MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **68** (2012) 611-620

ISSN 0340 - 6253

A Novel Graphical and Numerical Representation for Analyzing DNA Sequences Based on Codons

Nafiseh Jafarzadeh and Ali Iranmanesh*

Department of Mathematics, Faculty of Mathematical Sciences,
Tarbiat Modares University, P.O. Box: 14115-137,
Tehran, Iran

iranmanesh@modares.ac.ir

(Received November 7, 2011)

Abstract

One important task in the study of genome sequences and mutations is to determine densities of specific nucleotides and codons. The graphical representation of DNA sequences provide a simple way of viewing, storing, and comparing various sequences. In this paper, we first present for each kind of codon, a numerically representation as a 2D coordinate (x,y) and give a 2D graphical representation for DNA sequences. Then we transform the graphical representation into a Matrix to facilitate quantitative comparisons of DNA sequences and compute "ALE-index" for this matrix. It is an invariant of DNA sequences and using this index, we construct similarity/dissimilarity table based on this invariant for sequences of DNA of the first exon of β-globin gene from nine species for illustrating the utility of this representation.

1. Introduction

DNA (Deoxyribonucleic acid) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide". Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations as shorthand for the four bases: A is for adenine, G is for guanine, C is for cytosine, T is for thymine. They are often called bases. A and T are complement, also G and C.

* Corresponding author (Ali Iranmanesh)

In the recent years, a rapid growth of sequence data in DNA databases has been observed. Some graphical representations of DNA sequences have been given by Nandy [1], and Guo et al. based on 2D graphical representation of DNA sequences. Guo and Nandy introduced a novel 2D graphical representation of DNA sequences of low degeneracy [2].

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 sequences [3]. Since then, quite a few models based on different mechanism 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 [4].

In general, many advances in 2D, 3D, and 4D, DNA sequences representation appeared after the initial works [5-15]. Up to now, many papers published in DNA sequence. For example see [16-26].

Codon is a specific sequence of three adjacent nucleotides on the mRNA that specifies the genetic code information for synthesizing a particular amino acid. See the codon table below in Figure 1.

1st	T	C	A	G	3rd
T	TTT TTC TTA TTG	TCT TCC TCA TCG	TAT TAC TAA TAG	TGT TGC TGA TGG	T C A G
C	CTT CTC CTA CTG	CCT CCC CCA CCG	CAT CAC CAA CAG	CGT CGC CGA CGG	
A	ATT ATC ATA ATG	ACT ACC ACA ACG	AAT AAC AAA AAG	AGT AGC AGA AGG	
G	GTT GTC GTA GTG	GCT GCC GCA GCG	GAT GAC GAA GAG	GGT GGC GGA GGG	

Figure 1. Codon table

We want to present for each kind of codon, a numerically representation as a 2D coordinate (x,y). This is done in two steps:

Step 1: distribute each kind of 4 nucleotides in cartesian 2D coordinates as shown in Figure 2 and also for 4^2=16 dinucleotides as shown in Figure 3 [9].

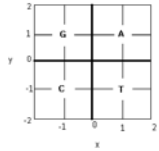

Figure 2. Distribution of the 4 kinds of bases in cartesian 2D coordinates

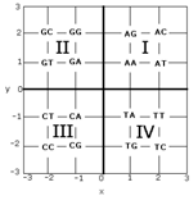

Figure 3. 16 kinds of di nucleotides distributed in cartesian 2D coordinates

Step 2: with development of previous step, we distribute each kind of $4^3 = 64$ codons in cartesian 2D coordinates.

For this goal, at first we add the base "A" at the first of dinucleotides of each region in Figure 3 and then we put all of the new table to the region (I) of Figure 4. In continue, we add the base "G" at the first of dinucleotides of each region in Figure 3. Then we put all of the new table to the region (II) of Figure 4, then we add the base "C" at the first of dinucleotides of each region in Figure 3 and then we put all of the new table to the region (III) of Figure 4 and in the last, we add the base "T" at the first of dinucleotides of each region in Figure 3 and then we put all of the new table to the region (IV) of Figure 4. The 64 kinds of codons can be divided in to the four quadrants of a Cartesian 2D coordinates, as shown in Figure 4.

