horai [18]. We can also find the distance between two species being connected is basically agree with the number of nucleotide substitutions in Figure 1.

We have made numerous computation with different parameters, and find the topology of the tree is quit stable when the different parameters satisfy $u = 2v$. By selecting suitable v and then u, we can obtain the divergence times among all species.

5 Conclusions

Different matrix has been proposed to mathematically characterize the DNA sequences, Such matrix allow one to make quantitative comparisons between different DNA sequences, whether between within the same or between different species. Our analysis of the sequences of mtDNAs base on the new distance matric has provided new insights into evolutionary relationships among primates. In addition, we have shown that similarities/dissimilarities matrix M_r is more fit for comparison of DNA sequences than M_e (base on D/D or E/G matrix). Choosing the right parameters for 3DD-Curve is very important. We should try to find out suitable parameters so that 3DD-Curve most appropriate to the problem under consideration.

Acknowledgements

This work was supported in part by the Shandong Natural Science Foundation(Y2006A14).

References

[1] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.*, 119(1986), 319-328.

[2] A. Nandy, Graphical representation of long DNA sequences, *Curr. Sci*, 66(1994), 821.

[3] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Applic. Biosci.* , 12(1995), 503-511.

[4] M. Randić, M. Vračko, N. Lerš, D. Plavšić, On 3-D graphical representation of DNA primary sequence and their numerical characterization, *J. Chem. Inf. Comput. Sci.*, 40(2000), 1235-1244.

[5] M. Randić, X. F. Guo, S. C. Basak, On the Characterization of DNA Primary Sequence by Triplet of Nucleic Acid Bases, *J. Chem. Inf. Comput. Sci.*, 41(2001), 619-626.

[6] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* , 368(2003), 1-6.

[7] Wu, Y., Liew, A.W., Yan, H., Yang, M., DB-Curve: a novel 2D method of DNA sequence visualization and representation, *Chem. Phys. Lett.*, 367(2003), 170-176.

[8] Bo Liao, A 2D graphical representation of DNA sequence,*Chem. Phys. Lett.*, 401(2005) 196-199.

[9] Bo Liao, Tianming Wang, 3-D graphical representation of DNA sequences and their numerical characterization, *J. Mol. Struct. (Theochem)* , 681(2004), 209-212.

[10] Bo Liao, Kequan Ding, A Graphical Approach to analyzing DNA sequences, *J. Comput. Chem.*, 26 (2005) 1519-1523.

[11] Bo Liao, Kequan Ding, A 3D graphical representation of DNA sequences and its application, *Theor. Comput. Sci.* , 358 (2006) 56 - 64.

[12] Bo Liao, Yusen Zhang, Kequan Ding, Tianming Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation,*J. Mol. Struct. (Theochem)* , 717(2005), 199-203.

[13] Yusen Zhang, Bo Liao, Kequan Ding, On 2D Graphical Representation of DNA Sequence of Nondegeneracy, *Chem. Phys. Lett.*, 411 (2005) 28-32.

[14] Yusen Zhang, Bo Liao, Kequan Ding, On 3DD-Curves of DNA Sequences, *Molecular Simulation*, 32(2006), 29-34

[15] Yusen Zhang, Wei Chen, Invariants of DNA Sequences Based on 2DD-Curves, *J. Theor. Biol.* , 242 (2006), 382-388.

[16] Yusen Zhang, On 3D Graphical Representation of RNA Secondary Structure, *MATCH Commun. Math. Comput. Chem.*, 57 (2007), 157-168.

[17] Yusen Zhang, On 2D Graphical Representation of RNA Secondary Structure, *MATCH Commun. Math. Comput. Chem.*, 57 (2007), 697-710.

[18] Hayasaka, K., T. Gojobori, and S. Horai, Molecular phylogeny and evolution of primate mitochondrial DNA, *Mol. Biol. Evol.* , 5(1988), 626-644.