# Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases

Bo Liao* and Tian-ming Wang

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China

Received November 20, 2003

We consider a 6-D representation of triplets of nucleotide bases of DNA sequences. Based on this representation, we outline an approach by constructing a 3-component vector whose components are the normalized leading eigenvalues of the L/L matrices associated with the triplets derived from DNA sequences. The examination of similarities/dissimilarities among the coding sequences of the first exon of $\beta$-globin gene of different species illustrates the utility of the approach.

## 1. INTRODUCTION

In recent years several authors outlined different graphical representations of DNA sequences based on 2-D, 3-D, or 4-D.[2,4,7−9,11−14] The advantage of graphical representation of DNA sequences is that they allow visual inspection of data, helping in recognizing major differences among similar DNA sequences, and allow one to construct a numerical characterization of the 2-D, 3-D, and even 4-D patterns of DNA. But both 2-D and 3-D graphical representations are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. Of course, each of the approaches has advantages and disadvantages that have to be considered when making comparative studies. Briefly, in some methods using codes to represent a DNA sequence there is no loss of information with the input data, but processing requires manipulation with codes that are in general computationally intensive. And, if a sequence is represented by a set of sequence invariants, then there exists some loss of information because different DNA sequences may have several identical invariants. Hence, one cannot reconstruct the sequence from a limited list of properties. However, if a sufficient number of sequence invariants are used, then one hopes to capture important features of DNA sequences. Recently, A. T. Balaban[1] et al. considered a characterization of DNA sequences based on nonoverlapping triplets of nucleotides bases. They associated each of the 64 possible triplets with single points with integer coordinates in a 3 × 3 × 3 cube. The $(x, y, z)$ coordinates of each triplet are restricted to $0 \leq x, y, z \leq 3$.

In this paper, we also consider a characterization of DNA sequences based on the fully overlapping triplets of nucleotide bases. We put forward a 6-D representation of the 64 possible triplets, which also avoids the limitations associated with crossing and overlapping. Our approach also allows a straightforward matrix representation in which the Euclidean distance between various nonoverlapping triplets of nucleotide bases gives the corresponding matrix elements. We make a comparison for the first exon of $\beta$-globin genes sequences belonging to eleven different species by constructing a 3-component vector consisting of the normalized leading eigenvalues of the L/L matrices of triplets associated with the DNA sequences. The similarities are computed by calculating the Euclidean distance between the end point of the vectors or calculating the correlation angle between two vectors.

## 2. 6-D REPRESENTATION OF TRIPLETS OF DNA SEQUENCES

Consideration of triplets of nucleotide bases instead of individual nucleotide bases has several reasons and advantages. There are two of them: (i) the genetic code consists of triplets (codons) of DNA (or RNA in some virus) nucleotides. (ii) The second advantage is that one can easily find the open reading frame as the longest sequence of triplets that contains no stop codons when read in a single reading frame.

We assign the 64 possible triplets with a single point with integer coordinates. The $(x, y, z, u, v, w)$ coordinates of each triplet are restricted to $x, y, z, u, v, w = -1, 0, 1$. Observing the 64 possible triplets, we find the following case: many pairs of codons that differ only in the third position base code for the same amino acid. On the other hand, a pair of codons differing only in the first or second position usually code for different amino acids. So we assign the first and the second nucleic base of triplet as follows

$$(0, 1) \rightarrow A; (1, 0) \rightarrow G; (0, -1) \rightarrow T; (-1, 0) \rightarrow C$$

and assign 0 or −1 or 1 to the third nucleic base. We consider all possible triplets of an arbitrary DNA primary, and the following map will reduce a DNA sequence into a plot set. In detail, let $G = g_1g_2\cdots$ be an arbitrary DNA primary sequence. Then we have a map $\phi$, which maps $G$ into a plot set. $\phi(G) = \phi(g_1g_2g_3)\phi(g_2g_3g_4)...\phi(g_ig_{i+1}g_{i+2})...$

* Corresponding author fax: (86)411-4706100; e-mail: dragonbw@163.com.

