MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 493-512

ISSN 0340 - 6253

Analysis of Similarities/Dissimilarities of DNA Sequences Based on a Novel Graphical Representation

Jia-Feng Yu [a, b], Ji-Hua Wang [b], Xiao Sun [a, *]

[a] *State Key Laboratory of Bioelectronics*, *School of Biological Science and Medical Engineering*, *Southeast University*, *Nanjing*, *Jiangsu 210096*, *China*

[b] *Key Laboratory of Biophysics in Universities of Shandong, Department of Physics*, *Dezhou University*, *Dezhou*, *Shandong 253023*, *China*

e–mail: **jfyu1979@126.com, jhwyh@yahoo.com.cn, xsun@seu.edu.cn.**

(Received September 11, 2009)

Abstract

According to the physiochemical property of the base at the first site, the 16 kinds of dinucleotides are classified into four groups. Based on such classification, we propose a novel graphical representation of DNA sequence without loss of information due to overlapping and crossing of the curve with itself. This representation allows direct inspection of compositions and distributions of dinucleotides and visual recognition of similarities/dissimilarities among different sequences. A 6D vector is exploited as quantitative descriptor from this representation, which can display both the global and local features of DNA sequences in a 6D phase space. The applications in similarities/dissimilarities analysis of the complete coding sequences of β globin genes of eleven species illustrate their utilities.

1. INTRODUCTION

Sequence alignment is the basic studying strategy for both DNA and protein sequences analysis in bioinformatics. With the exponential growth of DNA sequences resulted from the developments of sequencing techniques, many scientists from all kinds of researching fields are attracted to exploit the secrets of life. However, it is difficult to obtain biological informat-

* Corresponding author.

ion directly from these sequences composed of only four kinds of characters A, G, C and T. Facing such complicated sequences, one has to integrate several tools to do simple analysis of DNA sequences in many cases. Recently, graphical representations are well–regarded which can not only transform DNA sequences into visual curves but also offer effective numerical descriptors. Because of its convenience and excellent maneuverability, methods based on graphical representation have been extensively applied in relevant realms of bioinformatics. In 1983, Hamori and Ruskin firstly proposed a graphical representation to describe DNA sequence [1]. Since then, quite a few models based on different mechanisms have been outlined. According to the dimensions of the space in which the sequences are plotted, all the graphical representations can be classified into five categories ranging from 2D to 6D [2]. But this means of classification can not exhibit the essences of these models, because the dimensions are changeable sometimes in practical applications. In this paper, we classify these representations into three categories according to their biological basis, i.e., individual nucleotide [3–15], dinucleotide [16–20] and trinucleotide [21, 22], respectively.

Among all the representations based on individual nucleotides, Z curve is the most successful one [23, 24]. For a DNA sequence with N bases, the cumulative occurrence numbers of A, G, C and T are denoted by A_n, G_n, C_n and T_n, respectively. Then, the Z–transform can be accomplished by the following equations [3, 4]:

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n) \\ y_n = (A_n + C_n) - (G_n + T_n) \\ z_n = (A_n + T_n) - (G_n + C_n) \end{cases} \quad (n=1, 2, \ldots, N).$$

In this way, a DNA sequence is converted into a unique curve in 3D space based on plot sets (x_n, y_n, z_n). It seems that Z curve does not lose any biological information of the sequence, because it especially uses the classifications of chemical structure on purines/pyrimidines (A, G)/(C, T), amino/keto groups (A, C)/(G, T) and strong–weak hydrogen bonds (A, T)/(G, C), and one can recover the original DNA sequence from Z curve.

If assigning (1, 1), (−1, 1), (−1, −1) and (1, −1) to represent A, G, C and T, respectively, the

four kinds of bases ban be distributed into the four quadrants of Cartesian 2D coordinates as shown in Fig.1.

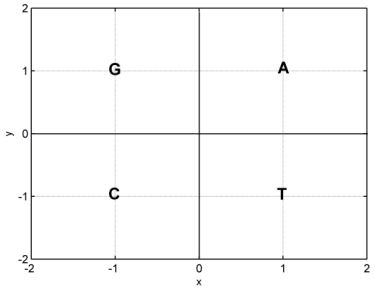


Figure1: distribution of the four kinds of bases in Cartesian 2D coordinates

In Fig.1, each kind of base is numerically represented by a 2D coordinate (x, y). Obviously, if x>0, the corresponding base must be A or T, otherwise G or C, then x>0/x<0 correspond to weak H–bond/strong H–bond groups. Similarly, y>0 denotes that the corresponding base is A or G, otherwise C or T, then y>0/y<0 correspond to purine/pyrimidine groups. In this way, an arbitrary DNA sequence can be transformed into a 3D curve by plotting sets (x_n, y_n, n). Here, (x_n, y_n) are the assignments in Fig.1, n=1, 2, 3, …, N, N is the length of given sequence. Obviously, the relation between the primary sequence and the 3D curve is unique. This kind of graphical representation is called walking model. Supposing z=x*y, x and y are positive or negative synchronously when z>0, which denotes the corresponding base is A or C according to Fig.1. Similarly, x and y are opposite numbers when z<0, the corresponding base is G or T. Therefore, z corresponds to the amino/keto groups. Now, we define the following three parameters:

$$u_{x,n}=\sum_{i=1}^{n}x_i\,,u_{y,n}=\sum_{i=1}^{n}y_i\,,u_{z,n}=\sum_{i=1}^{n}z_i\,,\ n=1,\,2,\,3,\,\ldots,\,N.$$

