Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
New 3D graphical representation of DNA sequence based on dual nucleotides

Xiao-Qin Qi*, Jie Wen, Zhao-Hui Qi

Department of Computer and Information Engineering, Shijiazhuang Railway Institute, Hebei, Shijiazhuang 050043, People's Republic of China

Received 22 January 2007; received in revised form 26 August 2007; accepted 27 August 2007

Available online 1 September 2007

Abstract

We introduce a 3D graphical representation of DNA sequences based on the pairs of dual nucleotides (DNs). Based on this representation, we consider some mathematical invariants and construct two 16-component vectors associated with these invariants. The vectors are used to characterize and compare the complete coding sequence part of beta globin gene of nine different species. The examination of similarities/dissimilarities illustrates the utility of the approach.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: 3D dual nucleotide curve; 3D quantitative characterization; Euclidean distance; Complete coding sequence; Similarities/dissimilarities

1. Introduction

The number of biological sequences is rapidly increasing in the biological database. It is one of the challenges for biologists to analyze mathematically the large volume of biological sequence data. It is good to use the graphic representation to study complicated biological systems because it can provide an intuitive picture and help people gain useful insights. Similar graphical approaches have also been used to deal with a wide variety of biological problems. For instance, various graphic approaches have been successfully used to study enzyme-catalyzed system (see, e.g., King and Altman, 1956; Chou et al., 1979; Chou and Forsen, 1980; Chou and Liu, 1981; Zhou and Deng, 1984, Chou, 1989, 1990; Kuzmic et al., 1992; Lin and Neet, 1990), protein folding kinetics (Chou, 1990, 1993), condon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), HIV reverse transcriptase inhibition mechanisms (Althaus et al., 1993a-c) and base frequency distribution in the anti-sense strands (Chou and Zhang, 1996). Recently, the images of cellular automata were also used to represent biological sequences (Xiao et al., 2005a), predict protein subcellular location (Xiao et al., 2006a), investigate HBV virus gene missense mutation (Xiao et al., 2005b) and HBV viral infections (Xiao et al., 2006b), as well as analyze the fingerprint of SARS coronavirus (Wang et al., 2005).

As for an important part of graphical techniques, graphical representations of DNA sequences have been proposed by several authors (Zhang, 1991; Nandy, 1994; Nandy and Nandy, 2003; Liao and Wang, 2004; Randic et al., 2003, Liu et al., 2006; Zhang and Chen, 2006). Some of them, for example Nandy' graphical representation (Nandy, 1994), are accompanied by some loss of visual information associated with crossing and overlapping of the curve with itself. In order to avoid the limitations related to crossing and overlapping, Liao (Liao and Wang, 2004) and Randic (Randic and Vracko, 2000; Randic et al., 2003) present their 2D or 3D graphical representations. However, their approaches are associated with the computations of D/D, L/L and leading eigenvalue, which need a great deal of running time and memory space.

Moreover, the dinucleotide analysis has also been tried by several previous authors. Randic (2000) proposed a condensed representation of DNA based on pairs of nucleotides. This approach can offer fast, qualitative comparisons of DNA and allow quantitative comparisons of DNA from different sources. Wu et al. (2003) and Liu et al. (2006) proposed their analysis approaches based on neighboring nucleotides of DNA sequence, which reveal...
Every element of the matrix has a corresponding index to the above four categories. The matrix is used to construct graphical representation, the graphical curve and numerical invariants characterizing DNA sequences. The introduced representation is simple and direct, and gives us more biology information based on DNs.