10.1021/ci034271f CCC: $27.50 © 2004 American Chemical Society
Published on Web 07/27/2004

Explicitly, $\phi_1(G) = \phi_1(g_1g_2g_3)\phi_1(g_2g_3g_4)\cdots$, where

$$\phi_1(g_ig_{i+1}g_{i+2}) = \begin{cases} (0,-1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTT, TTC\} \\ (0,-1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTA, TTG\} \\ (0,-1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TCT, TCC, TCA, TCG\} \\ (0,-1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAT, TAC\} \\ (0,-1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAA, TAG\} \\ (0,-1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGT, TGC\} \\ (0,-1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGA, TGG\} \\ (-1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CTT, CTC, CTA, CTG\} \\ (-1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CCT, CCC, CCA, CCG\} \\ (-1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAT, CAC\} \\ (-1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAA, CAG\} \\ (-1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CGT, CGC, CGA, CGG\} \\ (0,1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATT, ATC\} \\ (0,1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATA, ATG\} \\ (0,1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ACT, ACC, ACA, ACG\} \\ (0,1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAT, AAC\} \\ (0,1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAA, AAG\} \\ (0,1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGT, AGC\} \\ (0,1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGA, AGG\} \\ (1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GTT, GTC, GTA, GTG\} \\ (1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GCT, GCC, GCA, GCG\} \\ (1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GGT, GGC, GGA, GGG\} \\ (1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAT, GAC\} \\ (1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAA, GAG\} \end{cases}$$

In Table 2 we show (*x*, *y*, *z*, *u*, *v*, *w*,) coordinates of the 8 nonoverlapping triplets of the first 10 nucleotide bases making up the coding sequence of the first exon of human and goat $\beta$-globin gene.

## 3. CHARACTERIZATION OF TRIPLETS OF DNA SEQUENCES WITH 3-COMPONENT VECTORS

To find some of the invariants sensitive to the form of the 6D representation, we will transform the 6D representation of the triplets into another mathematical object, a matrix. Once we have a matrix representing a DNA sequence, we can use some of the matrix invariants as descriptors of the sequence. One of the matrices is the L/L matrix whose elements $l_{i,j}$ are defined as the quotient of the Euclidean distance between a pair of vertices (dots) representing triplets and the sum of the distances between the same pair of vertices measured between the triplets. In other words, $l_{i,j} = (d_{i,j}/\sum_{k=i}^{j-1} d_{k,k+1})$, where $d_{i,j}$ is the Euclidean distance between a pair of vertices. Its eigenvalues, and in particular its leading eigenvalue, can be used as descriptors of a DNA sequence.

The labels A, T, G, and C can be arranged in 4! ways. But, in fact, bases of DNA can be classed into groups, purine(A,G)/pyrimidine(C,T), amino(A,C)/keto(G,T), and weak-H bond(A,T)/strong-H band(G,C). We can obtain only three representations corresponding to the three classifications. By computing we find that the patterns ATGC, ATCG, CGAT, GCAT, CGTA, and GCTA have the same L/L matrices, which correspond to map $\phi_1$. Equally, patterns ACTG, ACGT, TGAC, GTAC, CATG, and CAGT correspond to

**Table 1.** Coding Sequences of the First Exon of $\beta$-Globin Gene of Eleven Different Species

| species | coding sequence |
|---|---|
| human | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGC<br>AAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG |
| goat | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGT<br>GAAAGT GGATGAAGTTGGTGCTGAGGCCCTGGGCAG |
| opossum | ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTA<br>AGGT GCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG |
| gallus | ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGC<br>AAGGT CAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG |
| lemur | ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCA<br>AGGT GGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG |
| mouse | ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCA<br>AAGG TGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| rabbit | ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGC<br>AAGGT GAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC |
| rat | ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGA<br>AAGGT GAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG |
| gorilla | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCA<br>AGGT GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| bovine | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGT<br>GAAA GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| chimpanzee | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT<br>GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG |

