MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **58** (2007) 197-208

ISSN 0340 - 6253

New Invariant of DNA Sequences

Yusen Zhang *, Wei Chen
Department of Mathematics, Shandong University at Weihai
Weihai 264209, China
(Received February 5, 2007)

Abstract. *Instead of leading eigenvalue of the matrix associated with a DNA sequence, we propose a new invariant base on the 2DD-Curves of DNA sequences which is simple for calculation and is proved to be highly effective for comparison of DNA sequences and construction of phylogenetic tree. The utility of the new invariant is illustrated on the primate mitochondrial DNA sequences for* 11 *different species. We also construct the phylogenetic tree for primate mitochondrial DNA sequences.*

1 Introduction

Comparison of different DNA primary sequences remains one of the important aspects of the analysis of DNA data banks. Usual representation of a DNA primary sequence is that of a string of letters A, G, C and T, which signify the four nucleic acid bases adenine, guanine, cytosine, and thymine, respectively. How similar/dissimilar the different sequences are may depend on how such strings of letters are encoded or characterized. The previous procedures consider differences between strings due to deletion-insertion, compression-expansion, and substitution of the string elements. These approaches, which have been hitherto widely used, are computer intensive[1][2]. Recently, an alternative approach for the comparison of sequences has been introduced. It is based on characterization of DNA by ordered sets of invariants derived from DNA sequences, rather than by a direct comparison of DNA sequences themselves [3-12]. However, as pointed in [11], this approach involves a number of as yet unresolved questions. In particular, questions that need our attention are as follows: how to obtain suitable invariants to characterize DNA sequences and how to select invariants suitable for sequence comparisons. A sequence invariant is usually a real number that is independent of the labels (bases) A, G, C, and T. For example, the length of the sequence is an invariant. But the length cannot capture the main information of the sequence considered, so it is regarded as a trivial invariant. Among other invariants, such as the average matrix element, the average matrix row sum, only the leading eigenvalue of a matrix associated with

*Corresponding author:zhangys@sdu.edu.cn

a DNA sequence is an important invariant and is proved to be effective for characterization of DNA sequences. However, some problems we must face are that the calculation of the eigenvalue will become more and more difficult with the order of the matrix large. Are there any other suitable descriptors for DNA sequences?

In this contribution we propose another invariant of DNA sequence based on the nondegeneracy 2DD-Curves [7] which is more simple and effective for calculation than others. The utility of the new invariant is illustrated on primate mitochondrial DNA sequences belonging to eleven different species. The phylogenetic relationships among primate groups shown by our analysis are generally consistent with result in [13].

2 Preparation

2.1 Definition of Invariants

A invariant of DNA sequences is usually a real number that is independent of the labels (bases) A, G, C, and T. As we know, once a symmetric matrix M is given, one often use some of matrix invariants, such as the average matrix element, the average row sum, the leading eigenvalue, as descriptors of the sequence [4][6][9-12]. For convenience, we denote these invariants by $Aver(M)$, $Avrow(M)$, and $\lambda(M)$, respectively. The average row sum $Avrow(M)$ is defined by

$$Avrow(M) = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{N} a_{ij}),$$

and the average matrix element is defined by

$$Aver(M) = \frac{1}{N^2}\sum_{i=1}^{N}(\sum_{j=1}^{N} a_{ij}).$$

The leading eigenvalue of the matrix associated with a DNA sequence as a important invariant, is effectively used in analysis of similarity of DNA sequences in better results. In order to avoid the complicated calculation of leading eigenvalue, the authors [5] propose a new sequence invariant ALE-index χ which is defined by

$$\chi = \chi(M) = \frac{1}{2}(Avrow(M) + \sqrt{\frac{N-1}{N}\sum_{i,j=1}^{N}|a_{ij}^2|})$$

The main advantage of this invariant is simple for calculation and it has good approach to the leading eigenvalue of the matrix associated with a DNA sequence. But we can not understand the biological meaning of it. So we want to propose another invariant which can also be used to characterize DNA sequences.

Here we propose a new descriptors based on the 2DD-Curves of the DNA sequences. Its definition is as follows:

Given a DNA sequence with N bases, we can always associate it with an $N \times N$ nonnegative real symmetric matrix whose diagonal entries are zero [4][6][9][10]. Let $M = (a_{ij})_{N\times N}$ be such a matrix, i.e., $a_{ij} \geq 0, a_{ij} = a_{ji}$, and $a_{ii} = 0$ for $i, j = 1, 2, \ldots, N$. Define

$$AEdis(M) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\sum_{j=1}^{N} a_{ij})$$

It is easy to see that $Aver(M) \leq AEdis(M) \leq Avrow(M)$ when $N \geq 2$. In the following part, we will show that the $Aver(M)$ can be used to characterize DNA sequences.