2. 3D graphical representation of DNA sequences based on dual nucleotides

Given a DNA primary sequence, there are 16 kinds of the pairs of the neighboring nucleotides. These pairs can be classified as four categories based on their chemical properties: purine-DN {AG, GA}/pyrimidine-DN {CT, TC}, amino-DN {AC, CA}/keto-DN {GT, TG}, weak-H bond DN {AT, TA}/strong-H bond DN {CG, GC} and repeat-DN {AA, CC, GG, TT}. Then we design a 4 × 4 matrix and give a new 3D graphical representation of DNA sequences. We arrange 16 DNs in a 4 × 4 matrix according to the above four categories. The matrix is

\[
\begin{array}{cccc}
\text{AG} & \text{GA} & \text{CT} & \text{TC} \\
\text{AC} & \text{CA} & \text{GT} & \text{TG} \\
\text{AT} & \text{TA} & \text{GC} & \text{CG} \\
\text{AA} & \text{CC} & \text{GG} & \text{TT} \\
\end{array}
\]

Every element of the matrix has a corresponding index \((i, j)\), \(i = 0, 1, 2, 3; j = 0, 1, 2, 3\). Based on the index, we assign one DN as follows:

\[
\phi(g_{i, j+1}) = \begin{cases}
(0, 0, i) & \text{if } g_{i, j+1} = \text{AG}, \\
(0, 1, i) & \text{if } g_{i, j+1} = \text{GA}, \\
(0, 2, i) & \text{if } g_{i, j+1} = \text{CT}, \\
(0, 3, i) & \text{if } g_{i, j+1} = \text{TC}, \\
(1, 0, i) & \text{if } g_{i, j+1} = \text{AC}, \\
(1, 1, i) & \text{if } g_{i, j+1} = \text{CA}, \\
(1, 2, i) & \text{if } g_{i, j+1} = \text{GT}, \\
(1, 3, i) & \text{if } g_{i, j+1} = \text{TG}, \\
(2, 0, i) & \text{if } g_{i, j+1} = \text{AT}, \\
(2, 1, i) & \text{if } g_{i, j+1} = \text{TA}, \\
(2, 2, i) & \text{if } g_{i, j+1} = \text{CG}, \\
(2, 3, i) & \text{if } g_{i, j+1} = \text{GC}, \\
(3, 0, i) & \text{if } g_{i, j+1} = \text{AA}, \\
(3, 1, i) & \text{if } g_{i, j+1} = \text{CC}, \\
(3, 2, i) & \text{if } g_{i, j+1} = \text{GG}, \\
(3, 3, i) & \text{if } g_{i, j+1} = \text{TT}.
\end{cases}
\]

The map \(\phi\) maps \(G\) into a plot set. For example, the corresponding plot set of the sequence ATGGTGACC is \{(2, 0, 1), (3, 2, 3), (1, 2, 4), (1, 3, 5), (2, 3, 6), (1, 1, 7), (1, 0, 8), (3, 1, 9)\}. The corresponding plot set is called as characteristic plot set. The curve connected all plots of the characteristic plot set in turn is called 3D-DN curve. In Table 1 and Fig. 1, we show the corresponding coordinates and the 3D graphical representation of the sequence, respectively.

From the construction of the 4 × 4 matrix, we know that their designs are not unique. There are 16 kinds of DNs, so they have 16! combinations. But we design the 4 × 4 matrix based on the classifications of nucleotides. In this paper, we only consider the above 4 × 4 matrix to illustrate our scheme.

3. Numerical characterization of DNA sequences

Given a DNA sequence with the length of \(N\). Based on the definition of the map \(\phi\), we can have a set of points \((x_i, y_i, z_i)\), \(i = 1, 2, \ldots, N - 1\), and the correspondence

| Base | DNs | \(x\) | \(y\) | \(z\) |
|------|-----|-------|-------|------|
| 1    | AT  | 2     | 0     | 1    |
| 2    | TG  | 1     | 3     | 2    |
| 3    | GG  | 3     | 2     | 3    |
| 4    | GT  | 1     | 2     | 4    |
| 5    | TG  | 1     | 3     | 5    |
| 6    | GC  | 2     | 3     | 6    |
| 7    | CA  | 1     | 1     | 7    |
| 8    | AC  | 1     | 0     | 8    |
| 9    | CC  | 3     | 1     | 9    |