**Table 2.** 6-D Coordinates for the First 10 Bases of Human and Goat

| | human | | | | | | | goat | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no. | base | x | y | z | u | v | w | base | x | y | z | u | v | w |
| 1 | ATG | 0 | 1 | 0 | −1 | 1 | 1 | ATG | 0 | 1 | 0 | −1 | 1 | 1 |
| 2 | TGG | 0 | −1 | 1 | 0 | 1 | 2 | TGC | 0 | −1 | 1 | 0 | −1 | 2 |
| 3 | GGT | 1 | 0 | 1 | 0 | 0 | 3 | GCT | 1 | 0 | −1 | 0 | 0 | 3 |
| 4 | GTG | 1 | 0 | 0 | −1 | 0 | 4 | CTG | −1 | 0 | 0 | −1 | 0 | 4 |
| 5 | TGC | 0 | −1 | 1 | 0 | −1 | 5 | TGA | 0 | −1 | 1 | 0 | 1 | 5 |
| 6 | GCA | 1 | 0 | −1 | 0 | 0 | 6 | GAC | 1 | 0 | 0 | 1 | −1 | 6 |
| 7 | CAC | −1 | 0 | 0 | 1 | −1 | 7 | ACT | 0 | 1 | −1 | 0 | 0 | 7 |
| 8 | ACC | 0 | 1 | −1 | 0 | 0 | 8 | CTG | −1 | 0 | 0 | −1 | 0 | 8 |

**Table 3.** Leading Eigenvalues of the L/L Matrices Associated with 3 Essentially Different Patterns for the Corresponding Triplets of Table 1

| patterns | human | goat | gallus | opossum | lemur | mouse | rabbit | rat | bovine | gorilla | chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ATGC | 42.7371 | 40.4410 | 41.3072 | 41.7015 | 41.1507 | 43.4264 | 41.8727 | 41.8568 | 41.3827 | 43.2450 | 49.0570 |
| ACGT | 40.1804 | 38.9056 | 40.6868 | 38.4661 | 38.9689 | 40.4513 | 38.5437 | 40.2589 | 39.4228 | 40.2695 | 45.6039 |
| AGTC | 41.2304 | 38.4891 | 40.4827 | 38.7402 | 38.6311 | 41.0299 | 39.3604 | 40.6079 | 39.7917 | 41.3176 | 47.0931 |

the following map $\phi_2$; patterns AGTC, AGCT, GATC, GACT, TCGA, and CTGA correspond to the following map $\phi_3$. Hence, there are three essentially different patterns of the matrix representations representing the same DNA sequence.

$$\boxed{\phi_2(g_ig_{i+1}g_{i+2}) = \begin{cases} (0,-1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTT, TTC\} \\ (0,-1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTA, TTG\} \\ (0,-1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TCT, TCC, TCA, TCG\} \\ (0,-1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAT, TAC\} \\ (0,-1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAA, TAG\} \\ (0,-1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGT, TGC\} \\ (0,-1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGA, TGG\} \\ (-1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CTT, CTC, CTA, CTG\} \\ (-1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CCT, CCC, CCA, CCG\} \\ (-1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAT, CAC\} \\ (-1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAA, CAG\} \\ (-1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CGT, CGC, CGA, CGG\} \\ (0,1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATT, ATC\} \\ (0,1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATA, ATG\} \\ (0,1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ACT, ACC, ACA, ACG\} \\ (0,1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAT, AAC\} \\ (0,1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAA, AAG\} \\ (0,1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGT, AGC\} \\ (0,1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGA, AGG\} \\ (1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GTT, GTC, GTA, GTG\} \\ (1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GCT, GCC, GCA, GCG\} \\ (1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GGT, GGC, GGA, GGG\} \\ (1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAT, GAC\} \\ (1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAA, GAG\} \end{cases}}$$

$$\boxed{\phi_3(g_ig_{i+1}g_{i+2}) = \begin{cases} (0,-1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTT, TTC\} \\ (0,-1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TTA, TTG\} \\ (0,-1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TCT, TCC, TCA, TCG\} \\ (0,-1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAT, TAC\} \\ (0,-1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TAA, TAG\} \\ (0,-1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGT, TGC\} \\ (0,-1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{TGA, TGG\} \\ (-1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CTT, CTC, CTA, CTG\} \\ (-1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CCT, CCC, CCA, CCG\} \\ (-1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAT, CAC\} \\ (-1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CAA, CAG\} \\ (-1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{CGT, CGC, CGA, CGG\} \\ (0,1,0,-1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATT, ATC\} \\ (0,1,0,-1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ATA, ATG\} \\ (0,1,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{ACT, ACC, ACA, ACG\} \\ (0,1,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAT, AAC\} \\ (0,1,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AAA, AAG\} \\ (0,1,1,0,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGT, AGC\} \\ (0,1,1,0,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{AGA, AGG\} \\ (1,0,0,-1,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GTT, GTC, GTA, GTG\} \\ (1,0,-1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GCT, GCC, GCA, GCG\} \\ (1,0,1,0,0,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GGT, GGC, GGA, GGG\} \\ (1,0,0,1,-1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAT, GAC\} \\ (1,0,0,1,1,i) & \text{if } g_ig_{i+1}g_{i+2} \in \{GAA, GAG\} \end{cases}}$$

