MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **61** (2009) 541-552

ISSN 0340 - 6253

A New Method to Analyze the Similarity Based on Dual Nucleotides of the DNA Sequence

Zanbo Liu*, Bo Liao, Wen Zhu

School of computer and communication, Hunan University, Changsha, Hunan, 410082, P R China

(Received May 5, 2008)

Abstract: In this paper, we propose a new method to analyze the similarity/dissimilarity of DNA sequences based on the neighboring dual nucleotides of DNA sequences. Therefore, we can decrease briefly a DNA sequence into a plot set in two-dimensional space. The proposed method is tested on the coding sequences of the first exon of β—globin gene belonging to eleven species. The utility of our approach can be illustrated by the examination of similarities /dissimilarities among the coding sequences.

1. Introduction

The graphical representation of DNA sequences was presented by Hamori and Ruskin[1]. In recent years, many methods have been proposed to numerically characterize DNA sequences based on multiple dimension space such as 2D, 3D, 4D and 6D. The idea of this approach is to transform a DNA sequence to a space curve. The graphical representation method is one of those method, which has such an important advantage over other methods: it allows visual inspection of data, helping to recognize major differences among similar DNA sequences. Many authors have presented new representation of DNA sequences to represent DNA sequences based on 2D, 3D, 4D and 6D [2-17,26]. The graphical techniques have emerged as a very powerful tool for the visualization and analysis of long DNA sequences [27]. These

* Corresponding author. Fax: +86 731 8821715
E-mail address: liuzanbo1981@163.com (Zanbo Liu)

techniques provide useful insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence that are not so easily obtainable by other methods. However, some graphical representations of DNA sequences are accompanied with some loss of information, due to overlapping and crossing of the curve representing DNA with itself. For the sake of defeat the limitation of both crossing and uniqueness, Liao in 2D [2] and Zhang in 2D [18] suggest their novel graphical representation approaches. In most of these approaches, they used the L/L matrix and leading eigenvalue of L/L and M/M matrices to compute the similarity of the sequences. But the computation is so complicated for a long sequence and authors only consider a simple nucleotide which corresponds a plot in space in the proposed methods. Recently, Qi[19] proposed a 2D graphical representation of DNA sequence based on dual nucleotides and Liao[17] also outlined a 4D representation of DNA sequences based on dual nucleotides. However, the coordinates of the plot is large and artificial for the method of Qi[19], and it is difficult to visualize for the approach by Liao[17].

Motivated by searching an efficient descriptor of DNA sequences, we propose a new method to analyze DNA sequence. In our paper, in order to provide a direct and simple graphical approach that can demonstrate the traits of DNA sequences clearly, we proposed a new 2D dual nucleotides approach of DNA graphical representation. A covariance matrix is applied in making comparison of DNA sequence. We consider the properties of the neighboring dual nucleotides and make analysis of similarity/dissimilarity among the coding sequences of the first exon of β- globin gene belonging to eleven species[21]. Furthermore, our method is rapid while it assures the validity because the whole process does not relate to complex algorithm.

2. Proposed approach

As we all know, in a DNA primary sequence, there are the four DNA bases A , C , G and T. They can be classified into classes: purine R={A , G} / primidine Y={C , T}, amino M={A , C}/keto K={G , T}, and weak H bond W={A , T}/strong H bond

S={C , G} according to their chemical properties. By considering neighboring two bases ,we can obtain sixteen combinations : AA , AT , AG , AC , TA , TT , TG , TC , GA , GT , GG , GC , CA , CT , CG and CC. So we can attain dual nucleotides set :{ AA , AT , AG , AC , TA , TT , TG , TC , GA , GT , GG , GC , CA , CT , CG , CC}. Randic [20] put forward 'magic circle' to uniformly distribute 20 amino acids on the circumference, therefore we also consider the 'magic circle' illustrated in Fig 1, which is of unit radius and on the circumference of which are uniformly distributed 16 dual nucleotides. Firstly, seen from Fig 1, we know that their orders are not unique, and there are 16! combinations. So we have to select a particular order of dual nucleotides. We adopted alphabetical ordering of dual nucleotides. The 16 points on the circumference of the circle have the coordinates given by

