MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **61** (2009) 767-780

ISSN 0340 - 6253

An improved binary representation of DNA sequences and its applications

Weiyang Chen, Bo Liao[1], Xuyu Xiang, Wen Zhu

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

(Received July 14, 2007)

Abstract:

We advanced one kind of binary coding method for DNA sequences. Based on this representation, we can transform a DNA sequence to three unique binary sequences. The introduced system can be applied to characterize and compare the DNA sequences. In our method, the operators $\oplus$ and $\wedge$ are used to judge mutations. Moreover, based on the result of our coding method, we present a 3D graphical representation of DNA sequences. The 3D graphical representation avoids loss of information accompanying alternative 2D and 3D representation in which the curve standing for DNA overlaps and intersects itself. And this graph can reflect the characters of DNA sequence well.

1. Introduction

Molecular biologists are currently engaged in some of the most impressive data collection projects. Recent genome sequencing projects are generating an enormous amount of data related to the function and the structure of biological molecules and sequences. Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists.

Bioinformatics data is mainly expressed by the form of sequence. Mutation

[1] Corresponding author: dragonbw@163.com

analysis is the most important tools of bioinformatics. Nowadays, the most of compare methods are based on the original DNA sequences which are composed of A (adenine), G (guanine), T (thymine), and C (cytosine). For two sequences comparison, there have many methods been used in sequence alignment. But these methods are not easy to measure the mutation between bases. Recently, many authors have presented different graphical representations of DNA sequences [1-17]. These graphical representations are also applied to the sequence alignment [1, 2] and mutation analysis [1, 3].

In the reference [4], we described a binary coding method for DNA sequences by two bit binary digits. Obviously, the coding rule will lengthen the sequence. In this paper, we will use one bit binary digit to represent the four bases A, C, G, and T, respectively. By this method, every DNA sequence also can be transformed to three sequences of binary system. Since the operating results are simple, the operator $\wedge$ is used to carry out some applications with the operator $\oplus$.Based on the system, we can judge base mutations between sequences. Moreover, based on the result of our coding method, we can give the DNA sequence one 3D graphical representation easily. And this graph can reflect the characters of every DNA sequence commendably.

2. Binary coding method for DNA sequence

Analysis and comparison of DNA sequences should consider not only the structures of strings but also their chemical structures. In a DNA primary sequences, the four bases A, C, G and T can be classed into three groups [4, 5], purine {A, G}/pyrimidine {C, T}, amino {A, C}/keto {G, T}, and weak-H bond {A, T} /strong-H band {C, G}. In the following, we will outline a new binary coding method of DNA sequences according to the three classifications of bases.

We will use the exclusive-OR operator and the and-operator. The exclusive-OR of x_1 and x_2 written $x1 \oplus x_2$ is defined by Table 1. The and-operator of x_1 and x_2 written $x_1 \wedge x_2$ is defined by Table 2 in the following.

Table1: The exclusive-OR

X_1	X_2	$X_1 \oplus X_2$
1	1	0
0	1	1
1	0	1
0	0	0

Table2: The and-operator

X_1	X_2	$X_1 \wedge X_2$
1	1	1
0	1	0
1	0	0
0	0	0

We will use one bit binary digit to represent the four bases A, C, G, and T, respectively. For the coding DNA sequence, the operating rules are defined as above. For one DNA sequence, there are three coding sequences corresponding to the three classifications of bases.

(i) Corresponding to the first classification: purine {A, G}/pyrimidine {C, T}, we define the first coding rule satisfied $A \oplus G=0$, $C \oplus T=0$.

A: 1, G: 1, C: 0, T: 0

(ii) Corresponding to the second classification: amino {A, C}/keto {G, T}, we define the second coding rule satisfied $A \oplus C=0$, $G \oplus T=0$.

A: 1, C: 1, G: 0, T: 0

(ii) Corresponding to the third classification: weak-H bond {A, T}/strong-H bond {G, C}, we define the third coding rule satisfied $A \oplus T=0$, $C \oplus G=0$.

A: 1, T: 1, C: 0, G: 0

For example, by our coding rules, the DNA sequence ACGT will be reduced into {1010, 1100, and 1001}. The three coding results are based on the three coding rules respectively.