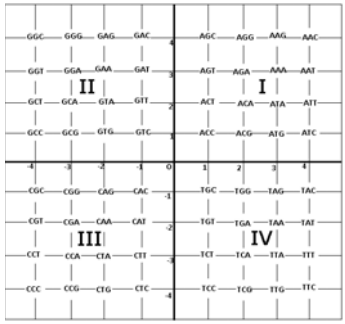

Figure 4. 16 kinds of codons distributed in cartesian 2D coordinates

In this paper, we give a new 2D-directed graphical representation of DNA sequences based on codons, after this we transform the graphical representation into a Matrix to facilitate quantitative comparisons of DNA sequences.

2. 2D-directed graphical representation of DNA

In this section, we give the detail of presentation of 2D directed graphical representation of DNA based on codons. For this purpose, let $S = C_1\ C_2\ .....C_N$ be the mRNA sequence which transcribed from a DNA sequence with N codons. With using 2D coordinate of each codon, we define N new points in Cartesian 2D coordinates, $(P_1, P_2, \ldots, P_N)$.

For each, $n \in \{1,2,\ldots,N\}$, we have $P_n = (\sum_{i=1}^{n} x_i , \sum_{i=1}^{n} y_i)$ which (x_i,y_i) is a coordinate of codon C_i.

Now we can plot a directed representation of DNA sequences.

Take S= ATGGTGCACCCC for example, here we have 4 codons : C_1=ATG, C_2= GTG, C_3=CAC C_4=CCC, and we obtain 2D coordinate of each codons: $(x_1,y_1)=(3,1)$,$(x_2,y_2)=(-2,1)$,$(x_3,y_3)=(-1,-1)$,$(x_4,y_4)= (-4,-4)$ and hence we have: $P_1=(3,1)$, $P_2=(1,2)$, $P_3=(0,1)$, $P_4=(-4,-3)$.

We showed its 2D directed graph in Figure 5.

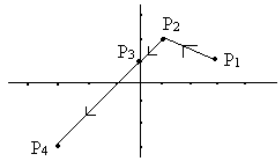

Figure 5. The 2D directed graph based on codon of the sequence ATGGGTGCACCCC

3. Numerical characterization

In this section, we construct a distance/distance matrix (D/D) according to 2D-directed graphical representation of DNA sequence. [27] The D/D matrix is defined as follows:

$[D/D]_{ij} = [ED]_{ij} / [GD]_{ij}$, $i \neq j$,

$[D/D]_{ij} = 0$, $i=j$

where ED is Euclidean-distance matrix which: the (i,j) – matrix element is defined to be the Euclidean – distance between vertices v_i and v_j of the curve in the 2D space., i.e.,

$$[ED]_{ij} = [(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2]^{1/2}$$ which $p_i = (x_{i1} , x_{i2})$, $p_j = (x_{j1} , x_{j2})$,

and GD is graph theoretical (topological) distance matrix which: the (i,j)-matrix element is defined to be the graph theoretical distance between vertices v_i and v_j ., i.e.,

$$[GD]_{ij} = \begin{cases} j - i & \text{if } j > i \\ \infty & \text{otherwise} \end{cases}$$

The D/D matrix associated with a directed graphical representation is an upper triangular matrix. We want to compute one invariant or index for this matrix. One of the important invariants for DNA sequences is leading eigenvalue (λ) and another invariant is ALE-index (χ). ALE-index is defined by Li and Wang [28], which is an invariant of DNA sequences:

Let $M = (a_{ij})_{n\times n}$ be such a matrix, with the following property:

$a_{ij} \geq 0$, $a_{ij} = a_{ji}$, and $a_{ii} = 0$ for i,j = 1,2,...,n.

The ALE-index of M is defined as follows:

$\chi = \chi(M) = \frac{1}{2}\left(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F\right)$, where $\|M\|_{m1} = \sum_{i,j} |a_{ij}|$ and $\|M\|_F = (tr(M^T M))^{1/2}$.

Leading eigenvalue and ALE-index, almost are used for undirected graphical representation.

In follows, we investigate whether or not these invariants are compatible for this 2D-directed representation based on codon.

In the following, by using [29], we will give some properties of this 2D-directed representation and prove them.

Property 1. ALE-index is compatible as a sequence invariant for the D/D matrix of 2D-directed graphical representation based on codon of a given mRNA sequence with N codons.

Proof. Let "S" be a given mRNA sequence and "M" is D/D matrix of 2D-directed graphical representation based on codon of "S". M is matrix of directed graph and $[GD]_{ij} = \infty$ when i>j , then M is an upper triangular matrix. ALE-index is defined for undirected graphical representation which is symmetric.

