J Math Chem (2008) 44:206–216
DOI 10.1007/s10910-007-9302-2

ORIGINAL PAPER

Visualization of DNA sequences based on 3DD-Curves

Yusen Zhang · Mingshu Tan

Received: 31 March 2006 / Accepted: 12 July 2006 / Published online: 21 September 2007

Abstract In this paper we (1) introduce a new 3D graphical representation of DNA sequences; (2) visualize DNA sequences based on 3DD-Curves; (3) provide a new invariant of DNA sequences based on our 3DD-Curve. All this represents a new development of graphical representation and numerical characterization for DNA sequences.

Keywords DNA · 3DD-Curve · Invariant

1 Introduction

One important task in the study of genome sequences is to determine densities of specific nucleotides and to understand the implications for exons, or coding regions. Several methods for addressing this problem graphically have been advanced [1]. Graphical representations of DNA sequences are useful because they allow visual observations of nucleotide composition, base pair patterns, and sequence evolution. Several authors outlined different graphical representation of DNA sequences based on 2D, 3D [2–8]. But both 2D and 3D graphical representation are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. Randic [3] present a novel 2D graphical representation, which avoids the limitation of Nandys approach, and outlined an approach to analysis the similarity among the coding sequences of the first exon of β-globin gene of 11 different species. We provide a 2D graphical representation without degeneracy [9]. The H-curve [10]

Y. Zhang (✉)
Department of Mathematics, Shandong University at Weihai, Weihai 264209, China
e-mail: zhangys@sdu.edu.cn

M. Tan
Chongqing Three Gorges University, Chongqing 404000, China

is another 3D graphical representation of DNA sequences. Other new 3D graphical representation can be found in [11–15].

In order to find some of the invariants sensitive to the form of the characteristic curve we can transform the graphic representation of the characteristic curve into another mathematical object, a matrix [2–9]. The leading eigenvalue of the matrix associated with a DNA sequence is an important invariant and is proved to be highly effective for characterization of DNA sequences. However, the biological meaning of the leading eigenvalue of a matrix associated with a DNA sequence is not easy to understand and the calculation of the eigenvalue will become more and more difficult with the order of the matrix large.

In this paper we introduce a novel 3DD-Curve of DNA sequences and provide a new invariant of DNA sequences based on the 3DD-Curve. We will make some data visualization for the first exon of β-globin genes sequences belonging to 11 different species.

2 Construction of 3DD-Curve

We construct a map between the bases of DNA sequences and plots in 3D space, then we will obtain a 3D representation of the corresponding DNA sequences. In 3D space points, vectors and directions have three components, and we will assign the following basic elementary directions to the four free bases.

We assign one nucleic base as follows:

$$(\sqrt{u}, \sqrt{u}, \sqrt{u}) \longrightarrow A, (-1, 0, 0) \longrightarrow G, (0, 1, 0) \longrightarrow C, (0, 0, -1) \longrightarrow T$$

where u is positive real number, but not perfect square number. So that we can reduce a DNA sequence into a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates x_i, y_i, z_i (i=0,1,2,…,N, where N is the length of the DNA sequence being studied) satisfy

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

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 ith base in the sequence. We define $A_0 = C_0 = G_0 = T_0 = 0$.

We called 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).

In Fig. 1, we show the characteristic curves that represent the first 10 bases of the coding sequence of the first exon of human and rabbit β-globin gene with $u = 2$.

As we know, 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).

According to the above definition, the 3DD-Curve is a three-dimensional space curve, which has three components, i.e. x_i, y_i and z_i. Each component has a clear biological implication. The component x_i displays the weighted distribution of bases