Meaningfully, it is found that $u_{x,n}$, $u_{y,n}$ and $u_{z,n}$ are equivalent to the three components z_n, x_n and y_n of Z curve, respectively. Therefore, Z curve is also a walking model from this viewpoint. Besides Z curve, some other models based on individual nucleotides are also

proposed [3–15]. Nandy [5] constructed a graphical model by assigning A, G, T and C to the four directions, $(-x)$, $(+x)$, $(-y)$ and $(+y)$, respectively, along the positive and the negative Cartesian coordinate axes. But such a representation of DNA is accompanied by some loss of information associated with crossing and overlapping of the resulting curve by itself. Recently, a model based on double vectors [15] is proposed to represent DNA sequence without loss of information, with which the author apply to do analysis of similarities/dissimilarities among different sequences, but there are some disappointed results in the similarities matrix.

Comparing with individual nucleotide, dinucleotide and trinucleotide have more advantages in sequence analysis [25–27]. Regretfully, those models based on individual nucleotides cannot represent dinucleotide and trinucleotide directly. To do this work, complicated statistics have to be done [24], some models even can not be used to denote the information of dinucleotide and trinucleotide at all. Because of the limitation of visualization, models based on trinucleotides are fewer [22]. Recently, some researchers outlined different graphical representations based on dinucleotides and apply them in similarity analysis of different sequences [16–20], but most of the calculations are complex and their initial assignments cannot reflect the distributions of dinucleotides.

Motivated by previous works, we propose a novel graphical representation without loss of information based on dinucleotides in this paper. Comparing with other models, this novel representation has following merits: (a). This model allows direct inspection of compositions and distributions of dinucleotides in DNA sequences. (b). It has excellent maneuverability. From this representation, quite a few alternative parameters can be deduced, which can be used to denote global and local information of DNA sequence. (c). This representation provides more information. Based on this representation, a simple approach is outlined for analysis of similarities/dissimilarities of DNA sequences among different species based on a 6D vector, which does not need complex calculations. This novel representation can provide convenient tools of sequence analysis for both computational scientists and molecular biologists.

2. CONSTRUCTION OF THE NOVEL GRAPHICAL REPRESENTATION

The four kinds of bases A, G, C and T can buildup $4^2 = 16$ kinds of dinucleotides. According to the category of the base at the first site of dinucleotide, the 16 kinds of dinucleotides can be divided into the four quadrants of a Cartesian 2D coordinates, as shown in Fig.2.

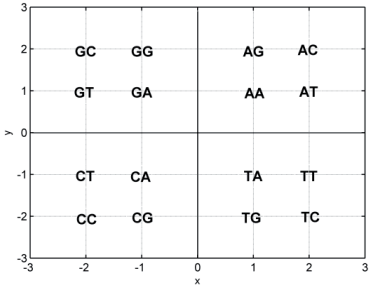


Figure 2: 16 kinds of dinucleotides distributed in Cartesian 2D coordinates

In this way, each kind of dinucleotide is numerically represented by a 2D coordinate (x, y). The signs of x and y are decided by the category of the base at the first site of dinucleotide, that is, {+, +}→A, {−, +}→G, {−, −}→C and {+, −}→T. The absolute values of x and y are decided by the base at the second site, that is, (|x|=1, |y|=1)→A, (|x|=1, |y|=2)→G, (|x|=2, |y|=2)→C, (|x|=2, |y|=1)→T. An intact dinucleotide is described by integrations of x and y. Take (−2, 1) for example, the negative sign of x and the positive sign of y denote the base at the first site of corresponding dinucleotide is G, the absolute value of x and y are 2 and 1, respectively, which denote the base at the second site is T, then (−2, 1) represents the dinucleotide of GT.

Now, let's consider all the dinucleotides of an arbitrary DNA sequence. Supposing $S = s_1s_2s_3\ldots s_N$ is a DNA sequence with N bases, then the number of dinucleotides is $N-1$. We have plot sets $\phi(S) = \phi(s_1s_2)\phi(s_2s_3)\ldots\phi(s_ns_{n+1})\ldots$ to convert S into a 3D curve, where, $\phi(s_ns_{n+1}) = (x_n, y_n, n)$, (x_n, y_n) is the 2D coordinates of dinucleotide of s_ns_{n+1} as introduced in Fig.2, and

n=1,2,3,…, N−1. The curve composed of all the dots of ϕ is the novel graphical representation (for convenience, we call it D–curve in this paper, i.e., curve based on dinucleotides). In Table 1, corresponding values of x, y and n of sequence ATGGTGCACC are listed, and Fig.3 presents its D–curve.