3. Mutation analysis

We will judge base mutations between sequences by the binary coding. We will give the method and process of judge base mutations. For two DNA sequences S1 and

S2, S11 and S21 are their coding sequences corresponding to the first coding rule (i), S12 and S22 are their coding sequences corresponding to the second classification (ii), S13 and S23 are their coding sequences corresponding to the third classification (iii).

(1) For the first kind of coding sequence, we do the exclusive-OR operation to the every bit of S11 and S21, and we can get a binary sequence L1 which is the result of S11 ⊕ S21. For example, S11=$a_1a_2\cdots a_n$, S21=$b_1b_2\cdots b_n$. After S11 ⊕ S21={ $a_i \oplus b_i$ | where $1 \leqslant i \leqslant n$} we can get L1=$c_1c_2\cdots c_n$, where $c_i=a_i \oplus b_i$.

(2) For the second kind of coding sequence, we do the exclusive-OR operation to the every bit of S12 and S22, and we can get a binary sequence L2 that is the result of S12 ⊕ S22.

(3) For the third kind of coding sequence, we do the exclusive-OR operation to the every bit of S13 and S23, and we can get a binary sequence L3 which is the result of S13 ⊕ S23.

These three resulting sequences are shown in table 3.

Table3: The three resulting sequence after exclusive-OR operation

S11	S21	S12	S22	S13	S23	L1	L2	L3
a_1	b_1	d_1	e_1	f_1	h_1	$a_1 \oplus b_1$	$d_1 \oplus e_1$	$f_1 \oplus h_1$
a_2	b_2	d_2	e_2	f_2	h_2	$a_2 \oplus b_2$	$d_2 \oplus e_2$	$f_2 \oplus h_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a_n	b_n	d_n	e_n	f_n	h_n	$a_n \oplus b_n$	$d_n \oplus e_n$	$f_n \oplus h_n$

Then by these results we will do mutation analysis between the two sequences S1 and S2. When three resulting sequences L1, L2, L3 are all 0, these segments are the similar regions of two DNA sequences. That is to say, there are not mutations in these corresponding places of two DNA sequences.

Besides these similar regions, other regions are the mutation places. And by every coding sequence we can judge the different mutations as follows.

(1) If the first resulting sequence L1 are 0 while other two resulting sequences are 1, then the mutations take place between purine and purine or between pyrimidine and pyrimidine. To distinguish the mutations

accurately, we need the and-operator. We do the ∧ operation in S11 and S21, the result is L1'. For example, let $S11=a_1a_2\cdots a_n$, $S21=b_1b_2\cdots b_n$. After $S11 \wedge S21$, we can get $L1'=g_1g_2\cdots g_n$, where $g_i=a_i \wedge b_i$. Then we can distinguish whether the mutation is from purine to purine or from pyrimidine to pyrimidine. In the mutation regions, if the L1' is 1 then the mutation takes place between A and G, if the L1' is 0 then the mutation takes place between C and T.

(2) If the second resulting sequence are 0 while other two resulting sequences are not 0, then the mutations take place between amino and amino or between keto and keto. To distinguish the mutations accurately, we do the ∧ operation in S12 and S22, the result is L2'. In the mutation regions, if the L2' is 1 then the mutation takes place between A and C, if the L2' is 0 then the mutation takes place between G and T.

(3) If the third resulting sequence are 0 while other two resulting sequences are not 0, the mutations take place between weak-H bond and weak-H bond or between strong-H band and strong-H band. We also do the ∧ operation in S13 and S23, the result is L3'. In the mutation regions, if the L3' is 1 then the mutation takes place between A and T, if the L3' is 0 then the mutation takes place between G and C.

For example, there are two sequences:

S1: AAAAAACCGGGGAGCT

S2: GGCCTTTTTTCCAGCT

The binary coding of these sequences are as follows:

S11: 1111110011111100

S21: 1100000000001100

S12: 1111111100001010

S22: 0011000000111010

S13: 1111110000001001

S23: 0000111111001001

Then we will get the sequences L1, L2, L3 by the ⊕ operation, and the sequences L1', L2', L3' after the ∧ operation.

```
L1: 00 11 11 00 11 11 0000
L2: 11 00 11 11 00 11 0000
L3: 11 11 00 11 11 00 0000
L1': 11 00 00 00 00 00 1100
L2': 00 11 00 00 00 00 1010
L3': 00 00 11 00 00 00 1001
     1° 2° 3° 4° 5° 6° 7°
```

Observing the result, we can find some interesting results.