Table 1
Cartesian 3D coordinates for the sequence ATGGTGACC of the coding sequence of the first exon of human \(\beta\)-globin gene
Matrix as the following:

\[
\bar{d}_{ab} = \sum_{k=1}^{N} d_{ab}^k / K_{ab}, \quad p_{ab} = K_{ab} / (N - 1),
\]

where \( d_{ab}^k \) denotes the sum of geometrical lengths of edges between vertices (dots) \( V_1 \) and \( V_{ab} \) of the 3D-DN curve, where \( V_{ab} \) denotes the vertex representing the DN \( ab \) appearing in the given sequence. The parameter \( p_{ab} \) is defined as the distribution of DN \( ab \) frequency. For 3D-DN curve, after simple computation, we can obtain \( \bar{d}_{ab} = \sum_{k=1}^{N} d_{ab}^k / K_{ab}, \) \( p_{ab} = K_{ab} / (N - 1), \) where \( d_{ab}^k \) denotes the sum of geometrical lengths of edges between vertices (dots) \( V_1 \) and \( V_{ab} \) of the 3D-DN curve when the DN \( ab \) appears \( k \)th time in the given sequence. Here, we calculate the \((\bar{d}, p)\)-Matrix as the following:

\[
(\bar{d}, p) - M = \begin{pmatrix}
(\bar{d}^{AG}, p^{AG}) & (\bar{d}^{AT}, p^{AT}) & (\bar{d}^{TC}, p^{TC}) \\
(\bar{d}^{AC}, p^{AC}) & (\bar{d}^{CA}, p^{CA}) & (\bar{d}^{TG}, p^{TG}) \\
(\bar{d}^{AA}, p^{AA}) & (\bar{d}^{CC}, p^{CC}) & (\bar{d}^{CG}, p^{CG})
\end{pmatrix}
\]

The direct biological meaning of the \((\bar{d}, p)\)-Matrix is that they indicate the mean spaces and the distributions of DN's in the graph of the given sequence, respectively. Here, we regard them as the invariants to numerically characterize the DNA sequences.

In Nandy et al. (2006), Nandy suggests that authors apply their techniques to complete genes, or at least the complete coding sequence part. The complete genes of the beta globin genes have an interrupted structure with three exons and two introns. Comparisons of related genes in different species show that the sequences of the corresponding exons are usually conserved but the sequences of the introns are much less well related. In this paper, we apply our method to the complete coding sequence part (i.e. three exons). For simplification, in Table 2 we only list the primary DNA sequences of the complete coding sequences of part species.

In Table 3, the \((\bar{d}, p)\)-Matrix is constructed for the nine species presented in Table 2. To compare conveniently, we list the comparison of the mean space \( \bar{d} \) and the distribution \( p \) of DNs among the nine species in Fig. 2. Taking a closer look at Fig. 2, we can find some common features of nine DNA primary sequences, which are not easily visible in Table 3. The features may give us more information about their evolution. The DNs AG, GA, TC, AC, CA, GT, GC, AA, CC and TT occur appropriately in all species presented in Table 2. The DNs CT, TG and GG occur more frequently in all species. The DNs AT, TA and CG occur more rarely in all species. Moreover, observing the lines whose localities and heights denote the distributions of the corresponding DNs and mean space of every two identical DNs, respectively, we find Gallus (the only nonmammalian species) and Opossum (the most remote species from the remaining mammals) show larger entries among these species.

4. Similarities/dissimilarities among the complete coding sequences of \( \beta \)-globin gene of different species

In this section, we will illustrate the use of the quantitative characterization of DNA sequences by an examination of similarities/dissimilarities among the nine complete coding sequences in Table 2. The analysis of similarity/dissimilarity between two DNA sequences represented by the vectors is based on the assumption that the two sequences are similar if the corresponding vectors point to a similar direction and have similar magnitudes. Similar assumption is done in Randic et al. (2001).