For general real symmetric matrix M whose diagonal entries are zero, It is hard to describe the characterization of the $AEdis(M)$. Different symmetric matrix associated with a DNA sequence includes different information about DNA sequence which have directly relations with the graphical representation of DNA sequence. In order to further discuss the invariant, It is necessary to consider the idiographic matrices associated with a DNA sequence based on a fixed graphical representation. Here we will consider the nondegeneracy 2D graphical representation of DNA sequence.

2.2 2DD-Curves

The DNA sequences have been converted to nondegeneracy 2D graphical representation using a mapping described in previous papers [7], reviewed here for convenience. The four nucleotides can be arranged in classes according to the three main dichotomies in their biochemical properties: purine(A, G)/pyrimidine(C, T), amino(A, C)/keto(G, T) and weekbond(A, T)/strong-H band(G, C). We will use the following three special representations corresponding to the three classifications, respectively.

1. Assigning the following vectors to the four bases:

$$(\frac{\sqrt{3m}}{2}, \frac{\sqrt{m}}{2}) \longrightarrow A, (\frac{\sqrt{3n}}{2}, \frac{\sqrt{n}}{2}) \longrightarrow C,$$
$$(\frac{\sqrt{2n}}{2}, \frac{\sqrt{2n}}{2}) \longrightarrow T, (\frac{\sqrt{2m}}{2}, \frac{\sqrt{2m}}{2}) \longrightarrow G,$$

where n, m are different positive real numbers, but not perfect square numbers. So that 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 (i = 0, 1, 2, . . ., N, where N is the length of the DNA sequence being studied) satisfy

$$\begin{cases} x_i = \frac{\sqrt{3m}}{2} A_i + \frac{\sqrt{3n}}{2} C_i + \frac{\sqrt{2n}}{2} T_i + \frac{\sqrt{2m}}{2} G_i \\ y_i = \frac{\sqrt{m}}{2} A_i + \frac{\sqrt{n}}{2} C_i + \frac{\sqrt{2n}}{2} T_i + \frac{\sqrt{2m}}{2} G_i \end{cases} \tag{1}$$

where A_i, C_i, G_i, and T_i are the cumulative occurrence numbers of A, C, G, and T, respectively, in the subsequence from the 1st base to the i-th base in the sequence. We define $A_0 = C_0 = G_0 = T_0 = 0$.

We call it 2DD-Curve of DNA sequences based on pattern ACTG.

2.Assigning the following vectors to the four bases:

$$(\frac{\sqrt{3m}}{2},\frac{\sqrt{m}}{2})\longrightarrow A,(\frac{\sqrt{3n}}{2},\frac{\sqrt{n}}{2})\longrightarrow G,$$
$$(\frac{\sqrt{2n}}{2},\frac{\sqrt{2n}}{2})\longrightarrow C,(\frac{\sqrt{2m}}{2},\frac{\sqrt{2m}}{2})\longrightarrow T,$$

then, we get 2DD-Curve of DNA sequences based on pattern AGCT:

$$\begin{cases} x_i=\frac{\sqrt{3m}}{2}A_i+\frac{\sqrt{3n}}{2}G_i+\frac{\sqrt{2n}}{2}C_i+\frac{\sqrt{2m}}{2}T_i \\ y_i=\frac{\sqrt{m}}{2}A_i+\frac{\sqrt{n}}{2}G_i+\frac{\sqrt{2n}}{2}C_i+\frac{\sqrt{2m}}{2}T_i \end{cases} \tag{2}$$

3. Assigning the following vectors to the four bases:

$$(\frac{\sqrt{3m}}{2},\frac{\sqrt{m}}{2})\longrightarrow A,(\frac{\sqrt{3n}}{2},\frac{\sqrt{n}}{2})\longrightarrow T,$$
$$(\frac{\sqrt{2n}}{2},\frac{\sqrt{2n}}{2})\longrightarrow G,(\frac{\sqrt{2m}}{2},\frac{\sqrt{2m}}{2})\longrightarrow C,$$

we get 2DD-Curve of DNA sequences based on pattern ATGC:

$$\begin{cases} x_i=\frac{\sqrt{3m}}{2}A_i+\frac{\sqrt{3n}}{2}T_i+\frac{\sqrt{2n}}{2}G_i+\frac{\sqrt{2m}}{2}C_i \\ y_i=\frac{\sqrt{m}}{2}A_i+\frac{\sqrt{n}}{2}T_i+\frac{\sqrt{2n}}{2}G_i+\frac{\sqrt{2m}}{2}C_i \end{cases} \tag{3}$$

Unless otherwise stated, we always use the pattern AGCT as default 2DD-Curve of DNA sequences in following parts.