Table 1: values of corresponding parameters of D–curve of sequence ATGGTGCACC

Dinucleotides	x	y	n	z	x'	y'	z'
AT	2	1	1	2	2	1	2
TG	1	−2	2	−2	3	−1	0
GG	−1	2	3	−2	2	1	−2
GT	−2	1	4	−2	0	2	−4
TG	1	−2	5	−2	1	0	−6
GC	−2	2	6	−4	−1	2	−10
CA	−1	−1	7	1	−2	1	−9
AC	2	2	8	4	0	3	−5
CC	−2	−2	9	4	−2	1	−1

From Fig.2, we find that when x>0, the first base of dinucleotide is A or T, while when x<0, the first base is G or C, and then x divides the 16 kinds of dinucleotides into two groups, i.e., weak H–bond/strong H–bond groups. Similarly, y>0 denotes the first base of dinucleotide is A or G, while y<0 denotes the first base is C or T, and then the 16 kinds of dinucleotides are also divided into two groups by y, i.e., purine/pyrimidine groups. Following the similar steps in the section 1, we define z= x*y and

$$x'_n = \sum_{i=1}^{n} x_i \,, \quad y'_n = \sum_{i=1}^{n} y_i \,, \quad z'_n = \sum_{i=1}^{n} z_i \,, \; n=1, 2, \ldots, N-1.$$

Obviously, when z>0, the first base of dinucleotide is A or C, while when z<0, the corresponding base is G or T, then z divides the 16 kinds of dinucleotides into two groups, i.e., amino/keto groups. In conclusion, the 16 kinds of dinucleotides can be classified into two groups in three ways by x, y and z, respectively, as presented in Table 2. Therefore, we can obtain detail information of dinucleotides from x, y and z, and x', y' and z' embody their cumulative effects, respectively, which can be used to exhibit the local and global features of corresponding sequence. The values of z, x', y' and z' of sequence ATGGTGCACC are also

presented in Table 1.

According to the construction of D–curve, a random DNA sequence can be converted into a unique 3D curve containing no loops based on (x, y, n), which avoids loss of information due to overlapping and crossing with itself. Besides, one can also obtain other forms of curves based on the alternative invariants inferred from D–curve, such as plot sets (x', y', z'), and so on. However, we notice that the initial assignments analogous to Fig.2 are not unique, according to statistical theory, there are $4!\times4!=576$ kinds of assignments. That is, different curves can be obtained by different initial assignments. Nevertheless, we can only obtain a unique 3D curve of DNA sequence according to the designation in Fig.2. Since the zigzag curve does not represent the genuine molecular geometry, we are not interested in the unique relationships between the initial assignments and the possible numbers of 3D curve, but are interested in its numerical representations that may facilitate analysis of DNA sequences.

Table 2: the 16 kinds of dinucleotides classified into two groups in three ways by x, y and z, respectively.

Groups	x				Groups	y				Groups	z			
Weak	AA	AG	AC	AT	Purine	AA	AG	AC	AT	Amino	AA	AG	AC	AT
H–bond	TA	TG	TC	TT		GA	GG	GC	GT		CA	CG	CC	CT
Strong	GA	GG	GC	GT	Pyrimidine	CA	CG	CC	CT	Keto	GA	GG	GC	GT
H–bond	CA	CG	CC	CT		TA	TG	TC	TT		TA	TG	TC	TT

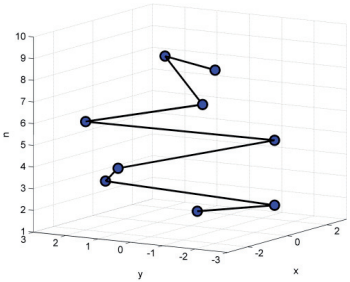


Figure 3: D–curve of sequence ATGGTGCACC.

3. ANALYSIS OF SIMILARITIES/DISSIMILARITIES AMONG DIFFERENT DNA SEQUENCES

Analysis of similarities/dissimilarities of DNA sequences among different species is one of the main motivations of graphical representations, as is reflected by related papers [10–13, 15–22, 28–31]. In previous works, most researchers applied their approaches to the coding sequences of the first exon of β globin genes. Nandy and his partners suggested that researchers should apply their graphical techniques to complete genes, or at least to the complete coding sequences, so that an unambiguous point of contact is available for comparing to the real world [2]. In this section, we perform similarities analysis on the complete coding sequences of β globin genes among 11 species based on D–curve. Table 3 lists the information of corresponding sequences.