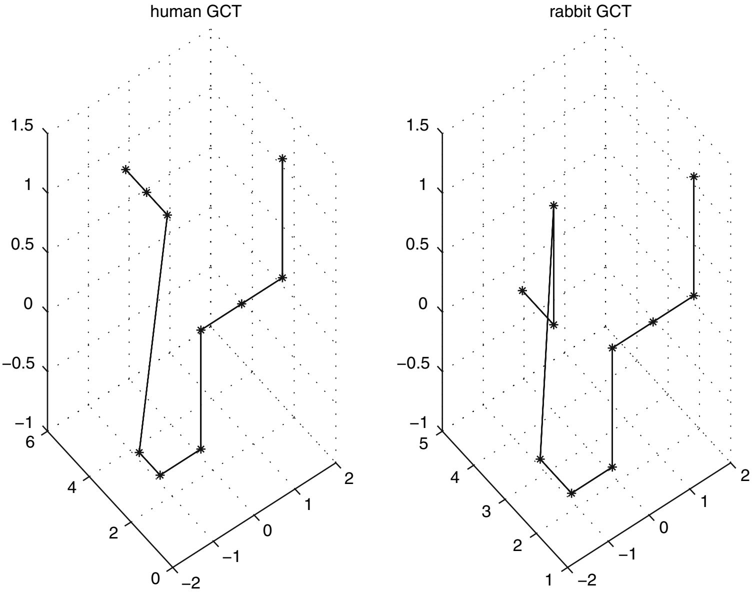

Fig. 1 Characteristic curve based on pattern GCT

of purine (A, G) along the DNA sequence. The component y_i displays the weighted distribution of bases of amino (A, C) along the sequence. The component z_i displays the weighted distribution of bases of weak-hydrogen bond (A, T) along the sequence. Consequently, the DNA sequence can be completely described by the three independent distributions.

We can obtain only three representations corresponding to the three classifications of four bases of DNA. Next two representations are as follows:

1. Assigning the following vectors to the four bases:
$(\sqrt{u}, \sqrt{u}, \sqrt{u}) \longrightarrow A$, $(-1, 0, 0) \longrightarrow C$, $(0, 1, 0) \longrightarrow G$, $(0, 0, -1) \longrightarrow T$, then, we get

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

2. Assigning the following vectors to the four bases:
$(\sqrt{u}, \sqrt{u}, \sqrt{u}) \longrightarrow A$, $(-1, 0, 0) \longrightarrow G$, $(0, 1, 0) \longrightarrow T$, $(0, 0, -1) \longrightarrow C$, we get

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

For any DNA sequence, we can obtain the three 3DD-curves which corresponding pattern GCT, CGT and GTC. That means different parameters can result in different visual clues to DNA sequence.

It is easy to see that, for given x-projection, y-projection and z-projection of any point $P = (x, y, z)$ on 3DD-Curve, after uniquely determining the number a_p, g_p, c_p and t_p of A, G, C, and T from the beginning of the sequence to the point P. By successive x-projection, y-projection and z-projection of points on the sequence, we can recover the original DNA sequence uniquely from the DNA graph. So we can 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.

In other words, the move in each of the three directions in our 3DD-Curve is equiprobable. It is not to move one unit along one of four directions representing the bases (A, C, G, T) but to move different unit for different bases. That is why our 3D graphical representation of DNA sequences is non degeneracy. That means the 3DD-Curve is a three-dimensional space curve constituting the unique representation of a given DNA sequence in the sense that each can be uniquely reconstructed given the other. 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.

3 Visualization of DNA sequences

In this section, we will make a comparison for the first exon of β-globin genes sequences belonging to 11 different species based on our 3DD-Curve. By the way, we also explain how to use the parameter u. In Table 1, the first exon-1 of the β-globin gene for 11 different species are listed, which were reported by Randic [16].

In Fig. 2, we show the 3DD-Curves of the first exon of β-globin gene of 11 different species in Table 1, which corresponding pattern GCT. By examining these 3DD-Curves, we find that gallus and opossum are dissimilar to others, and the more similar species should be human, gorilla and chimpanzee can be verified. However, we can also found that goat and rabbit also have some similar with human on this condition. So we need change the parameter u so that we can analyze these DNA sequences by corresponding different forms of 3DD-curve. For example, 3DD-Curves of the first exon of β-globin genes of human and rabbit which corresponding pattern GCT with $u = 2$, 1/3 and 1/5, respectively, are drawing in Fig. 3.

Observing Fig. 2, we can see the curves of goat and rabbit have some similar tendency with human, but in Fig. 3, we can see that rabbit has various degree of leaps comparing with human, especially when $u = 1/5$, the amplitude of 3DD-curve of rabbit is different from that of human. That is why human is more similar with gorilla and chimpanzee than rabbit and goat. Observing Fig. 4, we can easily find that human is similar with gorilla, but dissimilar with gallus in any pattern and we can conclude that different pattern can show us different information about the DNA sequences. Of course, it is necessary to analyze the similarity by other numerical characterizations of 3DD-Curves.