Table 1: The Upper Triangles of the ED Matrix of the Sequence ATGGTGCACC

ED	A	T	G	G	T	G	C	A	C	C
A	0	1.7321	4.3778	7.0236	8.7556	11.4014	13.0821	15.6710	17.3751	19.0843
T		0	2.6458	5.2915	7.0236	9.6693	11.3512	13.9440	15.6511	17.3630
G			0	2.6458	4.3778	7.0236	8.7089	11.3091	13.0216	14.7387
G				0	1.7321	4.3778	6.0674	8.6807	10.4004	12.1236
T					0	2.6458	4.3420	6.9671	9.6925	10.4200
G						0	1.7321	4.3778	6.1099	7.8419
C							0	2.6458	4.3778	6.1099
A								0	1.7321	3.4641
C									0	1.7321
C										0

For the 2DD-Curve of DNA sequences, we have a set of points (x_i, y_i), i = 1, 2, . . ., N, where N is the length of the sequence. We construct the Euclidean-distance matrix ED and quotient matrix E/G. The (i,j) element of matrix ED is defined to be the Euclidean-distance between vertices i and j of the 2DD-Curve. The the (i,j) element $[E/G]_{ij}$ of matrix E/G is

defined to be $[ED]_{ij}/|i-j|$. Clearly, the matrix ED and E/G are symmetric. The upper triangle elements of the ED matrix of the 2DD-Curve with $n = 3, m = 7$ for the DNA segment ATGGTGCACC are listed in Table 1. Observing Table 1, we can see that each entry in one column of matrix ED is bigger than corresponding one in previous columns.

Table 2: Database Source

species	ID/ ACCESSION	$length(bp)$	database
Saimiri sciureus	M22655	893	$NCBI$
Hylobates	V00659	896	$NCBI$
Lemur catta	M22657	895	$NCBI$
Macaca fascicular	M22653	896	$NCBI$
Gorilla	V00658	896	$NCBI$
Macaca fuscata	M22651	896	$NCBI$
Macaca mulatta	M22650	896	$NCBI$
Macaca sylvanus	M22654	896	$NCBI$
Macaca mulatta	V00672	896	$NCBI$
Orangutan	V00675	895	$NCBI$
Tarsius syrichta	M22656	895	$NCBI$

In fact, if denoting matrix ED of the 2DD-Curve based on pattern of AGCT, AGCT or ATGC as $(a_{ij})_{N\times N}$, it is not difficult to prove that

$$a_{ij+1} \geq a_{ij}, i = 1, 2, \ldots, N.$$

This property depends on the construction of the 2DD-Curves, different 2DD-Curve will lead to different results.

Considering that $\frac{1}{N-1}\sum_{j=1}^{N} a_{ij}$ is the average Euclidean-distance from vertex i to all other vertices on the 2DD-Curve of DNA sequences, the average Euclidean-distance of the 2DD-Curve can be defined as

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N-1}\sum_{j=1}^{N} a_{ij}),$$

which happened to be the $AEdis(ED)$. In other words, $AEdis(ED)$ is the average of all the Euclidean-distances between every pairs of points on 2DD-Curve of DNA sequence. The $AEdis$ and leading eigenvalue of ED matrix in Table 1 are 7.9964, 74.6486, respectively. One can see that $AEdis$ and leading eigenvalue are quietly different descriptors of the 2DD-Curves. Notices that the 2DD-Curve of a DNA sequence is nondegeneracy, which means the correspondence between DNA sequences and DNA graphs is one to one, we can regard $AEdis$ as a invariant of DNA sequences. It is easy to see that $AEdis$ is simple for calculation and thus facilitated for characterization of DNA sequences.

In Table 3, we show the $AEdis(ED)$ of 2DD-Curve for the mitochondrial DNA sequences for 11 different species are listed in table 2.