In the segment 7°, L1, L2, L3 are uniform, they are all 0. So there are the similar regions of two DNA sequences.

For the segment 1°, in the sequence L1 there are two 0 and in sequence L2, L3 they are not 0. So there are mutations. Because in the sequence L1' there are 1, so mutations take place between purine and purine.

For the segment 2°, in the sequence L2 there are two 0 and in sequence L1, L3 they are not 0. What's more, in the sequence L2' there are 1, so mutations occur between amino and amino.

For the segment 3°, in the sequence L3 there are two 0 and in sequence L1, L2 they are not 0. And in the sequence L3' there are 1, so mutations should be arise from weak-H bond to weak-H bond.

Using the similar method, we can judge the mutations should be take place between pyrimidine and pyrimidine in the segment 4°, between keto and keto in the segment 5°, between strong-H band and strong-H band in the segment 6°.

4. Graphical representation of DNA sequences and Similarity analysis

4.1 Graphical representation

According to our coding rules, we can define three-dimensional coordinate for the four bases respectively by combine their coding. The four bases are defined as follows.

A: (1, 1, 1), G: (1, 0, 0), C: (0, 1, 0), T: (0, 0, 1).

We assign A and G to +x, A and C to +y, A and T to +z. In detail, let $L=g_1g_2\cdots g_n$ be an arbitrary DNA primary sequence. Then we have a map Ø, which maps L into a plot set. Explicitly, $Ø(L)= Ø(g_1)\, Ø(g_2)\cdots\ Ø(g_n)$, where

$$\phi(g_i) = (A_i + G_i, A_i + C_i, A_i + T_i)$$

A_i, C_i, G_i, T_i are the cumulative occurrence numbers of A, C, G and T, respectively.

The corresponding plot set is called characteristic plot set. The curve connecting all plots of the characteristic plot set in turn is called a characteristic curve. This characteristic curve strictly displays the distributions of bases of different classifications in the corresponding DNA sequence.

For example, we consider the first ten bases of the first exon of human β-globin gene, the corresponding plot set of sequence ATGGTGCACC is {(1,1,1),(1,1,2),(2,1,2),(3,1,2),(3,1,3),
(4,1,3),(4,2,3),(5,3,4),(5,4,4),(5,5,4)}. The curve of this sequence is shown in fig.1.

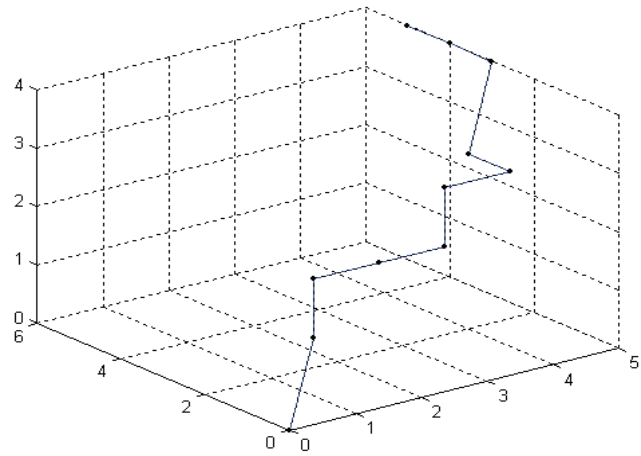

Fig.1. The corresponding curve of the sequence ATGGTGCACC.

Obviously, our graphical representation avoids the degeneracy totally. In our graphical representation, we considered the strings' structures and their chemical structures.

Theorem: (i) Our graph can reflect the characters of every DNA sequence commendably. That is to say, the coordinate can reflect the distribution of purine and pyrimidine, amino and keto, weak-H bond and strong-H bond.

(ii) Let the i-th base of a DNA sequence corresponds coordinate (x_i, y_i, z_i), then the value of pyrimidine, keto and strong-H bond is i- x_i , i- y_i,,i- z_i , respectively, where i is its length of the subsequence from the first base to the i-th base.

Proof. Suppose the cumulative numbers of A, C, G, and T are A_i, C_i, G_i, and T_i respectively. We can obtain the following equations based on our representation.