$$\theta_i = 2i\pi/16 \qquad \text{for} \quad i = 0,1,2,3,\cdots,15$$

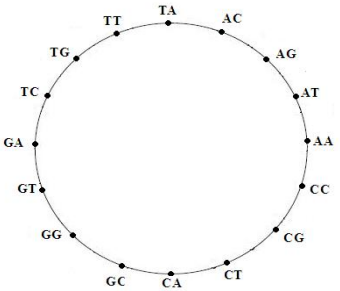

Fig 1. Sixteen dual nucleotides uniformly placed on the circumference of unit circle

Based on the above dual nucleotides set, we propose a new method to represent DNA sequences using 2D graphical representation. We also can obtain a curve on the Cartesian coordinate system. We put a positive integer i $(i = 1,2,3,\cdots,N-1)$,where N is the length of the DNA sequence being studied, to +x, and the circular measure angle of 16 dual nucleotides to +y, while the corresponding curve extends in the first quadrant. In detail, let $G = g_1g_2g_3\ldots g_N$ be an arbitrary DNA primary sequence. Then we define a map Φ, which maps G into a plot set. So that we will reduce a DNA

sequence into a series of nodes p_1 , p_2 ,p_3 ,...,p_{N-1}, whose coordinates is x_i ,y_i ($i = 1,2,3,\cdots,N-1$,where N is the length of the DNA sequence being studied) satisfy:

$$\Phi(g_i g_{i+1}) = \begin{cases} (i,0) & if \quad g_i g_{i+1} = AA \\ (i,\frac{2\pi}{16}) & if \quad g_i g_{i+1} = AT \\ (i,\frac{4\pi}{16}) & if \quad g_i g_{i+1} = AG \\ (i,\frac{6\pi}{16}) & if \quad g_i g_{i+1} = AC \\ (i,\frac{8\pi}{16}) & if \quad g_i g_{i+1} = TA \\ (i,\frac{10\pi}{16}) & if \quad g_i g_{i+1} = TT \\ (i,\frac{12\pi}{16}) & if \quad g_i g_{i+1} = TG \\ (i,\frac{14\pi}{16}) & if \quad g_i g_{i+1} = TC \\ (i,\frac{16\pi}{16}) & if \quad g_i g_{i+1} = GA \\ (i,\frac{18\pi}{16}) & if \quad g_i g_{i+1} = GT \\ (i,\frac{20\pi}{16}) & if \quad g_i g_{i+1} = GG \\ (i,\frac{22\pi}{16}) & if \quad g_i g_{i+1} = GC \\ (i,\frac{24\pi}{16}) & if \quad g_i g_{i+1} = CA \\ (i,\frac{26\pi}{16}) & if \quad g_i g_{i+1} = CT \\ (i,\frac{28\pi}{16}) & if \quad g_i g_{i+1} = CG \\ (i,\frac{30\pi}{16}) & if \quad g_i g_{i+1} = CC \end{cases} \qquad (1)$$

For example, the corresponding plot set of the sequence ATGGTGCACC, which is the first 10 bases of the coding sequence of the first exon of human β- globin gene, is $\{ (1,\frac{1}{8}\pi),(2,\frac{3}{4}\pi),(3,\frac{5}{4}\pi),(4,\frac{9}{8}\pi),(5,\frac{3}{4}\pi),(6,\frac{11}{8}\pi),(7,\frac{3}{2}\pi),(8,\frac{3}{8}\pi),(9,\frac{15}{8}\pi) \}$. We listed the Cartesian 2D coordinates for the sequence ATGGTGCACC of the coding sequence of the first exon of humanβ- globin gene in Table 1.

