MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **59** (2008) 647-652

ISSN 0340 - 6253

Analysis of Similarity/Dissimilarity of DNA Sequences Based on Dual Nucleotides

Bo Liao *, Cheng Zeng, Fuqiang Li , Yong Tang
School of Computer and Communication, Hunan University
Changsha Hunan 410082,China

(Received August 23, 2007)

Abstract. Based on the neighboring dual nucleotides, we reduce a DNA sequence into a plot set in four-dimensional space, and associate with the eigenvalues of the introduced covariance matrix. The examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene belonging to eleven species illustrates the utility of our approach.

1 Introduction

One important task in the study of genome sequences is to determine densities of specific nucleotides and to understand the implications for exons or coding regions. Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists. Thus more and more mathematical methods are applied in the gene research. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures. In recent years several authors outlined different graphical representation of DNA sequences based on 2-D, 3-D or 4D[1-5,7-15]. These techniques provide useful insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence which are not as easily obtainable by other methods. Based on these graphical representation, several authors outlined some approaches to make comparison of DNA sequences[6, 16-21]. In most of these approaches, the L/L matrix and leading eigenvalues of L/L matrices are used. For a long sequence, the computation is complicated. And in all the presented approaches, authors only consider a simple nucleotide which corresponds a plot in space. Recently, Qi[22] introduced a 2D graphical representation of DNA sequence based on dual nucleotides. But the coordinates of the plot is large and artificial.

Here, in order to provide a direct and simple approach that can display the features of DNA sequences, we consider the properties of the neighboring dual nucleotides and make analysis of similarities among the coding sequences of the first exon of β-globin gene belonging to eleven species[17]. A covariance matrix is applied in making comparison of DNA sequence.

*Corresponding author E-mail: dragonbw@163.com;Fax:+86-731-8821715

2 Mathematical model

As we all know, the four DNA bases A,C,G and T can be divided into three classes: purine R={A,G}/pyrimidine Y={C,T},amino M={A,C}/keto K={G,T}, and weak-H bond W={A,T}/strong-H bong S={C,G} according to their chemical properties. By considering neighboring two bases, we can obtain sixteen combinations: AG,GA,CT,TC,AC,CA,GT,TG, AT,TA,CG,GC, AA,CC,GG and TT. Consequently, we obtain dual nucleotides set: {AG,GA,CT,TC,AC,CA,GT,TG,AT,TA,CG,GC,AA,CC,GG, TT}. According to their chemical properties, the dual nucleotides can be divided into four classes: purine dual nucleotides {AG,GA}/pyrimidine dual nucleotides{CT,TC},amino dual nucleotides{AC,CA}/keto dual nucleotides{TG,GT},weak-H bond dual nucleotides{AT,TA}/strong-H bong dual nucleotides {CG,GC},and repeat dual nucleotides{AA,CC,GG,TT}.

Based on the above four class dual nucleotides sets, we present a mathematical model which translates a nucleotide into a plot in the four-dimensional space. In detail, let $G = g_1 g_2 \dots g_n$ 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_0 , P_1 , P_2 , $\dots$, P_{n-1} , whose coordinates $x_i, y_i, z_i, s_i (i = 1, 2, \dots, n-1$, where n is the length of the DNA sequence being studied) satisfy

$$\begin{cases} x_i = \frac{AG_i+GA_i+CT_i+TC_i}{i} \\ y_i = \frac{AC_i+CA_i+GT_i+TG_i}{i} \\ z_i = \frac{AT_i+TA_i+CG_i+GC_i}{i} \\ s_i = \frac{AA_i+CC_i+GG_i+TT_i}{i} \end{cases}$$

where $AG_i, GA_i, CT_i, TC_i, AC_i, CA_i, GT_i, TG_i, AT_i, TA_i, CG_i, GC_i, AA_i, CC_i, GG_i$ and TT_i are the cumulative occurrence numbers of AG,GA,CT,TC,AC,CA,GT,TG,AT,TA,CG,GC, AA,CC,GG and TT, respectively, in the subsequence from the 1st base to the i-th base in the sequence. We define $AG_0 = GA_0 = CT_0 = TC_0 = AC_0 = CA_0 = GT_0 = TG_0 = AT_0 = TA_0 = CG_0 = GC_0 = AA_0 = CC_0 = GG_0 = TT_0 = 0$.