Now let $M = (a_{ij})_{n\times n}$, where $a_{ij} = a_{ji} \geq 0$ and $a_{ii} = 0$ for i,j = 1,2,….,n . By $\acute{M} = (b_{ij})_{n\times n}$ we denote the corresponding upper triangular matrix of M, where

$b_{ij} = a_{ij}$ if $j \geq i$,

$b_{ij} = 0$ if $j < i$.

Then we have:

$\|M\|_{m1} = 2\left\|\acute{M}\right\|_{m1}$,

$\|M\|_F = \sqrt{2}\,\|\acute{M}\|_F$.

We denote vector $\alpha = (\frac{1}{2n}\|M\|_{m1}, \sqrt{\frac{n-1}{4n}}\|M\|_F)$. Then we have the following equation.

$$\chi(M) = \frac{1}{2}\left(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F\right) = (1,1)\cdot\alpha$$

$$\chi(\acute{M}) = \frac{1}{2}\left(\frac{1}{n}\|\acute{M}\|_{m1} + \sqrt{\frac{n-1}{n}}\|\acute{M}\|_F\right) = (\frac{1}{2},\sqrt{2})\cdot\alpha$$

where "." is the inner product between two vectors. This implies that the formal "ALE-index" of $\acute{M}$ and the "ALE-index" of M can determine each other uniquely. Therefore, one can use the formal "ALE-index" of an upper triangular matrix. Then ALE-index is compatible as a sequence invariant for D/D matrix of that 2D-directed graph. ■

Property 2. The leading eigenvalue is not compatible as a sequence invariant for the D/D matrix of 2D-directed graphical representation based on codon of a given mRNA sequence with N codons.

Proof. According to the information above, $\acute{M}$ is upper triangular matrix, then $\lambda(\acute{M}) \equiv 0$, whereas $\lambda(M)$ is usually not zero, therefor, $\lambda(M)$ cannot be reflected by $\lambda(\acute{M})$ directly. This implies that the leading eigenvalue cannot use as an invariant to describe 2D-directed graphical representation of DNA. ■

By the above proposition, we can compute ALE-index for analyzing similarities of DNA sequences based on codons. For example for directed graph in Figure 5, we compute ALE-index for "D/D matrix" of this graph.

Let S = ATGGTGCACCCC, then: $D/D = \begin{pmatrix} 0 & 2.236 & 1.500 & 2.687 \\ 0 & 0 & 1.414 & 3.536 \\ 0 & 0 & 0 & 2.828 \\ 0 & 0 & 0 & 0 \end{pmatrix}$.

Then,

$\|D/D\|_{m1} = 14.201$

$\|D/D\|_F = 6.080$ and therefore,

$\chi(D/D) = 1/2(1/4(14.201 + \sqrt{3}/4 \times 6.080)) = 2.104$

4. Discussion

In this section, to illustrate the utility of this novel graphical and numerical representation of DNA sequences based on codons, we will consider similarities and dissimilarities among the nine exons of Table 1. Following the method mentioned in Sections 2 and 3, we can get the corresponding directed graph and its D/D matrix, an upper triangular matrix, and then the

ALE-index χ of this matrix. To reduce variations caused by different lengths of sequences, one can consider the normalized ALE-index, i.e., $\chi' = \chi/n$, where n is the number of codons of DNA sequence considered and the order of the corresponding D/D matrix as well.

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

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTG GGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTGACCGGCTTCTGGGGC AAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTC TGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCC
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTG TGGGGAAAGGTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGG GGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGC
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGGCCCTGGGCAGGTTGGTATCAAGG

Table 2. Normalized ALE-Indices of Matrix D/D for first exon of β-Globin of the nine Species.

Species	n	χ	χ'
Human	30	45.5281	1.5176
Rabbit	30	41.5327	1.3844
Goat	28	45.7454	1.6337
Gallus	30	35.3984	1.1799
Mouse	30	41.2471	1.3749
Rat	30	41.3979	1.3799
Gorilla	31	46.6585	1.5051
Bovine	28	44.6586	1.5949
Chimpanzee	35	52.6120	1.5032

In Table 3, we will show the similarity/dissimilarity table for the coding sequences of Table 1 based on the normalized ALE-index of the D/D matrix of these sequences.

Table 3. The similarity/dissimilarity matrix for the coding sequences of table 1 based on the normalized ALE-index of the D/D matrix.