Table 1

Cartesian 2-D coordinates for the sequence ATGGTGCACC of the coding sequence of the first exon of humanβ- globin gene

Base	dual nucleotides	x	y
1	A T	1	$\pi/8$
2	T G	2	$3\pi/4$
3	G G	3	$5\pi/4$
4	G T	4	$9\pi/8$
5	T G	5	$3\pi/4$
6	G C	6	$11\pi/8$
7	C A	7	$3\pi/2$
8	A C	8	$3\pi/8$
9	C C	9	$15\pi/8$

We called the corresponding plot set be dual nucleotides plot set. The curve connected all plots of the dual nucleotides plot set in turn is called a dual nucleotides curve.

From the Fig 2, we can illustrate the DN curve of the DNA segment consisting of the first 10 bases, ATGGTGCACC, of the coding sequence of the first exon of humanβ- globin gene.

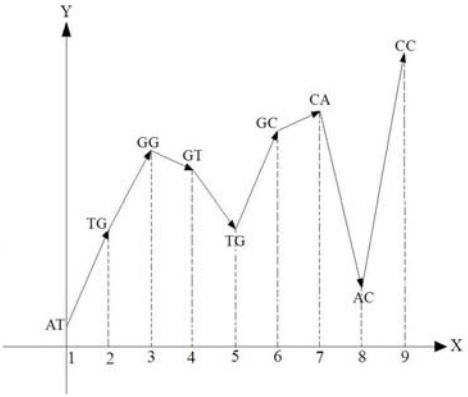

Fig 2. The lines representation of the sequence ATGGTGCACC of the coding sequence of the first exon of human β - globin gene

3. Application

For any DNA sequence, we can translate it to a set of points (x_i, y_i), $i = 1,2,3,\cdots,N-1$, where N is the length of the sequence. The coordinates of the geometrical center of the points, denoted by $\bar{x}$ and $\bar{y}$, may be calculated as follows:

$$\bar{x} = \frac{1}{N-1}\sum_{i=1}^{N-1} x_i \quad , \quad \bar{y} = \frac{1}{N-1}\sum_{i=1}^{N-1} y_i \qquad (2)$$

We construct a covariance matrix M, there

$$M = \begin{pmatrix} M_{xx} & M_{xy} \\ M_{yx} & M_{yy} \end{pmatrix}$$

The element of covariance matrix M of the points are defined:

$$\begin{cases} M_{xx} = \dfrac{\sum_{i=1}^{N-1}(x_i-\bar{x})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{N-1}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{N-1}(x_i-\bar{x})^2}} \\ M_{xy} = \dfrac{\sum_{i=1}^{N-1}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N-1}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{N-1}(y_i-\bar{y})^2}} = M_{yx} \\ M_{yy} = \dfrac{\sum_{i=1}^{N-1}(y_i-\bar{y})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N-1}(y_i-\bar{y})^2}\cdot\sqrt{\sum_{i=1}^{N-1}(y_i-\bar{y})^2}} \end{cases} \qquad (3)$$

The above four numbers give a quantitative description of a set of point (x_i, y_i), $i = 1,2,3,\cdots,N-1$, where N is the length of the sequence, scattering in a two-dimensional space. Obviously, the matrix is a real symmetric 2×2 one. There are two eigenvalues for giving a matrix M. We can apply the eigenvalues of M to make analysis of similarity/dissimilarity. In Table 2, we listed the first exon-1 of the β-globin gene for eleven different species, which were reported by Randic et al.[21].