3 Similarity analysis

For any sequence, we have a set of points $(x_i, y_i, z_i, s_i), i = 1, 2, 3, \dots, n-1$, where n is the length of the sequence. The coordinates of the geometrical center of the points, denoted by x^0, y^0, z^0 and s^0, may be calculated as follows:

$$x^0 = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i \quad , \quad y^0 = \frac{1}{n-1}\sum_{i=1}^{n-1} y_i \quad , \quad z^0 = \frac{1}{n-1}\sum_{i=1}^{n-1} z_i \quad , \quad s^0 = \frac{1}{n-1}\sum_{i=1}^{n-1} s_i \qquad (1)$$

We construct a covariance matrix CM, where, $CM = \begin{pmatrix} CM_{xx} & CM_{xy} & CM_{xz} & CM_{xs} \\ CM_{yx} & CM_{yy} & CM_{yz} & CM_{ys} \\ CM_{zx} & CM_{zy} & CM_{zz} & CM_{zs} \\ CM_{sx} & CM_{sy} & CM_{sz} & CM_{ss} \end{pmatrix}$.

And the element of covariance matrix are defined:

$$\begin{cases} CM_{xx} = \frac{1}{n-1}\sum_1^{n-1}(x_i - x^0)(x_i - x^0) \\ CM_{xy} = \frac{1}{n-1}\sum_1^{n-1}(x_i - x^0)(y_i - y^0) = CM_{yx} \\ CM_{xz} = \frac{1}{n-1}\sum_1^{n-1}(x_i - x^0)(z_i - z^0) = CM_{zx} \\ CM_{xs} = \frac{1}{n-1}\sum_1^{n-1}(x_i - x^0)(s_i - s^0) = CM_{sx} \\ CM_{yy} = \frac{1}{n-1}\sum_1^{n-1}(y_i - y^0)(y_i - y^0) \\ CM_{yz} = \frac{1}{n-1}\sum_1^{n-1}(y_i - y^0)(z_i - z^0) = CM_{zy} \\ CM_{ys} = \frac{1}{n-1}\sum_1^{n-1}(y_i - y^0)(s_i - s^0) = CM_{sy} \\ CM_{zz} = \frac{1}{n-1}\sum_1^{n-1}(z_i - z^0)(z_i - z^0) \\ CM_{zs} = \frac{1}{n-1}\sum_1^{n-1}(z_i - z^0)(s_i - s^0) = CM_{sz} \\ CM_{ss} = \frac{1}{n-1}\sum_1^{n-1}(s_i - s^0)(s_i - s^0) \end{cases} \tag{2}$$

The above four numbers give a quantitative description of a set of point $(x_i, y_i, z_i, s_i), i = 1, 2, \ldots, n-1$, scattering in a four-dimensional space. The eigenvalues of CM are applied to make analysis of similarity. In Table 1, the first exon-1 of the b-globin gene for 11 different species are listed, which were reported by Randic et al. [17]. In table 2, we list the eigenvalues belonging to 11 species.

Table 1: The coding sequences of the first exon of β-globin gene of eleven different species

Species	Coding sequence
human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGT GGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGT GCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGT CAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemmur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGT GGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGG TGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGT GAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGT GAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTG AACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we introduce an angle scale as defined below. Suppose that there are two species i and j, the parameters are $\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i, \lambda_1^j, \lambda_2^j, \lambda_3^j, \lambda_4^j$, respectively, where $\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i$ are the four eigenvalues of matrix CM_i corresponding to species i. We will illustrate the use of the 4-D quantitative characterization of DNA sequence with an examination of similarities/dissimilarties among the eleven species. We construct a four-component vector consisting of the four eigenvalues of matrix CM. The underlying assumption is that if two

Table 2: The eigenvalues of the first exon of β-globin gene belonging to eleven species