Table 3: the complete coding sequences of β globin genes of 11 species

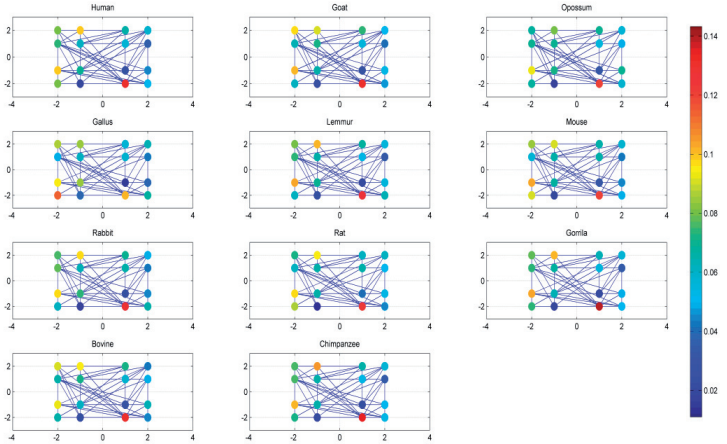
Species	NCBI ID	Location of each exon	Length of Complete CDS (bp)
Human	U01317	62187...62278, 62409...62631, 63482...63618	444
Goat	M15387	279...364, 493...715, 1621...1749	438
Opossum	J03643	467...558, 672...894, 2360...2488	444
Gallus	V00409	465...556, 649...871, 1682...1810	444
Lemur	M15734	154...245, 376...598, 1467...1595	444
Mouse	V00722	275...367, 484...705, 1334...1462	444
Rabbit	V00882	277...368, 495...717, 1291...1419	444
Rat	X06701	310...401, 517...739, 1377...1505	444
Gorilla	X61109	4538...4630, 4761...4982, 5833...5881	364
Bovine	X00376	278...363, 492...714, 1613...1741	438
Chimpanzee	X02345	4189...4293, 4412...4633, 5484...5532	376

3.1. Graphical alignments of different DNA sequences

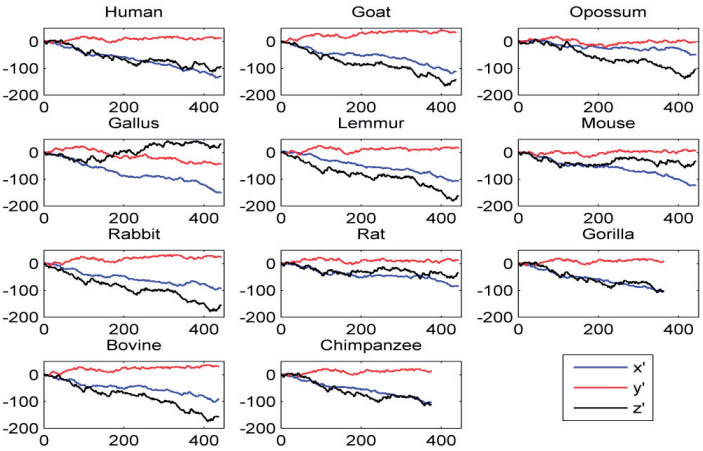
For a given DNA sequence, we can obtain its D–curve by plot sets (x, y, n). Projecting this 3D curve into the 2D x–y coordinates, one can direct observe the compositions and densities of all kinds of dinucleotides. Fig.4A shows the projections of D–curves of the 11 species presented in Table 3, where the compositions and densities of all the 16 kinds of dinucleotides in the primary sequence are marked by different colors. A close observation of Fig.4A shows

that most of the 11 sequences are rich in dinucleotide of TG, while lack of TA and CG, information of other dinucleotides can also be inspected intuitively according to the colorbar. In addition, we can see that Gorilla and Chimpanzee have the most similar compositions and densities of dinucleotides, then they have closest evolutionary distance, which is accordant with actual evolutionary evidence, while Opossum and Gallus have the most dissimilar compositions of dinucleotides with other species, this is also coincident with the fact that Gallus is non–mammal and Opossum is the most remote species from others. The evolutionary correlations among other species can also be inferred from Fig.4A. Furthermore, the lines linking the neighboring dinucleotides can imply their co–occurrence frequencies and correlations (the numbers of the lines observed may be less than practical situations because of overlapping), too.

The results of Fig.4A indicate that D–curve can provide convenient tool for exhibiting the compositions and densities of dinucleotides. Next, we introduce another utility of this novel representation in exhibiting distributions of dinucleotides in DNA sequence. As have been discussed in Table 2, x, y and z denote three types of groups of dinucleotides, which are weak H–bond/strong H–bond groups, purine/pyrimidine groups and amino/keto groups, respectively. Thereafter, we defined x', y' and z' as their accumulative effects. In Fig.4B, we present the 2D curves based on x', y' and z' vs. n of the 11 species, from which we can observe the distributions of each group of dinucleotides according to the fluctuations of the curves. Take Human for example, the curve based on x' decrease monotonously, which is caused by the dominant $x<0$, then the corresponding sequence must be richer in dinucleotides of strong H–bond group, and the results of sequence analysis show that the percentage of weak H–bond group is 43.57% and that of strong H–bond group is 56.43%. In the similar way, we can analyze other curves, such as curves based y' and z'. Still taking Human for example, the curve based on y' is almost a horizontal line around $y'=0$, which indicates that the content of dinucleotides of purine group is approximately equal to that of pyrimidine group, then this is consistent with the fact of 50.34% (percentage of purine group) and 49.66% (percentage of pyrimidine group); comparing with the former two kinds of curves,



