MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **59** (2008) 653-659

ISSN 0340 - 6253

A novel method for sequence similarity analysis based on the relative frequency of dual nucleotides

Jiawei Luo, Renfa Li *, Qingguang Zeng
School of Computer and Communication, Hunan University
Changsha, Hunan Province, 410082,China

(Received October 30, 2007)

Abstract. According to the three classifications of nucleotides, we divided the sixteen neighboring dual nucleotides into four classes. We associated each sequence with the relative frequencies of the dual nucleotides in each class, and obtain a sixteen-component vector relative to sixteen dual nucleotides. The introduced vector is applied to characterize and compare the coding sequences of the first exon of β-globin gene belonging to eleven species.

1 Introduction

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. In recent years, some researches proposed a class new method to view ,sort and compare sequences[1-24]. In these methods, a graphical representation of DNA sequence is introduced. Using the graphical representation, one can reduce a sequence into a series of nodes in two-dimension, three-dimension even high-dimension[1-4,6-12,19,23]. Based on the graphical representation, we can obtain some numerical characterization which can be applied to make similarity analysis, mutation analysis and alignment. In order to depict the numerical characterization and reduce the complexity of computation. Recently, Liao[20] and M.Randic[16] considered the triplets of nucleotide bases and proposed some methods to make similarity analysis of DNA sequences. Qi[23] introduced a 2D graphical representation of DNA sequence based on dual nucleotides. But the coordinates of the plot is large and

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

artificial.

In this letter, we consider the properties of the neighboring dual nucleotides and divide the dual nucleotides into four classes. The frequencies of the dual nucleotides in each class are applied to make similarities analysis among the coding sequences of the first exon of β-globin gene belonging to eleven species.

2 Similarity Analysis

In a DNA primary sequence, 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 and the base order, we can obtain sixteen combinations: AG,GA,CT,TC,AC,CA,GT,TG,AT,TA,CG,GC,AA,CC,GG and TT. According to the three classifications of the four DNA bases, 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 bond dual nucleotides{CG,GC},and repeat dual nucleotides{AA,CC,GG,TT}.

In each class, we consider the relative frequencies of the dual nucleotides. Consequently, the frequencies of the dual nucleotides were computed as following:

$$dn_1 = AG\% = \frac{AG_{n-1}}{AG_{n-1}+GA_{n-1}+CT_{n-1}+TC_{n-1}};\ dn_2 = GA\% = \frac{GA_{n-1}}{AG_{n-1}+GA_{n-1}+CT_{n-1}+TC_{n-1}};$$
$$dn_3 = CT\% = \frac{CT_{n-1}}{AG_{n-1}+GA_{n-1}+CT_{n-1}+TC_{n-1}};\ dn_4 = TC\% = \frac{TC_{n-1}}{AG_{n-1}+GA_{n-1}+CT_{n-1}+TC_{n-1}};$$
$$dn_5 = AC\% = \frac{AC_{n-1}}{AC_{n-1}+CA_{n-1}+GT_{n-1}+TG_{n-1}};\ dn_6 = CA\% = \frac{CA_{n-1}}{AC_{n-1}+CA_{n-1}+GT_{n-1}+TG_{n-1}};$$
$$dn_7 = GT\% = \frac{GT_{n-1}}{AC_{n-1}+CA_{n-1}+GT_{n-1}+TG_{n-1}};\ dn_8 = TG\% = \frac{TG_{n-1}}{AC_{n-1}+CA_{n-1}+GT_{n-1}+TG_{n-1}};$$
$$dn_9 = AT\% = \frac{AT_{n-1}}{AT_{n-1}+TA_{n-1}+GC_{n-1}+CG_{n-1}};\ dn_{10} = TA\% = \frac{TA_{n-1}}{AT_{n-1}+TA_{n-1}+GC_{n-1}+CG_{n-1}};$$
$$dn_{11} = GC\% = \frac{GC_{n-1}}{AT_{n-1}+TA_{n-1}+GC_{n-1}+CG_{n-1}};\ dn_{12} = CG\% = \frac{CG_{n-1}}{AT_{n-1}+TA_{n-1}+GC_{n-1}+CG_{n-1}};$$
$$dn_{13} = AA\% = \frac{AA_{n-1}}{AA_{n-1}+GG_{n-1}+CC_{n-1}+TT_{n-1}};\ dn_{14} = GG\% = \frac{GG_{n-1}}{AA_{n-1}+GG_{n-1}+CC_{n-1}+TT_{n-1}};$$
$$dn_{15} = CC\% = \frac{CC_{n-1}}{AA_{n-1}+GG_{n-1}+CC_{n-1}+TT_{n-1}};\ dn_{16} = TT\% = \frac{TT_{n-1}}{AA_{n-1}+GG_{n-1}+CC_{n-1}+TT_{n-1}};$$

where $AG_{n-1}, GA_{n-1}, CT_{n-1}, TC_{n-1}, AC_{n-1}, CA_{n-1}, GT_{n-1}, TG_{n-1}, AT_{n-1}, TA_{n-1},$