$x_i = A_i + G_i$; $y_i = A_i + C_i$; $z_i = A_i + T_i$ (1)

$A_i + G_i + C_i + T_i = i$, (2)

Combined the equation (1) with equation (2), we can get $A_i = (x_i + y_i + z_i - i)/2$.

So $G_i = x_i - A_i = x_i - (x_i + y_i + z_i - i)/2$; $C_i = y_i - A_i = y_i - (x_i + y_i + z_i - i)/2$; $T_i = z_i - A_i = z_i - (x_i + y_i + z_i - i)/2$

then (i) proved, (ii) is obvious.

In figure 2 , we show the graphical representation of the the first exon of gorilla β-globin gene. Its terminal coordinate is (54, 36, 37). So its content of purine is 54, the content of amino is 36, and the content of weak-H bond is 37.

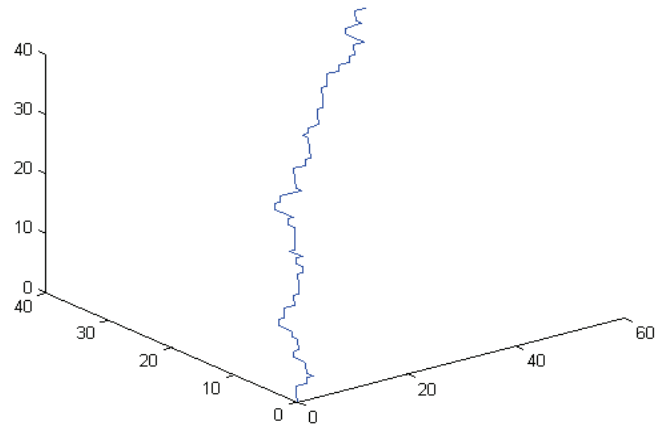

Fig.2. The 3D graphical representation of the first exon of gorilla β-globin gene.

4.2 Similarity analysis

In the following, we will make comparisons of similarities and dissimilarities for eleven exon-1 genes. We choose the commonly used coding sequences of the first exon of β-globin gene of eleven species (Human, Goat, Opossum, Gallus, Lemur, Mouse, Rabbit, Rat, Gorilla, Chimpanzee, Bovine), which can be found in the references [3, 11, 12, 18, 19].

For any sequence, we have a set of points (x_i, y_i, z_i), $i = 1, 2, 3 \ldots n$, where n is the length of the sequence. Similar to Nandy's index scheme [20], the coordinates of the geometrical center of the points, denoted by x_0, y_0 and z_0, may be calculated as follows:

$$x_0 = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ \mathrm{y}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \ \mathrm{z}_0 = \frac{1}{n}\sum_{i=1}^{n} z_i \ .$$

In Table 4 we present the geometrical centers of the first exon of β-globin gene belonging to 11 species.

Table 4. The geometrical centers of the first exon of β-globin gene belonging to 11 species

human	(24.826086,19.739130,20.043478)
goat	(25.581396,18.034883,17.906977)
opossum	(24.880434,21.543478,23.728260)
mouse	(23.957447,19.170214,22.244680)
gallus	(26.858696,21.358696,18.717392)
lemur	(25.543478,17.815218,23.119566)
rabbit	(26.211111,17.522223,20.011110)
rat	(26.413044,19.717392,22.608696)
gorilla	(25.419355,19.913979,19.946236)
chimpanzee	(29.028572,21.933332,22.380953)
bovine	(25.162790,17.895350,18.465117)

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 x_0(i), y_0 (i), z_0 (i) and x_0(j), y_0 (j), z_0 (j). We will illustrate the use of the 3D quantitative characterization of DNA sequences with an examination of similarities/dissimilarities among the 11 coding sequences.

The cosine value formula is shown as follows:

$$\cos\theta_{ij} = \frac{x_0(i)x_0(j) + y_0(i)y_0(j) + z_0(i)z_0(j)}{\sqrt{(x_0(i))^2 + (y_0(i))^2 + (z_0(i))^2}\ \sqrt{(x_0(j))^2 + (y_0(j))^2 + (z_0(j))^2}}$$

All the cosine values between any two species are shown in the table 5.

Obviously, the smaller the correlation angle is, the more similar the DNA sequences are. That is to say, the bigger cosine value is, the more similar the DNA sequences are.