A



B

Figure 4: Graphical representations of the complete coding sequences of β globin genes of 11 species. A: The projections of the D–curves on 2D x–y coordinates; B: The 2D curves based on x', y' and z'.

curve based on z' vibrates more sharply in some local regions, but the decreasing trend is still obvious, which indicates that the corresponding sequence is richer in dinucleotides of keto group, this is also consistent with the actual instances of 45.37% (percentage of amino group) and 54.63% (percentage of keto group). Among all the 11 species, the curve of Gallus based on z' fluctuate most greatly, from which we can find there is a decreasing trend before the position about n=100, then the former fragment with 100 bases is richer in keto group, while after n=100, the slope changes suddenly from negative to positive and the curve fluctuates occasionally with obvious increasing trend, then the amino group is becoming more dominant gradually. Furthermore, we can compare the distributions of each group of dinucleotides according to the magnitudes of the slopes of corresponding curves. For example, the curves based on x' of all the 11 species have trends of decreasing monotonously, from Fig.4B, one can easily find that Opossum has the smallest slope, which indicates its percentage of dinucleotides of strong H–bond group is the smallest, just the opposite, the slope of Gallus is the biggest one, which indicate its percentage of dinucleotides of strong H–bond group is the biggest, these results are entirely consistent with the actual situations (the percentages of strong H–bond group of the 11 species are listed in turn: Human–56.43%, Goat–55.38%, Opossum–51.69%, Gallus–59.37%, Lemur – 55.76%, Mouse – 55.76%, Rabbit – 54.18%, Rat – 53.27%, Gorilla – 56.20%, Bovine – 54.01%, Chimpanzee – 56.27%). According to the discussions above, we can also recognize similarities/dissimilarities among these sequences from the curves based on x', y' and z' in Fig.4B. A close observation shows that Human – Gorilla, Human – Chimpanzee, Gorilla – Chimpanzee are the most similar species, while Opossum and Gallus are the most dissimilar species with others, evolutionary correlations among other species can also be inferred in the same way, and these results coincide with that of Fig.4A perfectly.

3.2. Quantitative analysis of Similarities/dissimilarities of DNA sequences with numerical descriptors

Up to now, there are quite a few algorithms and programs based on different scoring matrices for sequence alignment, such as BLAST [36] and Clustal [37], which can

quantitatively describe similarities/dissimilarities of DNA sequences. Along with the extensive application of graphical representation in bioinformatics, more and more authors apply their presentations to numerically analyze similarities among different sequences. For example, Randic et al. have proposed E matrix, M/M matrix, L/L matrix and L^k/L^k matrix, and used their eigenvalues as descriptors to do analysis of similarities/dissimilarities of DNA sequences [6]. These methods were proved to be useful and used by many authors. However, these matrices become too large to calculate the eigenvalues when DNA sequence is very long, and the computations are very complex. Furthermore, there is some loss of information associated with these matrices [8]. Then, exploiting more convenient and precise methods to analyze similarities of different DNA sequences is necessary, which is helpful in evolutionary related researches. In this section, we put emphasis on the utilities of D–curve as numerical descriptors in quantitative analysis of similarities/dissimilarities among the 11 DNA sequences presented in Table 3.

The results of Fig.4 have validated the efficiencies of the invariants deduced from D–curve in qualitatively representing DNA sequences. With these parameters, one can not only obtain the information of compositions but also distributions of dinucleotides, which enable us to extract efficient numerical descriptors to represent DNA sequence. In section 2, we discussed the significances of x, y, z and x', y', z'. Here, we employ a vector composed of six components $\{m_x, m_y, m_z, m_{x'}, m_{y'}, m_{z'}\}$ as quantitative descriptor of DNA sequence, namely,

$$m_L = \frac{1}{N-1}\sum_{n=1}^{N-1} L_n ,$$

where m_L is the mean value of corresponding component, $L \in \{x, y, z, x_n', y_n', z_n'\}$. The underlying assumption is that if two vectors point to a similar direction, the corresponding DNA sequences are similar. Using Euclidean distance as criterion of sequence similarities/dissimilarities, the smaller the Euclidean distance is, the more similar the DNA sequences are. That is to say, the distances between evolutionary closely related species are smaller, while those between evolutionary disparate species are larger. The Euclidean

distance between two sequences can be defined as follows:

$$D(S_i, S_j) = \sqrt{\sum_{v=1}^{6} (V_v^{S_i} - V_v^{S_j})^2} \ .$$

Where, $V_v^{S_i}$ and $V_v^{S_j}$ are the vth component of the 6D vector V of sequences S_i and S_j, respectively. In this way, we obtain the matrix of similarities/dissimilarities of the 11 species presented in Table 4.