$CG_{n-1}, GC_{n-1}, AA_{n-1}, CC_{n-1}, GG_{n-1}$ and TT_{n-1} 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 (n-1)-th base in the sequence, n is the length of the studied sequence. We define $AG_0 = GA_0 = CT_0 = TC_0 = AC_0 = CA_0 = GT_0 = TG_0 = AT_0 = TA_0 = CG_0 = GC_0 = AA_0 = CC_0 = GG_0 = TT_0 = 0$. In table 1, we list the frequencies of the dual nucleotides in each class of the first exon of β-globin gene belonging to eleven species.

Table 1: The frequencies of the dual nucleotides in each class.

	purine/	pyrimidine	amino/	keto	weak-H bond/	strong-H bond	Repeat	
	AG%	CT%	AC%	GT%	AT%	CG%	AA%	CC%
	GA%	TC%	CA%	TG%	TA%	GC%	GG%	TT%
Human	0.3043	0.3043	0.1333	0.3000	0.1667	0.1667	0.1538	0.2692
	0.3043	0.0870	0.1000	0.4667	0.1667	0.5000	0.4615	0.1154
Goat	0.3077	0.3077	0.0870	0.2174	0.1538	0.1538	0.2174	0.1739
	0.3077	0.0769	0.1304	0.5652	0	0.6923	0.5217	0.0870
Gallus	0.2917	0.2917	0.1200	0.1600	0.2353	0.1765	0.2000	0.2800
	0.2500	0.1667	0.2800	0.4400	0	0.5882	0.5200	0
Opossum	0.2759	0.3103	0.2188	0.1875	0.3000	0	0.1500	0.2000
	0.2759	0.1379	0.2188	0.3750	0.2000	0.500	0.4500	0.2000
Lemur	0.3000	0.2667	0.0741	0.2593	0.3077	0.0769	0.1905	0.0952
	0.3000	0.1333	0.1481	0.5185	0.0769	0.5385	0.5238	0.1905
Mouse	0.2308	0.3462	0.0968	0.2581	0.2727	0.0909	0.2000	0.2800
	0.3077	0.1154	0.0968	0.5484	0	0.6364	0.4000	0.1200
Rabbit	0.3333	0.2083	0.0323	0.3548	0.3000	0.1000	0.2083	0.2083
	0.2917	0.1667	0.1290	0.4839	0	0.6000	0.5417	0.0417
Rat	0.2727	0.4091	0.1481	0.2222	0.2353	0.0588	0.2400	0.2400
	0.3182	0	0.0741	0.5556	0.2353	0.4706	0.4400	0.0800
Bovine	0.3478	0.2609	0.0833	0.2500	0.1667	0.1667	0.1923	0.1923
	0.3478	0.0435	0.1250	0.5417	0	0.6667	0.4615	0.1538
Gorilla	0.2917	0.2917	0.1290	0.2903	0.1818	0.1818	0.1538	0.2692
	0.3333	0.0833	0.0968	0.4839	0.0909	0.5455	0.5000	0.0769
Chimpanzee	0.3077	0.2692	0.1143	0.3143	0.2308	0.1538	0.1667	0.2333
	0.3077	0.1154	0.1143	0.4571	0.1538	0.4615	0.5000	0.1000

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we construct a sixteen-component vector consisting of the frequencies of the dual nucleotides in each class. The similarities between such vectors can be computed by calculating the Euclidean distance and by calculating the cosine of the angle between the vectors.

Suppose that there are two species i and j, the corresponding vectors are $(dn_1^i, dn_2^i, dn_3^i, dn_4^i, dn_5^i, dn_6^i, dn_7^i, dn_8^i, dn_9^i, dn_{10}^i, dn_{11}^i, dn_{12}^i, dn_{13}^i, dn_{14}^i, dn_{15}^i, dn_{16}^i)$ and $(dn_1^j, dn_2^j, dn_3^j, dn_4^j, dn_5^j, dn_6^j, dn_7^j, dn_8^j, dn_9^j, dn_{10}^j, dn_{11}^j, dn_{12}^j, dn_{13}^j, dn_{14}^j, dn_{15}^j, dn_{16}^j)$, respectively. The Euclidean distance between the two vectors is:

$$d_{ij} = \sqrt{\sum_{k=1}^{16}(dn_k^i - dn_k^j)^2} \tag{1}$$

The cosine of θ_{ij} between the two vectors is:

$$cos(\theta_{ij}) = \frac{\sum_{k=1}^{16} dn_k^i.dn_k^j}{\sqrt{\sum_{k=1}^{16}(dn_k^i)^2}.\sqrt{\sum_{k=1}^{16}(dn_k^j)^2}} \tag{2}$$

Obviously, the smaller Euclidean distance is,the more similar are the DNA sequences. The larger cosine is, the more similar are the DNA sequences. In table 2, we list the similarity/dissimilarity matrix for the coding sequences based on the Euclidian distance between the 16-component vectors consisting the frequencies of the dual nucleotides. In Table 3, we list the similarity/dissimilarity matrix for the coding sequences based on the cosine of the angle between the 16-component vectors consisting the frequencies of the dual nucleotides.