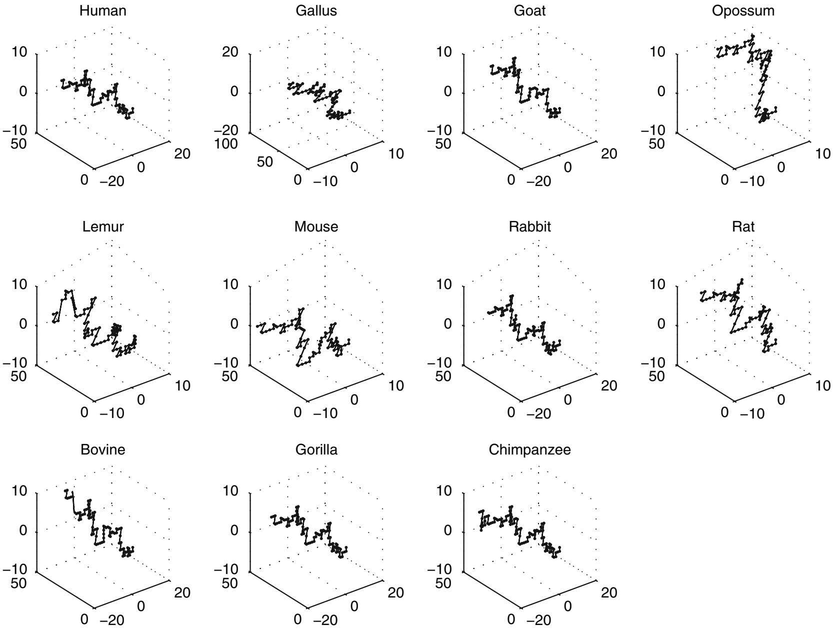

Fig. 2 3DD-Curves of the first exon of β-globin gene of 11 different species

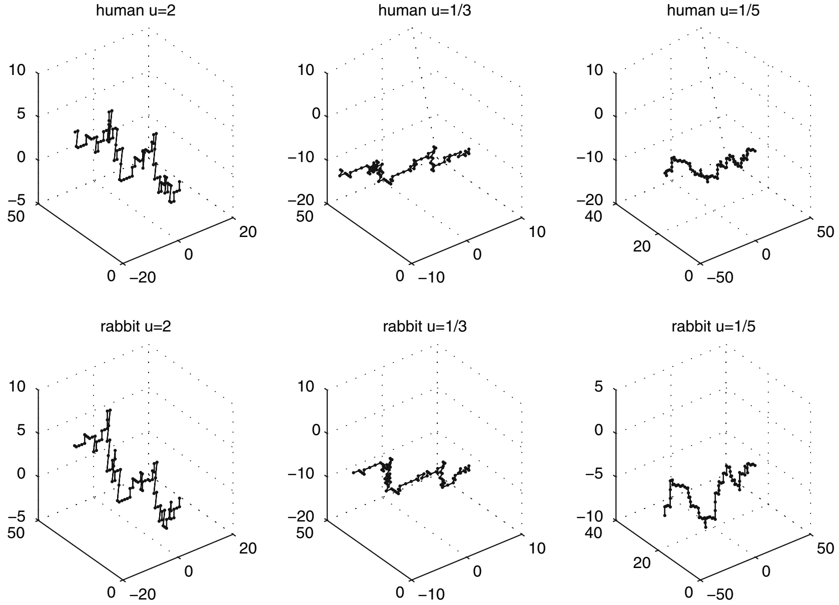

Fig. 3 3DD-Curves of the first exon of β-globin genes of human and rabbit corresponding GCT

Table 1 The coding sequences of the first exon of β-globin gene of 11 different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemmur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

We can see the 3DD-Curve with parameter u can provide more information about DNA sequence than existing graphic representation by choosing the appropriate parameter u. We can choose the system most appropriate to the problem at hand.