Observing Tables 5, we find gallus is very dissimilar to others among the 11 species because its corresponding row has smaller entries. And the more similar species pairs are human–gorilla, human–chimpanzee, goat–bovine, opossum–mouse, mouse–rat, rabbit–bovine, gorilla–chimpanzee, and bovine–chimpanzee. The similar results can be found in references [3, 11, 12, 18, 19].

Table 5. The similarity/dissimilarity matrix for the 11 coding sequences based on the cosine value of angle between the three-component vectors of the geometrical centers.

Species:	Human	goat	opossum	mouse	gallus	lemur	rabbit	rat	gorilla	chimpanzee	bovine
human	1.000000	0.997816	0.997533	0.997968	0.997937	0.995852	0.997572	0.998968	0.999929	0.999708	0.998657
goat		1.000000	0.991005	0.992916	0.998523	0.992926	0.998722	0.996412	0.998493	0.999092	0.999808
opossum			1.000000	0.999502	0.991534	0.996341	0.992963	0.997974	0.996626	0.995793	0.993161
mouse				1.000000	0.991821	0.998528	0.995593	0.999283	0.997230	0.996861	0.994986
gallus					1.000000	0.988928	0.995285	0.994477	0.998523	0.998412	0.998132
lemur						1.000000	0.997200	0.998957	0.995313	0.995712	0.995010
rabbit							1.000000	0.998423	0.997891	0.998701	0.999344
rat								1.000000	0.998664	0.998752	0.997874
gorilla									1.000000	0.999887	0.999112
chimpanzee										1.000000	0.999617
bovine											1.000000

5. Conclusions

In this paper, we introduced a sort of binary coding method of DNA sequences. By this method, every DNA sequence can be transformed to binary sequences. Based on the system, we can judge base mutations between sequences. And, based on the result of our coding method, we can give the DNA sequence a 3D graphical representation easily. This kind of graphical representation avoids the problem of degeneracy totally. And our graph can reflect the characters of every DNA sequence commendably. Based on the graph, we can do the analysis of similarities. It is helping in recognizing major similarities among different DNA sequences.

6. Acknowledgement

This work is supported in part by the National Nature Science Foundation of China(Grant 10571019) and the National Nature Science Foundation of Hunan province(Grant 07JJ5080).

References

[1] B. Liao, K. Q. Ding, Graphical approach to analyzing DNA sequences**,** J. Comput. Chem. 26 (2005) 1519-1523.

[2] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences, Chem. Phys. Lett. 431 (2006) 375-379.

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

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

[5] Y.H. Yao, X.Y. ,Nan, T.M. Wang, A new 2D graphical representation – Classification curve and the analysis of similarity/dissimilarity of DNA sequences. J. Mol. Struct. (Theochem) 764 (2006) 101–108.

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

[7] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, Curr. Sci. 66 (1994) 309–314.

[8] A. Nandy, P. Nandy, Graphical analysis of DNA sequences structure: II. Relative abundance of nucleotides in DNAs, gene evolution and duplication, Curr. Sci. 68 (1995) 75–85.

[9] 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.

[10] G. Huang, B. Liao, W. Zhang, F. Gong. A novel method for sequence alignment and mutation analysis. MATCH Commun. Math. Comput. Chem. 59 (2008) 635-645.

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

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

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

[14] M. Randić, Graphical representations of DNA as 2-D map, Chem. Phys. Lett. 386 (2004) 468-471.

[15] M. Randić, M. Vračko, J. Zupan, M. Nović, Compact 2-D graphical representation of DNA, Chem. Phys. Lett. 373 (2003) 558-562.

[16] X.Q. Liu, Q. Dai, Z.L. Xiu, T.M. Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its application, J. Theor. Biol. 243 (2006) 555–561.

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

[18] Chi, R., Ding, K.Q., Novel 4D numerical representation of DNA sequences. Chem. Phys. Lett. 407 (2005) 63-67.

[19] Q. Dai, X. Liu, T. Wang, A novel 2D graphical representation of DNA sequences and its application. J. Mol. Graph. Modell. 25 (2006) 340-344.

[20] C. Raychaudhury, A. Nandy, Indexing scheme and similarity measure for macromolecular sequences, J. Chem. Inf. Comput. Sci. 39 (1999) 243–247.