We will characterize the coding sequences of the first exon of $\beta$-globin gene of 11 species, shown in Table 1, by means of the leading eigenvalue of the L/L matrix. In Table 3 we give the leading eigenvalues of the L/L matrices associated with 3 essentially different patterns of the triplets representing each of the coding sequences. Observing Table 3, we can find that the largest leading eigenvalue occurs for pattern ATGC and the smallest value occurs for ACGT, except for goat, gallus, and lemur.

## 4. SIMILARITIES/DISSIMILARITIES AMONG THE CODING SEQUENCES OF THE FIRST EXON OF $\beta$-GLOBIN GENE OF ELEVEN SPECIES

We will illustrate the use of the 6-D quantitative characterization of DNA sequences with the examination of similarities/dissimilarities among the 11 coding sequences of Table 1. We construct a 3-component vector consisting of the normalized leading eigenvalue $\lambda/(N - 2)$, where $\lambda$ is the leading eigenvalue of matrix L/L of triplets of DNA sequences and $N$ is the number of bases making up the corresponding DNA sequence. The underlying assumption is that if two vectors point to a similar direction in the 3-dimensional space, then the two DNA sequences represented by the 3-component vectors are similar.

The similarities among such vectors can be computed in three ways: (1) we calculate the Euclidean distance between the end point of the vectors; (2) we calculate the correlation angle of two vectors, and (3) we calculate the cosine of the correlation angle of two vectors. Obviously, when one calculated the correlation angle of two vectors, the cosine of the correlation angle of two vectors is easily obtained. The more small Euclidean distance between the end points of two vectors, the more similar the DNA sequence. And, the more small correlation angle between two vectors, the more similar the DNA sequence. On the other hand, the more larger the cosine of the correlation angle between two vectors, the more similar the DNA sequence.

In Table 4, we give the similarities and dissimilarities for the coding sequences of Table 1 based on the Euclidean distances between the end points of the 3-component vectors of the normalized leading eigenvalues of the L/L matrices of the triplets. We believe that it is not accidental that the smallest entries in Table 4 are associated with the pairs (gorilla, chimpanzee), (human, chimpanzee), (mouse, rabbit), and (human, gorilla). On the other hand, the larger entries in the similarity/dissimilarity matrix appear in the rows belonging to opossum and gallus.

In Table 5, the similarities and dissimilarities for 11 coding sequences are based on the correlation angle between two vectors. Observing Table 5, we find gallus is very dissimilar to others among the 11 species because its corresponding row has lager entries. On the other hand, the more similar species pairs are gorilla-chimpanzee, human-chimpanzee, human-gorilla, (mouse rabbit), and (mouse, gorilla). Similar results have been obtained by M. Randic.[3,10,11]

Comparing Tables 4 and 5, we can find that there exists an overall qualitative agreement among similarities although there is small difference.

The Euclidean distance measure between the end points of vectors and the correlation angle between vectors are different measures of the similarity of RNA secondary structures. However, there are small differences between Tables 4 and 5. The reason for making these differences may be as follows: (1) There is some loss of information associated with the distance matrix. (2) Information extracted in each structure is not plenteous enough to compare with eleven species. In general, the correlation angle or the cosine of the correlation angle is the best tolerance for the similarities.

M. Randic uses a 12-component vector whose components are made up of the normalized leading value,[3] the 16-

**Table 4.** Similarity/Dissimilarity Matrix for the Coding Sequences of Table 1 based on the Euclidean Distances between the End Points of the 3-Component Vectors of the Normalized Leading Eigenvalues of the L/L Matrices of the Triplets of DNA Sequences

