DNA Sequences Analysis Based on Classifications of Nucleotide Bases

Long Shi[1,*] and Hailan Huang[2]

[1] School of Science, Central South University of Forest and Technology, Changsha, Hunan 410004, China
chu_9@163.com
[2] School of Mathematics and Computational Science, Xiangtan University, Xiangtan, Hunan 411105, China
huanghailan_1980@sina.com

Abstract. Four bases A, C, G and T of DNA sequences are divided into three kinds of classifications according to their chemical properties. We convert a DNA primary sequence into three symbolic sequences. The frequencies of group mutations of three symbolic sequences have been grouped into a twelve-component vector to represent the DNA sequence. The Euclidean distances among introduced vectors are applied to characterize and compare the coding sequences of the first exon of beta globin gene of 11 different species.

Keywords: bases classification, symbolic sequence, group mutation, similarity analysis.

1 Introduction

With the development of the sequencing technique, a large number of DNA primary sequence data are collected and become available in public databases. It is one of the challenges for bio-scientists to mathematically analyze the large volumes of biological data to find biological interest. In recent years, several authors outlined different graphical representations of DNA sequences based on 2D or 3D space[1-19]. 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. But some graphical representations of DNA sequence are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself[2-5]. To avoid the limitations associated with crossing and overlapping some authors provided many novel graphical representation methods recently[6-19]. These techniques provide useful insights into local and global characteristics along a sequence which are not easily obtained by other methods. Based on these graphical representations, several authors developed some numerical methods, for example E matrix, D/D matrix, M/M matrix, L/L matrix and their leading eigenvalues and so on, to compare DNA sequences[6-12]. But when the length of the

* Corresponding author.

J. Luo (Ed.): Affective Computing and Intelligent Interaction, AISC 137, pp. 379–384.
springerlink.com

sequences under studying is long the computation will be complicated and time-consuming.

In the past years the dinucleotide analysis has also been tried by several authors[18-20]. This approach can offer fast, qualitative comparisons and reveal the biology information of DNA sequences. In this paper, motivated by works on dinucleotide analysis, according to bases' chemical properties, we divide four bases into classes and use three mappings to convert a DNA primary sequence into three symbolic sequences. We construct a 12-component vector consisting of twelve frequencies of group mutations to represent a DNA sequence. The Euclidean distances between the end points of the vectors are applied to characterize and compare the coding sequences of the first exon of beta globin genes of 11 different species in Table 1.

Table 1. The coding sequences of the first exon of beta globin gene of 11 different species

Species	Coding sequences
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGG TCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGG GGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGG GCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGG GCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGACGAGAAGTCTGCGGTCACTGCCCTGTGG GGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGG GGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTT GGTATCAAGG

2 Proposed Method

For a DNA primary sequence, there are three kinds of classification methods to divide four bases A, C, G and T according to their chemical properties, i.e., purine group R={A,G} and pyrimidine group Y={C,T}; amino group M={A,C} and keto group K={G,T}; weak H-bond group W={A,T} and strong H-bond group S={C,G}. We call them RY classification, MK classification and WS classification correspondingly. Let $X = s_1 s_2 \cdots s_n$ be a DNA primary sequence with length n. we have three maps ϕ_{RY}, ϕ_{MK}, ϕ_{WS}, which map X into three symbolic sequences respectively. Explicitly,

$$\phi_{RY}(X) = \phi_{RY}(s_1)\phi_{RY}(s_2)\cdots\phi_{RY}(s_n), \tag{1}$$

where $\phi_{RY}(s_i) = \begin{cases} R & if\ s_i \in R \\ Y & if\ s_i \in Y \end{cases}$, i=1,2,…,n

$$\phi_{MK}(X) = \phi_{MK}(s_1)\phi_{MK}(s_2)\cdots\phi_{MK}(s_n), \tag{2}$$

where $\phi_{MK}(s_i) = \begin{cases} M & if\ s_i \in M \\ K & if\ s_i \in K \end{cases}$, i=1,2,…,n

$$\phi_{WS}(X) = \phi_{WS}(s_1)\phi_{WS}(s_2)\cdots\phi_{WS}(s_n), \tag{3}$$

where $\phi_{WS}(s_i) = \begin{cases} W & if\ s_i \in W \\ S & if\ s_i \in S \end{cases}$., i=1,2,…,n.

One example is three symbolic sequences corresponding sequence $X = ATGGTGCACCTGACT$ as follows:

$\phi_{RY}(X) = RYRRYRYRYY\ YRRYY$
$\phi_{MK}(X) = MKKKKKMMMMKKMMK$
$\phi_{WS}(X) = WWSSWSSWSS\ WSWSW$

We call obtained symbolic sequences RY sequence, MK sequence and WS sequence respectively. Then we analyze mutation information of above three symbolic sequences. In each classification we focus on group mutation information. From above symbolic sequences we find twice $R \to R$, five $R \to Y$, four $Y \to R$, and three times $Y \to Y$ group mutations; four $M \to M$, three times $M \to K$, twice $K \to M$ and five $K \to K$ group mutations; once $W \to W$, five $W \to S$, five $S \to W$ and three times $S \to S$ group mutations. It is difficult to obtain the information from DNA primary sequence directly. The frequency of group mutation will be applied to make comparisons among coding sequences belonging to eleven Species in Table 1 for analysis of similarity in next section.