In order to facilitate the quantitative comparison of different species, we extract some invariants with simple methods. Firstly, we calculate the space-sum matrix \((s-M)\) as follows:

\[
s-M = \begin{pmatrix}
d^{AG} & d^{GA} & d^{CT} & d^{TC} \\
d^{AC} & d^{CA} & d^{GT} & d^{TG} \\
d^{AT} & d^{TA} & d^{CG} & d^{GC} \\
d^{AA} & d^{CC} & d^{GG} & d^{TT}
\end{pmatrix}
\]

The element \( d_{ab} \) of the matrix \( s-M \) reveals the total sum of geometrical lengths of edges between vertices (dots) \( V_1 \) and all \( V_{ab} \) of the 3D-DN curve, where \( d_{ab} = \sum_{k=1}^{N} d_{ab}^k, \) \( a \in \{A, C, G, T\}, \) \( b \in \{A, C, G, T\}. \) Similarly, we have the
Table 2
The complete coding sequences of \(\beta\)-globin genes of nine species

| Species                  | Accession | Region: join | Exons | Exon 1 | Exon 2 | Exon 3 | ATGGTGCACTGTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGACCTGTGACA
|-------------------------|-----------|--------------|-------|--------|--------|--------|-----------------------------------------------|
| Human                   | U01317    | (62187 ... 62278; 62409 ... 62631; 63482 ... 63610) | 1     | 1-92   | 93-315 | 316-444 | ATGGTGCACTGTGACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGACCTGTGACA
| Goat                    | M15387    | (279 ... 364; 493 ... 715; 1621 ... 1749) | 1     | 1-86   | 87-309 | 310-438 | ATGCTGACTGCTGAGGAGGACTGGTGTGCTGTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| North American opossum  | J03643    | (467 ... 558; 672 ... 894; 2360 ... 2488) | 1     | 1-92   | 93-315 | 316-444 | ATGGTGCACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGACCTGTGACATGGGAACCTCCAAACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTGCAATGTCACCTGGCCACACTTCAAGATGGCATGGGAGGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| Gallus                  | V00409    | (465 ... 556; 649 ... 871; 1682 ... 1810) | 1     | 1-92   | 93-315 | 316-444 | ATGGTGCACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| Black lemur             | M15734    | (154 ... 245; 376 ... 598; 1365 ... 1419) | 1     | 1-92   | 93-315 | 316-444 | ATGGTGCACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| House mouse             | V00722    | (275 ... 367; 484 ... 705; 1334 ... 1462) | 1     | 1-92   | 93-315 | 316-444 | ATGCTGACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| Rabbit                  | V00882    | (277 ... 368; 493 ... 717; 1291 ... 1419) | 1     | 1-92   | 93-315 | 316-444 | ATGCTGACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| Norway rat              | X06701    | (310 ... 401; 517 ... 739; 1377 ... 1505) | 1     | 1-92   | 93-315 | 316-444 | ATGCTGACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA
| Cattle                  | X00376    | (278 ... 363; 492 ... 714; 1613 ... 1741) | 1     | 1-92   | 93-315 | 316-444 | ATGCTGACTGCTGAGGAGAAGGCTATGCTAGCTGGGCAAGGTGATTGTTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGGACAGGTTCTTTTGAGTCTTTGGGGATCTGTCCACTCCTGATGCTTATGGGCAACCCTAGGTGAAGGCTCATGGCAAGAAGGTGCTAGACTCCTTTAGTAACGGCATGAAGCATCTTGACGACCTCAAGGGCACCTTTGCCCAGCTGAGTGAGCTGCACTGTGATAAGCTGCACGTGGATCCTGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGGTTGTGCTGGCTCGCCACCATGGCAGTGAATTCACCCCGCTGCTGCAGGCTGAGTTTCAGAAGGTGGTGGCTGGTGTTGCAATGCCCTGGCCCACAAGATATCACTAA

X.-Q. Qi et al. / Journal of Theoretical Biology 249 (2007) 681–690
distribution matrix \((p-M)\) as follows:

\[
p-M = \begin{pmatrix}
p_{AG} & p_{GA} & p_{CT} & p_{TC} \\
p_{AC} & p_{CA} & p_{GT} & p_{TG} \\
p_{AT} & p_{TA} & p_{CG} & p_{GC} \\
p_{AA} & p_{CC} & p_{GG} & p_{TT}
\end{pmatrix}.
\]

The element \(p_{ab}\) in the matrix \(p-M\) indicates the distribution of DNs on the 3D-DN curve.