Table 3: $AEdis(ED)(\times 10^{-3})$ of the mitochondrial DNA sequences for 11 different species when $n = 3; m = 7$

	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
ACTG	0.6256	0.6272	0.6312	0.6346	0.6296	0.6305	0.6315	0.6267	0.6367	0.6384	0.6308
AGCT	0.6708	0.6680	0.6694	0.6845	0.6756	0.6745	0.6808	0.6651	0.6812	0.6887	0.6754
ATGC	0.6842	0.6862	0.6856	0.6795	0.6819	0.6859	0.6833	0.6896	0.6737	0.6768	0.6838

Table 4: The $AEdis(E/G)$ of 2DD-Curves of the mitochondrial DNA sequences for 11 species with $n = 3; m = 7$

	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
ACTG	2.0917	2.0972	2.1106	2.1236	2.1052	2.1079	2.1114	2.0981	2.1351	2.1364	2.1089
AGCT	2.2429	2.2339	2.2373	2.2908	2.2582	2.2552	2.2754	2.2254	2.2845	2.3040	2.2584
ATGC	2.2893	2.2956	2.2936	2.2756	2.2817	2.2944	2.2864	2.3104	2.2619	2.2667	2.2877

Table 5: The $AEdis(E/G)$ of 2DD-Curves of the mitochondrial DNA sequences for 11 species with $n = 3; m = 11$

	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
ACTG	2.3658	2.3752	2.3985	2.4214	2.3894	2.3939	2.4000	2.3765	2.4414	2.4437	2.3957
AGCT	2.6284	2.6129	2.6188	2.7118	2.6550	2.6499	2.6850	2.5982	2.7008	2.7346	2.6554
ATGC	2.7089	2.7199	2.7163	2.6847	2.6956	2.7178	2.7036	2.7456	2.6610	2.6690	2.7061

Table 6: The $AEdis(E/G)$ of 2DD-Curves of the mitochondrial DNA sequences for 11 species with $n = 1/5; m = 1/3$

	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
ACTG	0.4969	0.4977	0.4996	0.5014	0.4988	0.4992	0.4997	0.4978	0.5030	0.5032	0.4993
AGCT	0.5183	0.5170	0.5175	0.5251	0.5205	0.5201	0.5229	0.5158	0.5242	0.5270	0.5205
ATGC	0.5249	0.5258	0.5255	0.5231	0.5239	0.5257	0.5246	0.5279	0.5211	0.5218	0.5247

3 Application of Invariant

We will illustrate the use of the new invariants of DNA sequences with the examination of similarities and dissimilarities among the 11 coding sequences of Table 2 by constructing a vector having 3 components consisting of the invariants of $AEdis(E/G)$ the 2DD-Curves based on pattern ACTG, AGCT and ATGC. The underlying assumption is that if two vectors point to a similar direction in the 3-dimensional space and have similar magnitudes, then the two DNA sequences represented by the two vectors are similar. To reduce variations caused by a different length of sequences, one can consider the normalized $AEdis$, i.e. $AEdis' = AEdis/N$, where N is the length of the sequence and the order of the corresponding matrix as well. the normalized $AEdis$ of 2DD-Curves for the mitochondrial DNA sequences for 11 different species with different parameters are listed in table 4, 5, 6, respectively.

Observing theses tables, we find that, different m, n can result in different $AEdis(E/G)$ of 2DD-Curve, but for a fixed m, n, the $AEdis(E/G)$ provides us different information about the species and the tendency is similar when different m, n is chosen.

Table 7: The similarity/dissimilarity matrix for the coding sequences of Table 2 based on three normalized $AEdis(E/G)$ of 2DD-Curve with $n = 3, m = 7$

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0123	0.0201	0.0592	0.0218	0.0210	0.0381	0.0281	0.0661	0.0790	0.0232
Gorilla		0	0.0140	0.0659	0.0291	0.0239	0.0448	0.0171	0.0716	0.0854	0.0283
Hyl			0	0.0580	0.0246	0.0181	0.0388	0.0241	0.0619	0.0764	0.0219
Lemur				0	0.0380	0.0432	0.0224	0.0784	0.0190	0.0204	0.0376
M. Fas					0	0.0134	0.0189	0.0441	0.0444	0.0574	0.0071
M. Fus						0	0.0220	0.0352	0.0515	0.0629	0.0075
M. Syl							0	0.0570	0.0353	0.0428	0.0173
Ora								0	0.0849	0.0977	0.0414
S. Sci									0	0.0201	0.0451
T. Syr										0	0.0573
M. Mul											0

In Table 7, we give the similarities and dissimilarities for the coding sequences of Table 2 based on the Euclidean distances between the end points of the 3-component vectors of the normalized invariants $AEdis(ED)$ of 2DD-Curves with $n = 3, m = 7$.