3 Application

3.1 Mutation Analysis

For a symbolic sequence identified with a word over an alphabet {R,Y}, we define the frequency of group mutation as follows:

$$f_{UV} = \frac{the\ \ number\ \ \ of\ \ word\ \ UV}{n-1}, \tag{4}$$

where $U,V \in \{R,Y\}$, n is the length of the symbolic sequence. So we have four frequencies of group mutations in RY classification, denoted by $f_{RR}, f_{RY}, f_{YR}, f_{YY}$ respectively. Similarly, we can compute the frequencies of group mutations in MK classification and WS classification. In Table 2 we list frequencies of group mutations

based on three classifications of the coding sequences of the first exon of beta globin gene of eleven species in Table 1. From Table 2, we conclude:

1) The frequencies of $R \to Y$ group mutations and $Y \to R$ group mutations of all listed species are identical, although the lengths of coding sequences are different.

2) In RY sequences, (human, gallus) and (goat, bovine) have the same frequencies of group mutation respectively.

3) The frequencies of $R \to R$ group mutation and $K \to K$ group mutation are larger than 0.3.

3.2 Similarities and Dissimilarities

We construct a twelve-component vector consisting of twelve frequencies of group mutations to represent a DNA sequence. The analysis of similarity and dissimilarity between two DNA sequences represented by the twelve-component vectors is based on the assumption that two sequences are similar if the corresponding twelve-component vectors point to a similar direction in the 12D space and have similar magnitudes. The similarities between such vectors can be computed by calculating the Eucilidean distance between their end points. The smaller Euclidean distance is, the more similar are the DNA sequences.

In Table 2, we list the similarity/dissimilarity matrix for the coding sequences of Table 1 based on the Euclidean distances between the end points of the twelve-component vectors.

Table 2. Twelve frequencies of group mutation based on three classifications of the coding sequences of Table 1

Species	f_{RR}	f_{RY}	f_{YR}	f_{YY}	f_{MM}	f_{MK}	f_{KM}	f_{KK}	f_{WW}	f_{WS}	f_{SW}	f_{SS}
Human	0.3407	0.2308	0.2308	0.1978	0.1978	0.1978	0.1868	0.4176	0.0989	0.3077	0.2967	0.2967
Goat	0.3882	0.2118	0.2118	0.1882	0.1647	0.2353	0.2235	0.3765	0.1059	0.2941	0.2824	0.3176
Opossum	0.3077	0.2308	0.2308	0.2308	0.2308	0.2198	0.2088	0.3407	0.1319	0.3407	0.3297	0.1978
Gallus	0.3407	0.2308	0.2308	0.1978	0.2418	0.2308	0.2198	0.3077	0.0989	0.2747	0.2637	0.3626
Lemur	0.3626	0.2198	0.2198	0.1978	0.1319	0.2418	0.2308	0.3956	0.1429	0.3187	0.3077	0.2308
Mouse	0.3152	0.2283	0.2283	0.2283	0.1957	0.2065	0.1957	0.4022	0.1087	0.3152	0.3043	0.2717
Rabbit	0.3736	0.2308	0.2308	0.1648	0.1758	0.1978	0.1868	0.4396	0.0989	0.3187	0.3077	0.2747
Rat	0.3297	0.2418	0.2418	0.1868	0.1978	0.2198	0.2088	0.3736	0.1758	0.2747	0.2637	0.2857
Gorilla	0.3478	0.2283	0.2283	0.1957	0.1957	0.1957	0.1848	0.4239	0.0978	0.3043	0.2935	0.3043
Bovine	0.3882	0.2118	0.2118	0.1882	0.1765	0.2118	0.2000	0.4118	0.1294	0.2824	0.2706	0.3176
Chimpanzee	0.3462	0.2308	0.2308	0.1923	0.1923	0.1923	0.1827	0.4327	0.1250	0.2981	0.2885	0.2885

Observing from Table 3, we find the smallest entries are associated with the pairs (Human, Gorilla), (Gorilla, Chimpanzee) and (Human, Chimpanzee). On the other hand the largest entries in the similarity/dissimilarity matrix appear in the row belonging to Opossum (the most remote species from the remaining mammals) and Gallus (the only non-mammalian representative). We believe it is not an accident, but shows their relationship in evolutionary sense.