Table 2: The similarity/dissimilarity matrix for the coding sequences based on the Euclidian distance between the 16-component vectors consisting the frequencies of the dual nucleotides.

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat	Bovine	Gorilla	Chimpan
Human	0	0.1204	0.1219	0.1055	0.0924	0.0872	0.1095	0.0665	0.0833	0.0128	0.0120
Goat		0	0.0915	0.2131	0.0862	0.0607	0.0838	0.1563	0.0182	0.0611	0.1146
Gallus			0	0.1575	0.1339	0.1080	0.0984	0.2060	0.1184	0.0880	0.1122
Opossum				0	0.0955	0.1551	0.1944	0.1243	0.1980	0.1357	0.0948
Lemur					0	0.0870	0.0640	0.1307	0.0789	0.0843	0.0621
Mouse						0	0.0847	0.1153	0.0954	0.0618	0.0976
Rabbit							0	0.2034	0.0844	0.0735	0.0746
Rat								0	0.1588	0.0883	0.0804
Bovine									0	0.0519	0.0976
Gorilla										0	0.0205
Chimp											0

Table 3: The similarity/dissimilarity matrix for the coding sequences based on the cosine of the angle between the 16-component vectors consisting the frequencies of the dual nucleotides.

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat	Bovine	Gorilla	Chimpan
Human	1.0000	0.9690	0.9553	0.9580	0.9661	0.9699	0.9632	0.9757	0.9732	0.9956	0.9953
Goat		1.0000	0.9707	0.9306	0.9726	0.9806	0.9728	0.9488	0.9943	0.9816	0.9649
Gallus			1.0000	0.9418	0.9518	0.9622	0.9664	0.9257	0.9597	0.9681	0.9589
Opossum				1.0000	0.9652	0.9453	0.9327	0.9540	0.9322	0.9487	0.9620
Lemur					1.000	0.9695	0.9783	0.9527	0.9735	0.9693	0.9776
Mouse						1.0000	0.9712	0.9595	0.9800	0.9785	0.9662
Rabbit							1.0000	0.9297	0.9718	0.9752	0.9761
Rat								1.0000	0.9457	0.9678	0.9705
Bovine									1.0000	0.9831	0.9683
Gorilla										1.0000	0.9928
Chimp											1.0000

Observing Table 2 and Table 3, we find that the more similar species pairs are $Chimpanzee \sim Human, Chimpanzee \sim Gorilla$ and $Gorilla \sim Human$, while Lemur and Opossum are dissimilarity to others. The similar results can be found in references[5,13-16,19-20].

3 Conclusion

In this letter, we considered the properties of the neighboring dual nucleotides and outlined an approach to make similarity analysis of DNA sequences based on the frequencies of the dual nucleotides. 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 ia that allow visual inspection of data based on dual nucleotides and the computation is simple.

4 Acknowledgment

This work is supported in part by the National Nature Science Foundation of Hunan province (Grant 06JJ2090 and 06JJ4076).

References

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

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

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

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

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

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

[7] 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.

[8] 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.

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

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

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

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

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

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

[15] 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.

[16] M. Randić, X. F. Guo, S. C. Basak, On the Characterization of DNA Primary Sequence by Triplet of Nucleic Acid Bases, *J. Chem. Inf. Comput. Sci.* **41** (2001) 619-626.

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

[18] B. Liao, Y. Liu, R. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, *Chem. Phys. Lett.*, **412** (2006) 313-318.

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

[20] B. Liao, T. Wang, Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases, *J. Chem. Inf. Comput. Sci.*, **44** (2004) 1666-1670.

[21] B. Liao, K. Ding, Graphical Approach to Analyzing DNA Sequences, *J. Comput. Chem.*, **14** (2005) 1519-1523.

[22] B. Liao, R. Li, W. Zhu, On the similarity of DNA primary sequences based on 5-D representation, *J. Math. Chem.*, **1** (2007) 47-57.

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

[24] X. Zhang, J. Luo, L. Yang, New invariant of DNA sequence based on 3DD-curves and its application on phylogeny, *J. Comput. Chem.*, **28** (2007) 2342-2346.