We will illustrate the use of the 3D quantitative characterization of DNA sequences with an examination of similarities/dissimilarities among the nine complete coding sequences listed in Table 2. We construct two 16-component vectors \((s\text{-vector}, p\text{-vector})\): \(s\text{-vector}\) consisting of the 16 space-sums in the matrix \(s\text{-M}\); \(p\text{-vector}\) consisting of the 16 distributions in the matrix \(p\text{-M}\). Based on the assumption of similarity/dissimilarity between two DNA sequences, the similarities among such vectors can be computed in two ways: (1) calculating the Euclidean distance between the end point of the \(s\text{-vectors}\); (2) calculating the Euclidean distance between the end point of the \(p\text{-vectors}\). When comparing two DNA sequences, we suppose that there are two species \(G_1\) and \(G_2\), the
parameters $i$ and $j$ denote row number and column number of $4 \times 4$ matrix, respectively. The distance $D(G_1, G_2)$ between the two $s$-vectors is

$$D(G_1, G_2) = \left[ \sum_{i=0}^{3} \sum_{j=0}^{3} (d_{ij}^{G_1} - d_{ij}^{G_2})^2 \right]^{1/2}.$$  

The smaller the Euclidean distance is, the more similar the DNA sequences are. We list the similarities and

The distance $P(G_1, G_2)$ between the two $p$-vectors is

$$P(G_1, G_2) = \left[ \sum_{i=0}^{3} \sum_{j=0}^{3} (p_{ij}^{G_1} - p_{ij}^{G_2})^2 \right]^{1/2}.$$  

Fig. 2. The comparison of the mean spaces and the distributions of DNs among the nine species in Table 2; $i$ of x-coordinate denotes the $i$th species in Table 2, $i = 1, 2, \ldots, 9$; the value of y-coordinate denotes the distributions of DNs; the value of z-coordinate denotes the mean spaces of DNs.
dissimilarities for the nine complete coding sequences in Tables 4 and 5.

Observing Tables 4 and 5, we find Gallus and Opossum are very dissimilar to others among the nine species because their corresponding rows have larger entries. On the other hand, the most similar species pairs are Human–Rabbit, Goat–Cattle and Black lemur–Rabbit. The more similar species pairs are Human–Goat, House mouse–Norway rat, Human–Black lemur, Goat–North American opossum and North American opossum–Cattle. This is not an accident, but indicates that they have close evolutionary relationship. For comparison, we denote the degree of similarity of the pair Human–Gallus as 1. Then we list the above results of the examination of the degree of similarity between human and other several species in Fig. 3. As one can see there is an overall agreement among

Fig. 2. (Continued)
The similarity/dissimilarity matrix for the complete coding sequences of Table 2 based on the Euclidean distances between the end points of the 16-component vectors of the space-sums of 16 DNs