From Table 4, we find Gallus (the only non–mammal among them) and Opossum (the most remote species from the remaining mammals) are most dissimilar to others among the 11 species. On the other hand, Gorilla–Chimpanzee has the smallest distance, so they are the most similar species pairs. Take Human for example, Human–Gorilla, Human–Chimpanzee have smaller distance, so they are more similar species pairs. Evolutionary correlations among other species can also be obtained, these results coincide with Fig.4, and similar results are also obtained in recent papers by different approaches [18–20, 22, 32–35].

Table 4: matrix of similarities/dissimilarities of the complete coding sequences of β globin genes of 11 species based on 6D vector.

Species	Huma	Goat	Opossu	Gallu	Lemu	Mous	Rabbit	Rat	Gorill	Bovin	Chimpanze
Human	0	31.39	48.701	70.46	31.75	30.27	35.575	41.65	13.63	30.68	14.00
Goat		0	52.295	99.82	17.81	56.80	12.648	59.75	35.60	9.001	28.078
Opossum			0	85.79	41.98	44.69	51.136	38.80	37.57	43.88	35.876
Gallus				0	98.86	45.43	104.99	53.20	65.84	97.84	74.025
Lemur					0	54.38	10.123	58.59	33.74	11.17	25.943
Mouse						0	61.183	19.43	21.41	53.43	29.424
Rabbit							0	65.42	39.91	11.26	32.037
Rat								0	29.19	55.56	34.822
Gorilla									0	32.18	8.185
Bovine										0	24.071
Chimpanze											0

The main motivation of D–curve is to describe the information of dinucleotides of DNA sequence. To further test the correlation between D–curve and actual information of dinucleotides, we outline a 16D vector based method to analyze similarities/dissimilarities

among different sequences, which is defined as follows:

$$D(S_i, S_j) = \sqrt{\sum_{u=1}^{16} (P_u^{S_i} - P_u^{S_j})^2} \ .$$

Where, $P_u^{S_i}$ and $P_u^{S_j}$ are the usage frequency of each kind of dinucleotide in sequences S_i and S_j, respectively. For example, the occurring times of dinucleotide of GT in sequence ATGGTGCACC is 2, and there are altogether 9 dinucleotides, then the usage frequency of GT is 2/9 = 0.22. Based on this 16D vector, we calculate the matrix of similarities/dissimilarities of the complete coding sequences of β globin genes of the 11 species, and the results are presented in Table 5.

From Table 5, we find Gallus and Opossum are most dissimilar to others among the 11 species. On the other hand, Gorilla–Chimpanzee has the smallest evolutionary distance. Take Human for example, Human–Gorilla, Human–Chimpanzee have smaller distance, so they are more similar species pairs. Evolutionary distances among other species can also be obtained; these results are perfectly consistent with that of Table 4.

Table 5: matrix of similarities/dissimilarities of the complete coding sequences of β globin genes of 11 species based on 16D vector.

Species	Huma	Goat	Opossu	Gallus	Lemu	Mous	Rabbi	Rat	Gorill	Bovin	Chimpanze
Human	0	0.039	0.0481	0.071	0.031	0.031	0.034	0.035	0.018	0.0442	0.0188
Goat		0	0.0475	0.085	0.034	0.044	0.043	0.051	0.036	0.0242	0.039
Opossum			0	0.083	0.042	0.052	0.045	0.045	0.048	0.0416	0.0475
Gallus				0	0.083	0.055	0.085	0.071	0.082	0.09	0.0828
Lemur					0	0.050	0.026	0.054	0.026	0.0415	0.0224
Mouse						0	0.051	0.025	0.041	0.0479	0.0442
Rabbit							0	0.052	0.035	0.0451	0.0338
Rat								0	0.044	0.0468	0.0451
Gorilla									0	0.0411	0.0099
Bovine										0	0.0422
Chimpanze											0

To give a quantitative index describing the efficiency of D–curve, we calculate the Pearson's correlated coefficient (PCC) between Table 4 and Table 5. The PCC is a measure of

the correlation (linear dependence) between two variables or matrices X and Y, giving a value between +1 and −1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables or matrices [38], which is defined as

$$PCC(X,Y) = \frac{\sum XY - \frac{1}{N^2}\sum X \sum Y}{\sqrt{(\sum X^2 - \frac{1}{N^2}(\sum X)^2)(\sum Y^2 - \frac{1}{N^2}(\sum Y)^2)}},$$

N is the number of the samples of matrices X and Y, here, the value of N for Tables 4 and 5 is 11. The calculating result shows PCC = 0.9032 and P–value = 1.54×10^{-45}, that is, the efficiencies of the equations based on the 6D and 16D vectors are equivalent, which further verify the ability of D–curve in exhibiting the information of dinucleotides.