Table 2

The coding sequence of the first exon of human β - globin gene of eleven different species

Species	Coding sequence
human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGC CCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGG CCCTGGGCAG
goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTG GGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTG GGCAG
gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGG CCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAG CCCTGGCCAG
opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTAC CATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGG CCCTTGGCAG
lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCT CTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGC CTTGGGCAG
mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTG CCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGG CCCTGGGCAGG
rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGC CCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAG GCCCTGGGC
rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGC CTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGC CCTGGGCAG
bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTG GGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTG GGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGC CCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGG CCCTGGGCAGG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGC CCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGG CCCTGGGCAGGTTGGTATCAAGG

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we extract an angle scale as defined below. In Table 3, we listed the two eigenvalues belonging to 11 species. Supposed that there are two species i and j (i, j = 1, 2, … , L, where L is the total number of the different species),

the parameters are λ_1^i , λ_2^i , λ_1^j , λ_2^j , where λ_1^i , λ_2^i and λ_1^j , λ_2^j are the two eigenvalues of matrix M_i and M_j , respectively, corresponding to species i and j. To reflect the different between the trends of every two 2D curves, the angles between the corresponding eigenvectors of every two species are used. The 2D vectors are denoted as follows:

$$\alpha_i = (\lambda_1^i, \lambda_2^i) \quad , \quad \alpha_j = (\lambda_1^j, \lambda_2^j) , \quad (i, j = 1, 2, \cdots, L) \qquad (4)$$

Table 3

The eigenvalues of the first exon of human β - globin gene belonging to eleven different species

Species	λ_1	λ_2
human	1.0046	0.9954
goat	1.0187	0.9813
gallus	1.1347	0.8653
opossum	1.0872	0.9128
lemur	1.1528	0.8472
mouse	1.0713	0.9287
rabbit	1.0818	0.9182
rat	1.1244	0.8756
bovine	1.0390	0.9610
Gorilla	1.0144	0.9856
Chimpanzee	1.0029	0.9971

The similarities among such vectors can be computed in this way: by calculating the cosine of the correlation angle. The cosine of θ_{ij} between the two vectors is:

$$\cos(\theta_{ij}) = \frac{\alpha_i \cdot \alpha_j}{|\alpha_i| \cdot |\alpha_j|} \quad , \quad (i, j = 1, 2, \cdots, L) \qquad (5)$$

The cosine of θ_{ij} denotes the angle between the geometric centers of the *i*th and the *j*th genomes, and M is the total number of all genomes (L=11, here). Then we obtain a real L×L symmetric matrix whose elements are $\cos(\theta_{ij})$. The larger the

cosine of the correlation angle between two species, the more similar are the DNA sequences. That is to say, the cosine of the angle among evolutionary closely related species is larger, moreover those among evolutionary disparate species are smaller.

Observing Table 4, we find that human, gorilla and chimpanzee are greatly similar to each other. The pairs human ~ chimpanzee, human ~ gorilla, and gorilla ~ chimpanzee are the most similar species pairs according to the smallest entries, while lemur and opossum are dissimilarity to others. The results are not accidental, but show these species have closely evolutionary relationship. The similarity results can be found in references [21-25].

Table 4

The similarity/dissimilarity matrix for the coding sequence based on the cosine of the angle between the 2-component vectors of the eigenvalues of M matrices

Species	Human	goat	gallus	opossum	lemur	mouse	rabbit	rat	bovine	Gorilla	Chimpanzee
human	1.000000	0.999901	0.991653	0.996609	0.989211	0.997784	0.997036	0.992908	0.999409	0.999952	0.999999
goat		1.000000	0.993372	0.997670	0.991178	0.998623	0.998021	0.994486	0.999794	0.999991	0.999875
gallus			1.000000	0.998900	0.999843	0.998034	0.998634	0.999949	0.995499	0.992869	0.991433
opossum				1.000000	0.997911	0.999875	0.999986	0.999324	0.998848	0.997367	0.996467
lemur					1.000000	0.996766	0.997550	0.999612	0.993662	0.990599	0.988960
mouse						1.000000	0.999946	0.998618	0.999482	0.998388	0.997670
rabbit							1.000000	0.999112	0.999091	0.997742	0.996903
rat								1.000000	0.996408	0.994026	0.992705
bovine									1.000000	0.999698	0.999349
Gorilla										1.000000	0.999934
Chimpanzee											1.000000