| species | human | goat | gallus | opossum | lemur | mouse | rabbit | rat | bovine | gorilla | chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| human | 0 | 0.017963 | 0.018791 | 0.035508 | 0.036415 | 0.014180 | 0.013779 | 0.012011 | 0.032909 | 0.005679 | 0.004059 |
| goat | | 0 | 0.026427 | 0.048750 | 0.048330 | 0.028092 | 0.028004 | 0.023829 | 0.020103 | 0.021966 | 0.021071 |
| gallus | | | 0 | 0.031668 | 0.028118 | 0.018401 | 0.022109 | 0.007863 | 0.044758 | 0.019326 | 0.021010 |
| opossum | | | | 0 | 0.008375 | 0.021610 | 0.023477 | 0.028817 | 0.066990 | 0.030427 | 0.033459 |
| lemur | | | | | 0 | 0.023327 | 0.026390 | 0.027376 | 0.067476 | 0.032088 | 0.035232 |
| mouse | | | | | | 0 | 0.004358 | 0.011569 | 0.045527 | 0.009120 | 0.012402 |
| rabbit | | | | | | | 0 | 0.014761 | 0.044306 | 0.008153 | 0.011028 |
| rat | | | | | | | | 0 | 0.041848 | 0.011581 | 0.013513 |
| bovine | | | | | | | | | 0 | 0.037544 | 0.035295 |
| gorilla | | | | | | | | | | 0 | 0.003357 |
| chimpanzee | | | | | | | | | | | 0 |

**Table 5.** Similarity/Dissimilarity Matrix for the Coding Sequences of Table 1 based on the Angle between the End Points of the 3-Component Vectors for the Normalized Leading Eigenvalues of the L/L Matrices of the Triplets of DNA Sequences

| species | human | goat | gallus | opossum | lemur | mouse | rabbit | rat | bovine | gorilla | chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| human | 0 | 0.014850 | 0.019304 | 0.015629 | 0.014662 | 0.008640 | 0.011507 | 0.009625 | 0.007004 | 0.004632 | 0.004615 |
| goat | | 0 | 0.012788 | 0.017442 | 0.007114 | 0.013776 | 0.018551 | 0.009025 | 0.008076 | 0.016141 | 0.018237 |
| gallus | | | 0 | 0.028903 | 0.019698 | 0.023367 | 0.028117 | 0.009770 | 0.013752 | 0.023011 | 0.023890 |
| opossum | | | | 0 | 0.010862 | 0.007036 | 0.006008 | 0.020443 | 0.016044 | 0.011742 | 0.014042 |
| lemur | | | | | 0 | 0.009466 | 0.013489 | 0.013519 | 0.010183 | 0.013773 | 0.016373 |
| mouse | | | | | | 0 | 0.004871 | 0.014168 | 0.009774 | 0.005314 | 0.007947 |
| rabbit | | | | | | | 0 | 0.018712 | 0.014408 | 0.006975 | 0.008716 |
| rat | | | | | | | | 0 | 0.008423 | 0.013251 | 0.014163 |
| bovine | | | | | | | | | 0 | 0.009547 | 0.011022 |
| gorilla | | | | | | | | | | 0 | 0.002681 |
| chimpanzee | | | | | | | | | | | 0 |

**Table 6.** Similarity/Dissimilarity of the Coding Sequences of the First Exon of the Human $\beta$-Gene based on the Euclidean Distances between the End Points of (A) the 3-Component Vectors, (B) the 12-Component Vectors, (c) the 16-Component Vectors, (D) the 64-Component Vectors, (E) the 12-Component Vectors, (F) the 5-Component Vectors, and (G) the 15-Component Vectors Representing the Coding Sequences

| species | A[a] | B[b] | C[c] | D[d] | E[e] | F[f] | G[g] |
|---|---|---|---|---|---|---|---|
| goat | 0.014850 | 0.061 | 6.928 | 8.944 | 4.996 | 0.2066 | 0.007723 |
| gallus | 0.019304 | 0.109 | 9.592 | 10.630 | 5.015 | 0.0494 | 0.009273 |
| opossum | 0.015629 | 0.148 | 8.602 | 11.402 | 4.491 | 0.0402 | 0.004880 |
| lemur | 0.014662 | 0.087 | 7.483 | 10.100 | 2.970 | 0.0536 | 0.009288 |
| rabbit | 0.011507 | 0.042 | 6.083 | 6.708 | 3.171 | 0.0329 | 0.004669 |
| rat | 0.009625 | 0.043 | 6.325 | 8.246 | 4.857 | 0.0303 | 0.004277 |
| gorilla | 0.004632 | 0.021 | | | | | 0.003893 |
| chimpanzee | 0.004615 | 0.017 | | | | | 0.004679 |