Observing Table 7, 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 8, 9, we give the similarities and dissimilarities for the coding sequences of Table 2 based on the Euclidean distances between the end points of the 3-component vectors of the normalized invariants $AEdis(ED)$ with $n = 3$ and $m = 11$, $n = 1/5$ and $m = 1/3$, respectively.

Table 8: The similarity/dissimilarity matrix for the coding sequences of Table 2 based on three normalized $AEdis(E/G)$ of 2DD-Curve with $n = 3$ and $m = 11$

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0212	0.0348	0.1030	0.0379	0.0364	0.0663	0.0487	0.1151	0.1376	0.0403
Gorilla		0	0.0243	0.1147	0.0507	0.0415	0.0780	0.0296	0.1249	0.1487	0.0492
Hyl			0	0.1008	0.0427	0.0314	0.0674	0.0421	0.1078	0.1330	0.0381
Lemur				0	0.0660	0.0753	0.0391	0.1364	0.0329	0.0356	0.0655
M. Fas					0	0.0232	0.0328	0.0768	0.0775	0.1000	0.0122
M. Fus						0	0.0383	0.0612	0.0899	0.1097	0.0131
M. Syl							0	0.0992	0.0615	0.0746	0.0300
Ora								0	0.1480	0.1703	0.0721
S. Sci									0	0.0348	0.0787
T. Syr										0	0.0998
M. Mul											0

Table 9: The similarity/dissimilarity matrix for the coding sequences of Table 2 based on three normalized $AEdis(E/G)$ of 2DD-Curve with $n = 1/5$ and $m = 1/3$

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0018	0.0029	0.0084	0.0031	0.0030	0.0054	0.0040	0.0093	0.0112	0.0033
Gorilla		0	0.0020	0.0093	0.0041	0.0034	0.0064	0.0024	0.0101	0.0121	0.0040
Hyl			0	0.0082	0.0035	0.0026	0.0055	0.0034	0.0087	0.0108	0.0031
Lemur				0	0.0054	0.0061	0.0032	0.0111	0.0027	0.0029	0.0053
M. Fas					0	0.0019	0.0027	0.0062	0.0063	0.0081	0.0010
M. Fus						0	0.0031	0.0050	0.0073	0.0089	0.0011
M. Syl							0	0.0081	0.0050	0.0060	0.0025
Ora								0	0.0120	0.0138	0.0059
S. Sci									0	0.0029	0.0064
T. Syr										0	0.0081
M. Mul											0

Comparing Table 7, 8 and 9, we can see there exists an overall qualitative agreement among similarities based $AEdis(E/G)$ of 2DD-Curve with different parameters. That means the result of comparison doesn't rely heavily on the parameters of 2DD-Curves.

In Table 10, we also compute the row average values of similarity/dissimilarity matrix of table 7, 8 and 9. Observe table 10, we can find that Tarsius syrichta is very dissimilar to others among the 11 species because its corresponding average value has large entries in three cases and Macaca fascicular is very similar with Macaca fuscata. These results are generally consistent with that reported in [13].

Table 10: The row average of the mitochondrial DNA sequences for 11 species based on three curves

	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Table 7	0.0369	0.0392	0.0358	0.0442	0.0299	0.0299	0.0338	0.0508	0.0500	0.0599	0.0287
Table 8	0.0642	0.0683	0.0622	0.0769	0.0520	0.0520	0.0587	0.0885	0.0871	0.1044	0.0499
Table 9	0.0052	0.0056	0.0051	0.0063	0.0042	0.0042	0.0048	0.0072	0.0071	0.0085	0.0041