| Species            | Human | Goat | North American opossum | Gallus | Black lemur | House mouse | Rabbit | Norway rat | Cattle |
|--------------------|-------|------|-------------------------|--------|-------------|-------------|--------|------------|--------|
| Human              | 0     | 0.0398 | 0.0480                   | 0.0713 | 0.0320      | 0.0311      | 0.0348 | 0.0358     | 0.0442 |
| Goat               | 0     | 0.0474 | 0.0834                   | 0.0857 | 0.0435      | 0.0522      | 0.0453 | 0.0450     | 0.0416 |
| North American opossum | 0     | 0.0599 | 0.0424                   | 0.0834 | 0.0522      | 0.0453      | 0.0453 | 0.0519     | 0.0900 |
| Gallus             | 0     | 0.0583 | 0.0715                   | 0.0583 | 0.0522      | 0.0265      | 0.0515 | 0.0525     | 0.0450 |
| Black lemur        | 0     | 0.0565 | 0.0345                   | 0.0563 | 0.0522      | 0.0547      | 0.0525 | 0.0547     | 0.0479 |
| House mouse        | 0     | 0.0525 | 0.0265                   | 0.0563 | 0.0522      | 0.0547      | 0.0525 | 0.0547     | 0.0479 |
| Rabbit             | 0     | 0.0525 | 0.0265                   | 0.0563 | 0.0522      | 0.0547      | 0.0525 | 0.0547     | 0.0479 |
| Norway rat         | 0     | 0.0525 | 0.0265                   | 0.0563 | 0.0522      | 0.0547      | 0.0525 | 0.0547     | 0.0479 |
| Cattle             | 0     | 0.0525 | 0.0265                   | 0.0563 | 0.0522      | 0.0547      | 0.0525 | 0.0547     | 0.0479 |

The degree of similarity of the complete coding sequences of several species with the complete coding sequence of human (a: from [this work, Table 4]; b: from [this work, Table 5]); i of x-coordinate denotes the species of Table 4 (x-coordinate 1: Goat, x-coordinate 2: North American opossum, x-coordinate 3: Gallus, x-coordinate 4: Black lemur, x-coordinate 5: House mouse, x-coordinate 6: Rabbit, x-coordinate 7: Norway rat, x-coordinate 8: Cattle).

5. Discussion

In this paper we arrange 16 DNs in a 4 × 4 matrix according to the four categories. The matrix $M$ is

\[
\begin{bmatrix}
AG & GA & CT & TC \\
AC & CA & GT & TG \\
AT & TA & CG & GC \\
AA & CC & GG & TT
\end{bmatrix}
\]

From the construction of the 4 × 4 matrix, we know that their designs are not unique. There are 16 kinds of DNs, so they have 16! combinations. Similarly, we find out the same phenomenon in Randic (2000) and Liu et al. (2006). Randic (2000) think that the ordering of the nucleic bases in his matrix is not important. Liu et al. (2006) only consider an ordering matrix to illustrate their method.

We suggest a novel approach based on DNs to compute parameters to determine similarity/dissimilarity between two DNA sequences. The ordering of the nucleic bases in our matrix is not important. But we want to know whether
Table 6
The similarity/dissimilarity matrix for the complete coding sequences of Table 2 based on the Euclidean distances between the end points of the 16-component vectors of the space-sums of 16 DNs (by using the s-vector derived from the new 3D-DN curve C' based on the new random matrix M')

| Species                      | Human | Goat | North American opossum | Gallus | Black lemur | House mouse | Rabbit | Norway rat | Cattle |
|------------------------------|-------|------|------------------------|--------|-------------|-------------|--------|------------|--------|
| Human                        | 0     | 10.379 | 12.545                | 19.172 | 9.210       | 10.814      | 7.897  | 9.602      | 12.403 |
| Goat                         | 0     | 8054 |                        | 20.541 | 11.125      | 13.402      | 10.123 | 12.887     | 6.669  |
| North American opossum       | 0     |      |                        | 19.740 | 13.302      | 14.171      | 12.391 | 9.086      | 8.976  |
| Gallus                       | 0     |      |                        | 0      | 21.693      | 11.952      | 20.071 | 16.045     | 23.525 |
| Black lemur                  | 0     |      |                        | 0      | 15.087      | 6.044       | 13.759 | 13.335     | 16.032 |
| House mouse                  | 0     |      |                        | 0      | 13.184      | 9.340       | 16.032 | 11.375     | 0      |
| Rabbit                       | 0     |      |                        | 0      | 10.148      | 11.375      |        |            | 0      |
| Norway rat                   | 0     |      |                        | 0      | 14.140      | 0           |        |            | 0      |

Based on the designation we can draw another 3D-DN curve to represent the same DNA sequence, which is named as C'.