[a] This work [Table 5]. [b] From ref 2,Table 2. [c] From ref 10, Table 6. [d] From ref 11, Table 9. [e] From ref 11, Table 12. [f] From ref 6, Table 9. [g] From ref 5, Table 8.

component vector whose components are made up of the frequency of occurrence of all possible ordered pairs of adjacent bases,[10] the 5-component vector whose components are made up of the average bandwidths,[6] and the 64 components consisting of the frequency of occurrence of all ordered triplets of bases (segments of length 3).[11] In a previous paper,[4] we constructed 15-component vectors consisting of the average bandwidths. While in this paper we use a 3-component vector whose components are made up of the normalized leading eigenvalue of L/L matrices of the triplets. But there exists an overall qualitative agreement among similarities based on different descriptors despite there are some variations among them. Our approach also can be applied in the comparison among RNA sequences and protein sequences. Our approach may be more convenient since the alternative descriptor is a 3-component vector, while the other approaches used higher dimensional vectors. The recently reported results of the examination of the degree of similarity/dissimilarity of the coding sequences of the first exon of $\beta$-globin gene of several species with the coding sequence of the first exon of the human $\beta$-globin gene by means of approaches using alternative DNA sequence descriptors are listed in Table 6 for comparison.

## 5. CONCLUSION

We proposed a 6-D representation of triplets of DNA sequences and presented a similarity measure among DNA sequences. In our approach, the characteristic plot sets of DNA sequences and the similarity can be computed easily, and also our approach allows visual inspection of data, helping in recognizing major similarities among different DNA sequences, and allows one to construct numerical characterization. In our approach, the insertions, deletions, and substitutions of plots representing triplets associated with DNA sequences in 6D space correspond to the insertions, deletions, and substitutions of letters in the compared sequences, respectively. One difference from the alignments of DNA sequences is that our approach shall consider not only sequences' structure but also chemical structure for DNA sequences. The mathematic invariants, normalized leading eigenvalues, are applied to compare DNA sequences, rather than strings' sequence themselves. For the full sequence, the coordinates of all bases (nucleotides) of any DNA sequences and the L/L matrices are easily computed, so our approach also can be applied to compute complete gene sequences.

## ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for many valuable suggestions that have improved this manuscript.

## REFERENCES AND NOTES

(1) Balaban, A. T.; Plavsic, D.; Randic, M. DNA invariants based on nonoverlapping triplets of nucleotide bases. *Chem. Phys. Lett.* **2003**, *379*, 147−154.

(2) Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Novel 2-D graphical representation of DNA sequences and their numberical characterization. *Chem. Phys. Lett.* **2003**, *368*, 1−6.

(3) Randic, M.; Vracko, M.; Lers, N. Dejanplavsic, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **2003**, *371*, 202−207.

(4) Yuan, C.; Liao, B.; Wang, T. New 3-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, *379*, 412−417.

(5) Liao, B.; Wang, T. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett.* **2004**, *388*, 195−200.

(6) Randic, M.; Balanba, A. T. On A Four-Dimensional Representation of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci* **2000**, *40*, 50−56.

(7) Nandy, A. A new graphical representation and analysis of DNA sequence structure: I.Methodology and Application to Globin Genes. *Curr. Sci.* **1994**, *66*, 309−314.

(8) Nandy, A.; Nandy, P. Graphical analysis of DNA sequences structure: II. Relative abundance of nucleotides in DNAs, gene evolution and duplication. *Curr. Sci.* **1995**, *68*, 75−85.

(9) Nandy, A. Graphical analysis of DNA sequence structure: III. Indication of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci.* **1996**, *70*, 661−668.

(10) Randic, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.

(11) Randic, M.; Guo, X. F.; Basak, S. C. On the Characterization of DNA Primary Sequence by Triplet of Nucleic Acid Bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619−626.

(12) Zhang, R.; Zhang, C. T. Z-curve, An Intuitive Tool for Visualizing and Analyzing the DNA sequences. *J. Biomol. Str. Dyn.* **1994**, *11* (4), 767−782.

(13) Lan, M. L.; Carpendale, M. S. T. Supporting Detail-in-Context for the DNA Representation. *H.-Curves* **1998**.

(14) Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* **1983**, *25*8, 1318−1327.

CI034271F