	λ_1	λ_2	λ_3	λ_4
Human	0	0.0037	0.0094	0.0207
Goat	0	0.0042	0.0088	0.0215
Gallus	0	0.0048	0.0117	0.0182
Opossum	0	0.0014	0.0094	0.0225
Lemur	0	0.0030	0.0033	0.0140
Mouse	0	0.0039	0.0098	0.0179
Rabbit	0	0.0040	0.0069	0.0201
Rat	0	0.0032	0.0091	0.0142
Bovine	0	0.0045	0.0123	0.0224
Gorilla	0	0.0031	0.0094	0.0206
Chimpanzee	0	0.0034	0.0083	0.0182

vectors point to a similar direction in the four-dimensional space, and then the two DNA sequences represented by the four-component vectors are similar. The similarities among such vectors can be computed by calculating the cosine of the angle between the vectors. The cosine of θ_{ij} between the two vectors is:

$$cos(\theta_{ij}) = \frac{\lambda_1^i\lambda_1^j + \lambda_2^i\lambda_2^j + \lambda_3^i\lambda_3^j + \lambda_4^i\lambda_4^j}{\sqrt{(\lambda_1^i)^2 + (\lambda_2^i)^2 + (\lambda_3^i)^2 + (\lambda_4^i)^2}\sqrt{(\lambda_1^j)^2 + (\lambda_2^j)^2 + (\lambda_3^j)^2 + (\lambda_4^j)^2}} \tag{3}$$

The larger cosine is, the more similar are the DNA sequences. That is to say, the cosine between evolutionary closely related species is larger, while those between evolutionary disparate species are smaller.

Observing Table 3, we find that the more similar species pairs are $Chimpanzee \sim Human, Chimpanzee \sim Gorilla, Gorilla \sim Human$ and $Mouse \sim Bovine$, while Lemur and Opossum are dissimilarity to others. The similar results can be found in references[6,16,17,21].

4 Conclusion

In this paper, we considered the properties of the neighboring dual nucleotides and outlined an approach to make analysis of DNA sequences. It is useful for computational scientists and biologists to visualize the local and global features of long or short DNA sequences. The advantage of our method is that also allow visual inspection of data based on dual nucleotides and the computation is simple. Comparing the previous methods, the advantage of our method is that for a long sequence, a large D/D matrix or a large L/L matrix is needn't computed, while the computation of the covariance matrix is simple, and we considered the properties of the neighboring dual nucleotides not a single nucleotide, more information will be obtained.

Table 3: The similarity/dissimilarity matrix for the coding sequences based on the cosine of the angle between the 4-component vectors consisting the eigenvalues of the CM matrices

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat	Bovine	Gorilla	Chimpan
Human	1.0000	0.9992	0.9883	0.9941	0.9807	0.9969	0.9953	0.9897	0.9971	0.9997	1.0000
Goat		1.0000	0.9832	0.9926	0.9878	0.9938	0.9984	0.9841	0.9937	0.9983	0.9992
Gallus			1.0000	0.9721	0.9453	0.9972	0.9718	0.9995	0.9967	0.9867	0.9889
Opossum				1.0000	0.9759	0.9860	0.9897	0.9767	0.9876	0.9964	0.9933
Lemur					1.000	0.9652	0.9951	0.9453	0.9644	0.9789	0.9808
Mouse						1.0000	0.9860	0.9977	0.9999	0.9960	0.9972
Rabbit							1.0000	0.9725	0.9857	0.9941	0.9953
Rat								1.0000	0.9977	0.9889	0.9901
Bovine									1.0000	0.9966	0.9973
Gorilla										1.0000	0.9995
Chimp											1.0000

5 Acknowledgment

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

References

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

[2] S. S. T. Yau, 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.

[3] 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.

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

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

[7] B. Liao, M. S. Tan, K. Q. Ding, Application of 2-D graphical representation of DNA sequence, *Chem. Phys. Lett.* **414** (2005) 296–300.

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

[15] B. Liao, M. S. Tan, K. Q. Ding, A 4D representation of DNA sequences and its application, *Chem. Phys. Lett.* **402** (2005) 380–383.

[16] 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.

[17] 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.

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

[19] B. Liao, X. Y. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196–1202.

[20] B. Liao, Y. S. Liu, R. F. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, *Chem. Phys. Lett.* **421** (2006) 313–318.

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

[22] Z. H. Qi, X. Q. Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides, *Chem. Phys. Lett.* **440** (2007) 139–144.