4 Invariants of DNA sequences

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 leading eigenvalue $\lambda(M)$, the average matrix element, as descriptors of the sequence [5,8]. The average matrix element, denoted as $Aver(M)$, is defined by

$$Aver(M) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\right).$$

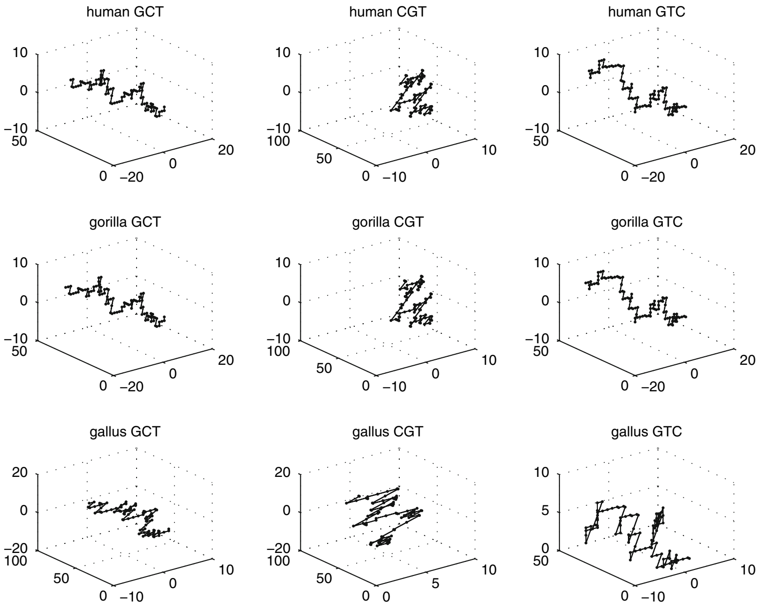

Fig. 4 3DD-curves of human, gorilla and gallus with $u = 2$

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. But the calculation of leading eigenvalue is not easy. In order to avoid tolerance between the two methods although the calculation of the average matrix element associated with a DNA sequence is not difficult, tolerance become insufficient for characterization of DNA sequences.

Here we propose a new descriptors for the characterization of 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 [5,8]. 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

$$Inv(M) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\right).$$

It is easy to see that $Inv(M) > Aver(M)$.

For any 3DD-Curve of DNA sequences (suppose $u = 2$), we have a set of points (x_i, y_i, z_i), i=1, 2, 3,...,n, where n is the length of the sequence. we construct the quotient matrix E/P and E/G [5,8]. The (i,j) element of matrix E/P is defined to be the quotient of the Euclidean-distance between vertices i and j of the 3DD-Curve and the sum of the distances between the same pair of vertices. In other words, $[E/P]_{ij} =$

Table 2 Three invariants of matrix E/G for the first exon of β-globin gene of 11 different species

Species	GCT			CGT			GTC		
	Inv	λ	$Aver$	Inv	λ	$Aver$	Inv	λ	$Aver$
Human	50.8098	50.5208	50.2575	64.7751	64.4969	64.0710	54.7415	54.4868	54.1465
Goat	49.2782	48.9757	48.7052	66.5346	66.3060	65.7610	51.3524	51.3568	50.7553
Opossum	60.3218	60.5034	59.6662	65.5430	65.3151	64.8306	60.9615	60.9226	60.2989
Gallus	58.9713	58.5853	58.3303	66.7802	66.2700	66.0544	50.2114	49.8940	49.6656
Lemur	51.0772	50.8132	50.5220	69.1748	69.1472	68.4229	59.0077	58.7506	58.3663
Mouse	53.8802	53.7226	53.3070	63.9319	63.7253	63.2517	55.9195	55.5627	55.3246
Rabbit	49.9799	49.6728	49.4245	69.3207	69.2557	68.5505	55.2517	55.1027	54.6378
Rat	53.6766	53.4091	53.0932	68.2890	67.8751	67.5468	59.4821	59.2256	58.8355
Gorilla	51.6181	51.2988	51.0631	66.6796	66.4677	65.9627	54.3140	53.9907	53.7299
Bovine	48.7805	48.5029	48.2133	66.6170	66.4062	65.8424	52.8545	52.8449	52.2399
Chimpanzee	56.7353	56.4102	56.1950	74.9612	74.6685	74.2473	60.8957	60.5296	60.3158

$[ED]_{ij}/\sum_{k=i}^{j-1}[ED]_{k,k+1}$, where $[ED]_{ij}$ is the Euclidean distance between a pair of vertices and the (i,j) element $[E/G]_{ij}$ of matrix E/G is defined to be $[ED]_{ij}/|i-j|$.