Table 6: scoring matrix of the complete coding sequences of β globin genes of 11 species based on ClustalW2.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	100	84	73	74	84	82	88	81	99	86	96
Goat		100	71	70	83	79	84	79	82	94	80
Opossum			100	70	69	72	73	71	72	71	70
Gallus				100	71	72	72	71	74	71	72
Lemur					100	78	84	78	84	83	81
Mouse						100	81	89	82	79	79
Rabbit							100	80	88	85	86
Rat								100	81	78	78
Gorilla									100	85	99
Bovine										100	82
Chimpanzee											100

The basis of the methods for similarities analysis of DNA sequence introduced above is numerical descriptor, which is different from the programs based on scoring matrices. Among the programs on the basis of scoring matrices, Clustal is a typical one. ClustalW2 is the newest release with a general purpose multiple sequence alignment program for DNA or proteins. For comparison, we perform multiple sequence alignment for the 11 sequences with ClustalW2 (http://www.ebi.ac.uk/clustalw/) [37], and the results are presented in Table 6.

Thing to note is that the scoring in Table 6 are opposite to Tables 4 and 5. In the latter, the smaller the Euclidean distance is, the more similar the DNA sequences are, while in Table 6, the bigger the scoring is, the more similar the DNA sequences are. From the scoring matrix, we find the scores of Gallus and Opossum are far less than others, which imply they are most dissimilar to other species. On the other hand, Chimpanzee – Gorilla and Human – Gorilla are the most similar pairs. Human – Chimpanzee and Goat – Bovine have bigger scorings, and then they have smaller evolutionary distances. Other evolutionary correlations among other species can also be inferred from Table 6, which are perfectly consistent with the results obtained in Tables 4 and 5. In addition, we also calculate the Pearson's correlated coefficient (PCC) between Table 4 and Table 6, the result shows PCC = −0.7710, P–value = 4.52×10^{-25}.

3.3. Discussion

In this section, we analyze similarities/dissimilarities among eleven orthologous DNA sequences of the complete coding genes of β globin genes using the novel representation, which comprise two efforts: (a). With the help of the visual function of D–curve, we display the features of dinucleotides of DNA sequences in a visible space, which allows to direct identifying similarities among the studied sequences. (b). Based on D–curve, 6 invariants are exploited as descriptors to numerically represent DNA sequences, which can embody the information of dinucleotides perfectly, as is verified by the comparison of the results of similarities analysis using different algorithms and program. Furthermore, another advantage lies in its excellent operability and accuracy, and then D–curve can provide more convenient tools to facilitate related researches.

4. CONCLUSION

Visualization and numerical descriptors are the essential functions of graphical representations. In the past several years, quite a few presentations have been outlined. Most of the incipient models are based on individual nucleotides, but with the development of bioinformatics, the specific significances of dinucleotides are highlighted for special attentions, whereupon more and more graphical representations based on dinucleotides are

outlined. Considering dinucleotides instead of individual nucleotides has more advantages, which can be seen from recent researching papers [39, 40]. According to the physiochemical property of the base at the first position of dinucleotide, the 16 kinds of dinucleotides are classified into four groups. Based on such classifications, we propose a novel graphical representation of DNA sequence without loss of information due to overlapping and crossing of the curve with itself in this paper. This representation reasonably utilize the positive and negative signs of x and y as well as their derivatives to embody DNA sequences, which can not only direct describe information of dinucleotides without complex calculations but also provide more information, such as compositions and distributions of dinucleotides. From this model, six variables are deduced as quantitative descriptors, which help to exhibit global and local features of DNA sequence. Furthermore, two simple methods based on 6D and 16D vectors are outlined for similarities/dissimilarities analysis among different sequences, respectively, and the results validate the abilities of D–curve in describing DNA sequence quantitatively. On the other hand, we compare the similarities/dissimilarities matrices based on the 6D and 16D vectors with the scoring matrix obtained by ClustalW2, a typical multiple sequence alignment program, and the results show that the evolutionary correlations obtained by the three different methods are perfect consistent. Therefore, this novel representation can provide more effective and convenient tools for sequence analysis, such as sequences alignment, extracting features of DNA sequences.

5. ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation of China (Project No. 60671018 and No. 60121101) and partially supported by National Natural Science Foundation of China (Project No. 30970561). The authors would like to thank the anonymous referees for any valuable suggestions that have improved this manuscript.

References

[1] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (1983) 1318–1327.

[2] A. Nandy, M. Harle, S. C. Basak, Mathematical Descriptors of DNA Sequences: Development and Applications, ARKIVOC. 9 (2006) 211–238.

[3] C.T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucl. Acids Res. 19 (1991) 6313–6317.

[4] R. Zhang, C. T. Zhang, Z curves, an intuitive tool for visualizing and analyzing the DNA sequences, J. Biomol. Struct. Dyn. 11 (1994) 767–782.

[5] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, Curr. Sci. 66 (1994) 309–314.

[6] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2–D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.

[7] M. Randic, M. Vracko, J. Zupan, M. Novic, Compact 2–D graphical representation of DNA, Chem. Phys. Lett. 373 (2003) 558–562.