4. Conclusion

In this letter, based on the neighboring dual nucleotides we proposed a novel 2D graphical representation of DNA sequences: DN curve, and outlined an approach to make analysis of similarity/dissimilarity of DNA sequences. The advantage of our approach is that allow visual inspection of data based on dual nucleotides and the sequence invariant easily computed. And our representation provides a direct plotting

method to denote DNA sequences without degeneracy. Graphical representation of DNA sequences provides a tool to the molecular biologists to analysis the local and global features of long or short DNA sequences in order to derive some kind of relative ranking of the sequence, for evolutionary or prediction of functional properties. Comparing the previous scheme[19,26], the advantage of our method is that for a long sequence, a large D/D matrix or a large L/L matrix needn't to be computed, while the computation of the covariance matrix is simple, and we considered that the neighboring dual nucleotides is not a single nucleotide, so more information will be obtained.

5. Acknowledgement

This work is supported in part by the National Nature Science Foundation of China (Grant 10571019)

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.

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

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

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

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

[6] S.S.T.Yan, J.S.Wang, A.Niknejad, C.X.Lu, N Jin, Y.K.Ho, DNA sequence representation without degeneracy, *Nucl. Aci. Res.* **31** (2003) 3078-3080.

[7] M.Randić, M.Vračko, A.Nandy, S.C.Basak, On 3-D graphical representation of

DNA primary sequence and their numerical characterization, *J. Chem. Inf. Comput. Sci*. **40** (2000) 1235-1244.

[8] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numberical characterization, *Chem. Phys. Lett.* **368** (2003) 1-6.

[9] A. Nandy, A new graphical representation and analysis of DNA sequence structure: methodology and Application to Globin Genes, *Curr. Sci.* **66** (1994) 309-314.

[10] A. Nandy, Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences, *Comput. Appl. Biosci.* **12** (1996) 55-62.

[11] E. Hamori, J. Ruskin, H curves, a novel method of representaion of nucleotides series especially suited for long DNA sequences. *J. Biol. Chem.* **258** (1983) 1318-1327.

[12] E. Hamori, Novel DNA sequence representations, *Nature* **314** (1985) 585-586.

[13] M.A. Gates, Simple DNA sequence representations, *Nature* **316** (1985) 219.

[14] J. Wang, Y. Zhang, Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation, *Chem. Phys. Lett.* **423** (2003) 50-53.

[15] H. J. Jeffrey, Nucleic. Chaos game representation of gene structure, *Acids. Res.* **18** (1990) 2163-2170.

[16] J. Hu, A. H. Wang, Comment on 'Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation' [Chem. Phys. Lett., 411 (2005) 248], *Chem. Phys. Lett.* **424** (2006) 453-455.

[17] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/Dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comp. Chem.* **59** (2008) 647-652.

[18] Y. S. Zhang, B. Liao, K. Q, Ding, On 2D graphical representation of DNA sequence of nondegeneracy, *Chem. Phys. Lett.* **411** (2005) 28-32.

[19] Z.H. Qi, X.Q. Qi, Novel 2D graphical representation of DNA sequence based on

dual nucleotides, *Chem. Phys. Lett.* **440** (2007) 139-144.

[20] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **419** (2006) 528-532.

[21] 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.

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

[23] B. Liao, Y.S. Zhang, K.Q. Ding, T.M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *J. Mol. Struct. (THEOCHEM)* **717** (2005) 199-203.

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

[25] Y. H. Yao, X. Y. Nan, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, *Chem. Phys. Lett.* **411** (2005) 248-255.

[26] Z.H. Qi, X.Q. Qi, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434-440.

[27] B. Liao, X. Z. Shan, W. Zhu, R. F. Li, phylogenetic tree construction based on 2D graphical representation, *Chem. Phys. Lett.* **422** (2006) 282-288.