For these concrete matrices associated with a DNA sequence, Inv happened to be the sum of the average Euclidean distance from any point to all other points on the 3DD-Curves of the DNA sequence. So Inv can be regarded as a invariant of DNA sequences. Clearly, Inv is simple for calculation and thus facilitated for characterization of DNA sequences. In Table 2, 3, we list the Inv, λ, $Aver$ of matrix E/G and E/G for the first exon of β-globin gene of 11 different species, respectively. Observing Table 2, 3, we can see that the relative difference between Inv and λ is less than that between λ and $Aver$ and find that the Inv is slightly bigger than the corresponding leading eigenvalue. we wonder whether this is always true for real symmetric matrix whose diagonal entries are zero. We find the observation that $Inv > \lambda > Aver$ for all matrices E/G and E/G. From it follows that λ is approximately given by $(Inv + Aver)/2$. Define

$$INV(M) = (Inv(M) + Aver(M))/2,$$

where M is E/G or E/G.

In Table 4, 5, we show the INV and λ of matrix E/G for the first exon of β-globin gene of 11 different species and that of matrix E/G. Observing these tables, we can see that the INV has nice approach to corresponding leading eigenvalue. Both INV of matrix E/G and that of matrix E/G can be used to approach the corresponding leading eigenvalue in the analysis of similarities and dissimilarities of DNA sequence and the calculation of INV is very simple.

In order to compare DNA sequences, we can construct a 3-component vector made by using the invariant INV of 3 matrices obtained from the three 3DD-curves which corresponding pattern GCT, CGT and GTC of the first exon of β-globin gene of 11 dif-

Table 3 Three invariants of matrix E/P for the first exon of β-globin gene of 11 different species

Species	GCT			CGT			GTC		
	Inv	λ	*Aver*	*Inv*	λ	*Aver*	*Inv*	λ	*Aver*
Human	40.2845	39.9900	39.8467	50.8835	50.5444	50.3304	43.2165	42.9178	42.7468
Goat	38.4265	38.1233	37.9797	51.1145	50.6476	50.5202	39.7377	39.5007	39.2756
Opossum	44.4339	44.1291	43.9509	48.4178	48.0251	47.8915	45.0672	44.6910	44.5773
Gallus	45.3313	44.9238	44.8385	51.1859	50.6922	50.6296	38.6675	38.3498	38.2472
Lemur	39.3787	39.0788	38.9506	52.7195	52.4323	52.1465	45.2443	44.8488	44.7525
Mouse	42.6851	42.4020	42.2310	50.3465	50.0260	49.8109	44.3665	43.9975	43.8945
Rabbit	39.0170	38.7003	38.5834	53.3279	52.9788	52.7354	42.8044	42.4753	42.3288
Rat	40.6384	40.2898	40.1967	51.5571	51.0967	50.9967	44.9639	44.5918	44.4752
Gorilla	40.9438	40.6298	40.5035	52.3758	51.0811	51.8126	42.9507	42.6221	42.4888
Bovine	38.0933	37.8283	37.6503	51.2133	50.7907	50.6178	40.9598	40.7464	40.4836
Chimpanzee	45.0622	44.7586	44.6330	59.1053	58.7797	58.5424	48.2439	47.8964	47.7844

Table 4 INV and λ of matrix E/G for the first exon of β-globin gene of 11 different species

Species	GCT		CGT		GTC	
	INV	λ	*INV*	λ	*INV*	λ
Human	50.5337	50.5208	64.4231	64.4969	54.4440	54.4868
Goat	48.9917	48.9757	66.1478	66.3060	51.0538	51.3568
Opossum	59.9940	60.5034	65.1868	65.3151	60.6302	60.9226
Gallus	58.6508	58.5853	66.4173	66.2700	49.9385	49.8940
Lemur	50.7996	50.8132	68.7989	69.1472	58.6870	58.7506
Mouse	53.5936	53.7226	63.5918	63.7253	55.6220	55.5627
Rabbit	49.7022	49.6728	68.9356	69.2557	54.9447	55.1027
Rat	53.3849	53.4091	67.9179	67.8751	59.1588	59.2256
Gorilla	51.3406	51.2988	66.3212	66.4677	54.0219	53.9907
Bovine	48.4969	48.5029	66.2297	66.4062	52.5472	52.8449
Chimpanzee	56.4651	56.4102	74.6042	74.6685	60.6058	60.5296