4 Construction of the Phylogenetic Trees

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

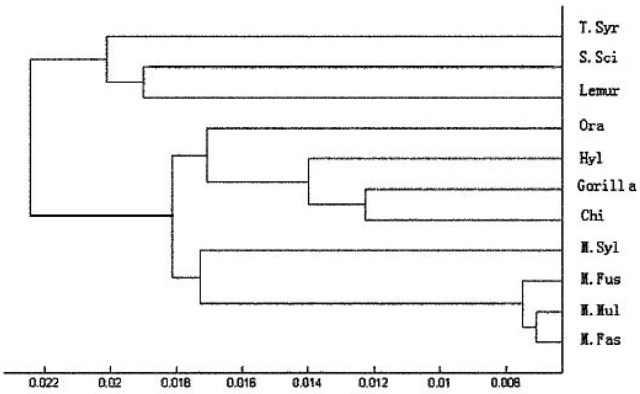

Figure 1: Dendrogram tree based on 2DD-Curve with $n = 3, m = 7$

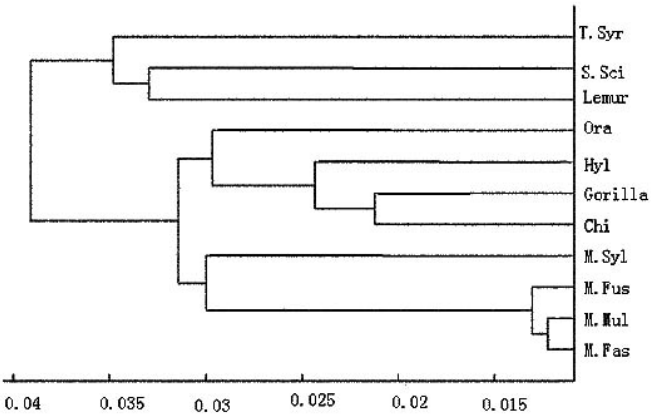

Figure 2: Dendrogram tree based on 2DD-Curve with $n = 3, m = 11$

Hayasaka, Gojobori and Horai [13] 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 topology of the tree, except for the position of the tarsier, is generally in agreement with the widely accepted classification of primates that is based on fossil records and other molecular analysis. The phylogenetic relationships derived from these mtDNA sequence comparisons appear reliable.

In Figure 1, 2 and 3, we have presented the dendrogram tree based on linkage cluster analysis using Euclidean distances of these 3-dimensional vectors which consist of Table 4, Table 5 and Table 6, respectively, for the 11 different species. The phylogenetic relationships among primate groups shown by our analysis are generally consistent with results in [13]. But, all the previous methods require a multiple alignment of the sequences and assume some sort of an evolutionary model. most existing phylogeny construction methods, the proposed method does not require multiple alignment.

One can also find that these three trees based on 2DD-Curves with different parameters give us phylogenetic trees with the same topology. So the construction of phylogenetic tree doesn't rely heavily on the parameters of 2DD-Curves too.

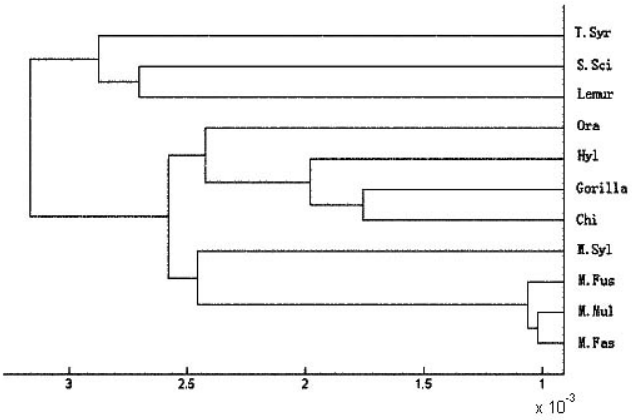

Figure 3: Dendrogram tree based on 2DD-Curve with $n = 1/5, m = 1/3$

5 Conclusions

We propose a new invariant of DNA sequences based on ED(or E/G) of 2DD-Curves which is simple for calculation and is proved to be highly effective for comparison of DNA sequences and construction of phylogenetic tree. when it is used in comparison of DNA sequences, results are much similar with previous studies. We also show the results of choosing different parameters n, m.

Acknowledgements

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

References

[1] A. Arneodo, Y. dAubenton Carafa, B. Audit, E. Bacry, J.F. Muzy, C. Thermes, What can we learn with wavelets about DNA sequences?,*Physica A*, 249 (1998) 439-448.

[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, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* , 368(2003), 1-6.

[5] Chun Li, Jun Wang, New Invariant of DNA Sequences, *J. Chem. Inf. Model.*, 45(2005), 115-120.

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

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

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

[9] 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.

[10] 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.

[11] 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.

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

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