Table 3. Similarity/dissimilarity matrix for the coding sequences of Table 1 based on the Euclidean distances between the end points of the twelve-component vectors

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.0974	0.1523	0.1507	0.1263	0.0531	0.0622	0.1077	**0.0143**	0.0814	**0.0357**
Goat		0	0.1859	0.1287	0.1121	0.1172	0.1071	0.1159	0.0948	0.0576	0.1059
Opossum			0	0.2014	0.1424	0.1128	0.1744	0.1523	0.1653	0.1960	0.1616
Gallus				0	0.2097	0.1596	0.1941	0.1373	0.1527	0.1483	0.1683
Lemur					0	0.1130	0.1142	0.1263	0.1316	0.1241	0.1215
Mouse						0	0.0976	0.1070	0.0652	0.1129	0.0683
Rabbit							0	0.1355	0.0597	0.0900	0.0595
Rat								0	0.1125	0.1028	0.0962
Gorilla									0	0.0736	**0.0344**
Bovine										0	0.0729

4 Conclusion

Mutation analysis and similarity analysis of DNA sequences are still important subjects in bioinformatics. According to chemical properties of nucleotide bases we divide four bases into two classes in each classification. For a DNA primary sequence we obtain three symbolic sequences and we compute the frequencies of group mutations of each symbolic sequence. From Table 2 we find some useful conclusions about group mutation. The frequencies of group mutations are applied to construct a 12-component vector to represent the sequence. The Euclidean distances among the introduced vectors are applied to compare the similarity and dissimilarity of the coding sequence of the first exon of beta globin gene of 11 different species. The results are coincident with what we expect in the evolutionary sense. Method we introduced is simple and clear so that we can quickly make comparisons of the similarity/dissimilarity among DNA sequences and its computation is easy, even for a long sequence.

Acknowledgment. This research is supported by Scientific Research Fund of Hunan Provincial Education Department (No.09C1015, No.08C882).

References

1. Hamori, E., Ruskin, J., Curves, H.: A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. J. Biol. Chem. 258, 1318–1327 (1983)
2. Gates, M.A.: A Simple way to look at DNA. J. Theor. Biol. 119, 319–328 (1986)
3. Nandy, A.: A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr. Sci. 66, 309–314 (1994)
4. Leong, P.M., Morgenthaler, S.: Random walk and gap plots of DNA sequences. Comput. Appl. Biosci. 11, 503–507 (1995)
5. Guo, X.F., Randic, M., Basak, S.C.: A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem. Phys. Lett. 350, 106–112 (2001)

6. Randic, M., Vrakoc, M., Lers, N., Plsvsic, D.: Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6 (2003)
7. Randic, M., Vrakoc, M., Lers, N., Plsvsic, D.: Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem. Phys. Lett. 371, 202–207 (2003)
8. Wu, Y.H., Liew, A.W., Yan, H., Yang, M.S.: DB-Curve: a novel 2D method of DNA sequence visualization and representation. Chem. Phys. Lett. 367, 170–176 (2003)
9. Liao, B., Wang, T.M.: New 2D graphical representation of DNA Sequences. J. Comput. Chem. 25, 1364–1368 (2004)
10. Liao, B., Wang, T.M.: Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chem. Phys. Lett. 388, 195–200 (2004)
11. Liao, B., Wang, T.M.: 3-D graphical representation of DNA sequences and their numerical characterization. J. Mol. Struct. Theochem. 681, 209–212 (2004)
12. Yao, Y.H., Wang, T.M.: A class of new 2-D graphical representation of DNA sequences and their application. Chem. Phys. Lett. 398, 318–323 (2004)
13. Liao, B., Tang, M.S., Ding, K.Q., Wang, T.M.: Analysis of similarity /dissimilarity of DNA sequences based on a condensed curve representation. J. Mol. Struct. Theochem. 717, 199–203 (2005)
14. Song, J., Tang, H.W.: A new 2-D graphical representation of DNA sequences and their numerical characterization. J. Biochem. Biophys. Methods 63, 228–239 (2005)
15. Li, C., Tang, N.N., Wang, J.: Directed graphs of DNA sequences and their numerical characterization. J. Theor. Biol. 241, 173–177 (2006)
16. Yao, Y.H., Nan, X.Y., Wang, T.M.: A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences. J. Mol. Struct. Theochem. 764, 101–108 (2006)
17. Liao, B., Ding, K.: A 3D graphical representation of DNA sequences and its application. Theor. Comput. Sci. 358, 56–64 (2006)
18. Liu, X.Q., Dai, Q., Xiu, Z.L., Wang, T.M.: PNN-curve: A new 2D graphical representation of DNA sequences and its application. J. Theor. Biol. 243, 555–561 (2006)
19. Qi, Z., Qi, X.: Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chem. Phys. Lett. 440, 139–144 (2007)
20. Qi, Z., Fan, T.: PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 442, 434–440 (2007)