For comparison, we list the similarities and dissimilarities for the nine complete coding sequences in Table 4 by using the s-vector derived from the new 3D-DN curve C' based on the new random matrix M'. As one can see there is an overall agreement among similarities obtained by different methods, despite some variation of numerical value among them. The variation of numerical value is not important. It is important whether there exit divergent trends in computing parameters to determine similarity/dissimilarity between two DNA sequences. We list the results of the examination of the degree of similarity between human and other several species in Fig. 4.

According to the results of the examination of Fig. 4 we can draw a conclusion that there exit the same trends in computing parameters to determine similarity/dissimilarity between two DNA sequences by using the s-vectors derived from different 3D-DN curves. The ordering of the nucleic bases in the suggested approach is not important.

6. Conclusion

In this paper, we give a novel approach to graphically characterize DNA primary sequences. The properties of DNs in a DNA sequence based on the 4 × 4 matrix consisting of 16 DNs are presented in the 3D graphical representation. Based on this representation, we construct two 16-component vectors and employ the vectors in characterizing and comparing the complete coding sequence part of beta globin gene of nine different species. The results of examination show that our method is useful for visualizing the local and global features of long or short DNA sequences and can reveal the visual characteristic in a DNA sequence. The advantage of our approach is that it allows visual inspection of DNs data, helping in recognizing major similarities among different DNA sequences.

References

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reuss, F., 1993a. Biochemistry 32, 6548.
Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romer, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993b. J. Biol. Chem. 268, 6119.

Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993c. J. Biol. Chem. 268, 14875.

Chou, K.C., 1989. J. Biol. Chem. 264, 12074.

Chou, K.C., 1990. Biophys. Chem. 35, 1.

Chou, K.C., 1993. J. Math. Chem. 12, 97.

Chou, K.C., Forsen, S., 1980. Biochem. J. 187, 829.

Chou, K.C., Liu, W.M., 1981. J. Theor. Biol. 91, 637.

Chou, K.C., Zhang, C.T., 1992. AIDS Res. Hum. Retrovirus 8, 1967.

Chou, K.C., Jiang, S.P., Liu, W.M., Fee, C.H., 1979. Sci. Sin. 22, 341.

Chou, K.C., Zhang, C.T., Elrod, D.W., 1996. J. Protein Chem. 15, 59.

King, E.L., Altman, C., 1956. J. Phys. Chem. 60, 1375.

Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. J. Chem. Inf. Comput. Sci. 40, 50.

King, E.L., Altman, C., 1956. J. Phys. Chem. 60, 1375.

Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. J. Chem. Inf. Comput. Sci. 40, 50.

Qi, Z.H., Qi, X.Q., 2007. Chem. Phys. Lett. 440, 139.

Randic, M., 2000. J. Chem. Inf. Comput. Sci. 40, 59.

Randic, M., Vracko, M., 2000. J. Chem. Inf. Comput. Sci. 40, 59.

Randic, M., Guo, X.F., Basak, S.C., 2001. J. Chem. Inf. Comput. Sci. 41, 619.

Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003. Chem. Phys. Lett. 371, 202.

Wang, M., Yao, J.S., Huang, Z.D., Xu, Z.J., Liu, G.P., Zhao, H.Y., Wang, X.Y., Yang, J., Zhu, Y.S., Chou, K.C., 2005. Med. Chem. 1, 39.

Wu, Y.H., Liew, A., Wee-C, Yan, H., Yang, M.S., 2003. Chem. Phys. Lett. 367, 170.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005a. Amino Acids 28, 29.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005b. J. Theor. Biol. 235, 555.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006a. Amino Acids 30, 49.

Xiao, X., Shao, S.H., Chou, K.C., 2006b. Biochem. Biophys. Res. Commun. 342, 605.

Zhang, Y.S., Chen, W., 2006. J. Theor. Biol. 242, 382.

Zhang, C.T., Chou, K.C., 1994. J. Mol. Biol. 238, 1.

Zhang, C.T., Zhang, R., 1991. Nucleic Acids Res. 19, 6113.

Zhou, G.P., Deng, M.H., 1984. Biochem. J. 222, 169.