Species	Human	Rabbit	Goat	Gallus	Mouse	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.1332	0.1161	0.3377	0.1427	0.1377	0.0125	0.0773	0.0144
Rabbit		0	0.2493	0.2045	0.0095	0.0045	0.1207	0.2105	0.1188
Goat			0	0.4538	0.2588	0.2538	0.1286	0.0388	0.1305
Gallus				0	0.1950	0.2000	0.3252	0.4150	0.3233
Mouse					0	0.0050	0.1302	0.2200	0.1283
Rat						0	0.1252	0.2150	0.1233
Gorilla							0	0.0898	0.0019
Bovine								0	0.0917
Chimpanzee									0

On observing Table 3, we note that the sequence of human and chimpanzee are similar so are human and gorilla, mouse and rat, mouse and rabbit, while gallus has great dissimilarity with others. The fact that gallus is a non-mammalian while all others are mammals in the above table might be a reason for this very different result. This is analogous to the results reported by other authors [8, 30-33].

5. Conclusion

We give a new graphical representation for DNA sequence based on codons and then we use ALE-index for presenting numerical representation. ALE-index is an invariant for DNA sequences which is very simple for calculation so that it can be directly used to handle long DNA sequences. Construction of similarity/dissimilarity table based on this invariant for DNA sequences of the first exon of the β-globin gene from nine species illustrates the utility of this representation.

Acknowledgement

The authors would like to thank the referee for the valuable comments.

References

[1] A. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, *Chem. Phys. Lett.* **368** (2003) 102–107.

[2] X. Guo, A. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, *Chem. Phys. Lett.* **369** (2003) 361– 366.

[3] 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.

[4] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences, *ARKIVOC* **9** (2006) 211–238.

[5] B. Liao, T. M. Wang, New 2D graphical representation of DNA sequences, *J. Comput. Chem.* **25** (2004) 1364–1368.

[6] 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.

[7] B. Liao, A 2D graphical representation of DNA sequence, *Chem. Phys. Lett.* **401** (2005) 196–199.

[8] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3- D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[9] J. F. Yu, J. H.Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 493–512.

[10] M. Randić, A. T. Balaban, On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci*. **43** (2003) 532–539.

[11] 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.

[12] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol*. **119** (1986) 319–328.

[13] A. Nandy, A new graphical representation and analysis of DNA sequence structure I. Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–313.

[14] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci*. **11** (1995) 503–507.

[15] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci*. **70** (1996) 661–668.

[16] R. Wu, Q. Hu, R. Li, G. Yue, A novel composition coding method of DNA sequence and its application, *MATCH Commun. Math. Comput. Chem*. **67** (2012) 269–276.

[17] X. Zhou, K. Li, M. Goodman, A. Sallam, A novel approach for the classical Ramsey number problem on DNA-based supercomputing, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 347–370.

[18] Q. Zhang, B. Wang, On the bounds of DNA coding with H-distance, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 371–380.

[19] W. Wang, T. Wang, Conditional LZ complexity and its application in mtDNA sequence analysis, *MATCH Commun*. *Math. Comput. Chem*. **66** (2011) 425–443.

[20] Q. Zhang, B. Wang, X. Wei, Evaluating the different combinatorial constraints in DNA computing based on minimum free energy, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 291–308.

[21] Y. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 477–488.

[22] R. Wu, R. Li, B. Liao, G. Yue, A novel method for visualizing and analyzing DNA sequences, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 679–690.

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

[24] J. Pesek, A. Žerovnik, Numerical characterization of modified Hamori curve representation of DNA sequences, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 301–312.

[25] Y. Zhang, W. Chen, New invariant of DNA sequences*, MATCH Commun. Math. Comput. Chem.* **58** (2007) 197–208.

[26] V. Aram, A. Iranmanesh, 3D-dynamic representation of DNA sequences, *MATCH Commun. Math. Comput. Chem*. **67** (2012) 809–816.

[27] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATH Commun. Math. Comput. Chem*. **56** (2006) 209–216.

[28] C. Li, J. Wang, New invariant of DNA sequences, *J. Chem. Inf. Model*. **45** (2005) 115–120.

[29] C. Li, N. Tang, J. Wang, Directed graphs of DNA sequences and their numerical characterization, *J. Theor. Biol*. **241** (2006) 173–177.

[30] M. Randić, M. Vračko, On the similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci*. **40** (2000) 599–606.

[31] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **371** (2003) 202–207.

[32] Y. Liu, The numerical characterization and similarity analysis of DNA primary sequences, *Internet El. J. Mol. Des.* **1** (2002) 675–684.

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