ferent species. 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 point to a similar direction in the 3D-space and have similar magnitudes. The similarity between these two vectors can be measured by calculating the Euclidean distance between their end points. Clearly, the smaller is the Euclidean distance the more similar are the two DNA sequences.

Table 5 INV and λ of matrix E/P for the first exon of β-globin gene of 11 different species

Species	GCT		CGT		GTC	
	INV	λ	INV	λ	INV	λ
Human	40.0656	39.9900	50.6069	50.5444	42.9816	42.9178
Goat	38.2031	38.1233	50.8173	50.6476	39.5067	39.5007
Opossum	44.1924	44.1291	48.1546	48.0251	44.8223	44.6910
Gallus	45.0849	44.9238	50.9078	50.6922	38.4573	38.3498
Lemur	39.1646	39.0788	52.4330	52.4323	44.9984	44.8488
Mouse	42.4581	42.4020	50.0787	50.0260	44.1305	43.9975
Rabbit	38.8002	38.7003	53.0317	52.9788	42.5666	42.4753
Rat	40.4176	40.2898	51.2769	51.0967	44.7195	44.5918
Gorilla	40.7236	40.6298	52.0942	51.0811	42.7197	42.6221
Bovine	37.8718	37.8283	50.9155	50.7907	40.7217	40.7464
Chimpanzee	44.8476	44.7586	58.8238	58.7797	48.0142	47.8964

As for analysis of the similarities and dissimilarities for 11 coding sequences that based on Euclidean distances between the end points of the 3-component vectors of the invariant INV of the E/P matrices and E/G matrices, the methods and results are much similar with that in [12,13]. Here we don't further discuss.

5 Conclusions

In this paper, a new 3DD-Curve for Visualizing DNA sequences have been constructed. it is very easy to observe the similarity and difference between these sequences. We also propose a new invariant of DNA sequences based on the 3DD-Curve. the proposed 3DD-Curve and invariant successfully demonstrates the effectiveness for sequence visualization and comparison. It has been tested on several DNA sequences and the results have been verified to match results reported in the literature.

Acknowledgments The authors would like to thank the referees for several helpful suggestions and the Shandong Natural Science Foundation(Y2006A14) for supporting this research.

References

1. A. Arneodo, Y. d'Aubenton Carafa, B. Audit, E. Bacry, J.F. Muzy, C. Thermes, Physica A **249**, 439–448 (1998)
2. M.A. Gates, J. Theor. Biol. **119**, 319–328 (1986)
3. A. Nandy, Curr. Sci. **66**, 821 (1994)
4. P.M. Leong, S. Morgenthaler, Comput. Appl. Biosci. **12**, 503–511 (1995)
5. M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. **368**, 1–6 (2003)
6. Y. Zhang, B. Liao, K. Ding, Chem. Phys. Lett. **411**, 28–32 (2005)
7. B. Liao, Chem. Phys. Lett. **401**, 196–199 (2005)
8. Y. Wu, A.W. Liew, H. Yan, M. Yang, Chem. Phys. Lett. **367**, 170–176 (2003)

9. M. Randic, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 1235–1244 (2000)
10. E. Hamori, J. Ruskin, J. Biol. Chem. **258**, 1318–1327 (1983)
11. M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. **371**, 202–207 (2003)
12. C. Li, J. Wang, Combin. Chem. High Throughput Screen. **7**, 23–27 (2004)
13. M. Randic, X.F. Guo, S.C. Basak, J. Chem. Inf. Comput. Sci. **41**, 619–626 (2001)
14. M. Randic, J. Chem. Inf. Comput. Sci. **40**, 50–56 (2000)
15. B. Liao, T. Wang, J. Mol. Struc.: THEOCHEM **681**, 209–212 (2004)
16. Y. Zhang, B. Liao, K. Ding, Mol. Simul. **32**, 29–34 (2006)