[8] B. Liao, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3–D graphical representation., Chem. Phys. Lett. 388 (2004) 195–200.

[9] M. Randic, Graphical representations of DNA as 2–D map, Chem. Phys. Lett. 386 (2004) 468–471.

[10] B. Liao, T. M. Wang, 3–D graphical representation of DNA sequences and their numerical characterization, J. Mol. Struct. (Theochem) 681 (2004) 209–212.

[11] R. Chi, K. Q. Ding, Novel 4D numerical representation of DNA sequences, Chem. Phys. Lett. 407 (2005) 63–67.

[12] Y. H. Yao, X. Y. Nan, T. M. Wang, A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences, J. Mol. Struct. (Theochem) 764 (2006) 101–108.

[13] B. Liao, K. Q. Ding, A 3D graphical representation of DNA sequences and its application, Theor. Comput. Sci. 358 (2006) 56–64.

[14] J. Song, H. W. Tang, A new 2–D graphical representation of DNA sequences and their numerical characterization, J. Biochem. Biophys. Methods. 63 (2005) 228–239.

[15] Z. J. Zhang, DV–Curve: a novel intuitive tool for visualizing and analyzing DNA sequences, Bioinformatics. 25 (2009) 1112–1117.

[16] X.Q. Liu, Q. Dai, Z.L. Xiu, T.M. Wang, PNN–curve: A new 2D graphical representation of DNA sequences and its application, J. Theor. Biol. 243 (2006) 555–561.

[17] X. Q. Qi, J. Wen, Z. H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, J. Theor. Biol. 249 (2007) 681–690.

[18] Z. H Qi, T. R. Fan, PN–curve: A 3D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 442 (2007) 434–440.

[19] Z. Cao, B. Liao, R. F. Li, A Group of 3D graphical representation of DNA sequences based on dual nucleotides, Int. J. Quantum Chem. 108 (2008) 1485–1490.

[20] Z. B. Liu, B. Liao, W. Zhu, G. H. Huang, A 2–D graphical representation of DNA sequence based on dual nucleotides and its application, Int. J. Quantum Chem. 109 (2009) 948–958.

[21] B. Liao, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases, J. Chem. Inf. Comput. Sci. 44 (2004) 1666–1670.

[22] J. F. Yu, X. Sun, J. H. Wang, TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, J. Theor. Biol. (In press) doi:10.1016/j.jtbi.2009.08.005.

[23] R. Zhang, C. T. Zhang, Identification of replication origins in archaeal genomes based on the Z–curve method, Archaea 1 (2005) 335–346.

[24] F. B. Guo, H. Y. Ou, C. T. Zhang, ZCURVE: a new system for recognizing protein–coding genes in bacterial and archaeal genomes, Nucl. Acids. Res. 31 (2003) 1780–1789.

[25] C. Workman, A. Krogh, No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution, Nucl. Acids Res. 27 (1999) 4186–4822.

[26] P. Clote, F. Ferre, E. Kranakis, D. Krizanc, Structural RNA has a lower folding energy than random RNA of the same dinucleotide frequency, RNA. 11 (2005) 578–591.

[27] D. R. Forsdyke, Calculation of folding energies of single–stranded nucleic acid sequences: Conceptual issues, J. Theor. Biol. 248 (2007) 745–753

[28] W. Y Chen, B. Liao, Y. H. Liu, W. Zhu, Z. Z. Su, A numerical representation of DNA sequence and its applications, MATCH Commun. Math. Comput. Chem. 60 (2008) 291–300.

[29] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, MATCH Commun. Math. Comput. Chem. 56 (2006) 209–216.

[30] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, MATCH Commun. Math. Comput. Chem. 59 (2008) 647–652.

[31] M. Randic, J. Zupan, D. Vikic–Topic, D. Plavsic, A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences, Chem. Phys. Lett. 431 (2006) 375–379.

[32] Y. Guo, T.M. Wang, A new method to analyze the similarity of the DNA sequences, J. Mol. Struct. (Theochem) 853 (2008) 62–67.

[33] P.A. He, J. Wang, Characteristic sequences for DNA primary sequence, J. Chem. Inf. Comput. Sci. 42 (2002) 1080–1085.

[34] J. Wang, Y. Zhang, Characterization and similarity analysis of DNA sequences based on mutually direct–complementary triplets, Chem. Phys. Lett. 425 (2006) 324–328.

[35] Y. S. Zhang, W. Chen, Invariants of DNA sequences based on 2DD–curves, J. Theor. Biol. 242 (2006) 382–388.

[36] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[37] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics. 23 (2007) 2947–2948.

[38] J. L. Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Statist. 42 (1988) 59–66.

[39] R. H. Baran, H. KO, Detecting horizontally transferred and essential genes based on dinucleotide relative abundance, DNA Res. 15 (2008) 267–276.

[40] G. Q. Liu, H. Li, The correlation between recombination rate and dinucleotide bias in drosophila melanogaster, J. Mol. Evol. 67 (2008) 358–367.