MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **59** (2008) 635-645

ISSN 0340 - 6253

A Novel Method for Sequence Alignment and Mutation Analysis

Guohua Huang, Bo Liao[1], Wu Zhang, Fei Gong

School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China

(Received September 25, 2007)

Abstract

We introduced a representation of DNA sequences and outlined an approach to search optimal alignments and to judge mutations based on the presented representation.

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. Other complementary high-throughput technologies, such as DNA microarrays, are rapidly generating large amounts of data that are too overwhelming for conventional approaches to biological data analysis[1]. Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists. Graphical representation of DNA sequences provides a single way of sorting and comparing various gene sequences, initiated by Hamori and Ruskin[2,3], and by Gates[4,5]. In recent years, many authors have present different representations of DNA sequences based on 2-D,3-D, even higher dimension space [6-21] .Most representations are applied to these cases as follows: (i) similarity analysis . The invariants of D/D matrix[9,15,19], M/M matrix[10], L/L matrix [7,10,14] , covariance matrix, even 'higher order' matrices [7,9-10,15] (such as eigenvalues[10,14], leading eigenvalue[7,9,15] , or average band width[14,16,18]), are used to compute the similarity between sequences. (ii) sequence alignment. Recently, Liao[20] and Randić [21] utilize the representation of sequences to make alignment between two DNA sequences that is one of the crucial sequence-comparison operation in the bioinformatics and genetic research . (iii) mutation analysis. The characteristic vectors are used to measure the mutation between bases in the representations [6, 20].

In the paper, we propose a representation of DNA sequences, and introduce an

[1] Corresponding author: dragonbw@163.com

approach to search optimal alignment based on the presented representation .Our alignment algorithm improves the efficiency of alignment [21], and predigest the difficulty of alignment [20]. Based on our proposed representation, we also can judge base mutations between sequences.

Representation of DNA sequences

Recently, Milan Randić [21] put forward a representation of DNA sequence that the four base types of DNA sequences are assigned to four horizontal lines separated by a unit distance interval in a plane. The bases consisting of DNA sequences are represented by dots along the horizontal lines. Associating adjacent dots with short lines, one can obtain a zigzag curve. Here, we will present a universal representation of DNA sequence.

Let sequence $a = a_1 a_2 \cdots a_n, a_i \in \{A, C, G, T, -\}, 1 \le i \le n$, we define a map ρ as follows:

$$\rho(a_i) = \begin{cases} (i, h_1) & a_i = A \\ (i, h_2) & a_i = T \\ (i, h_3) & a_i = G \\ (i, h_4) & a_i = C \\ (i, \infty) & a_i = '\_' \end{cases} \quad (1)$$

where n is the length of the studied sequence, h_i (i=1,2,3 and 4) are integers and are not equal with the others. For example ,we can denote h_1(h_2, h_3 and h_4) as 1(2,4 and 8)respectively, -1(0,2 and 6) respectively or 0(1,3 and 7)respectively ,and so on . In order to express the result of alignment, we suppose that the character '-' is defined as (i, ∞) which corresponds to an empty dot. Obviously, each base of sequence corresponds to a dot in Cartesian coordinate system in the plane. For example, let $h_1 = 1, h_2 = 2, h_3 = 4, h_4 = 8$, then the sequence g=ATGGCATTGACAAACTCG will be represented as an ordered array {(1,1), (2,2), (3,4), (4,4), (5,8), (6,1), (7,2), (8,2), (9,4), (10,1), (11,8), (12,1), (13,1), (14,1), (15,8), (16,2), (17,8), (18,4)}. The array is shown in Fig.1.

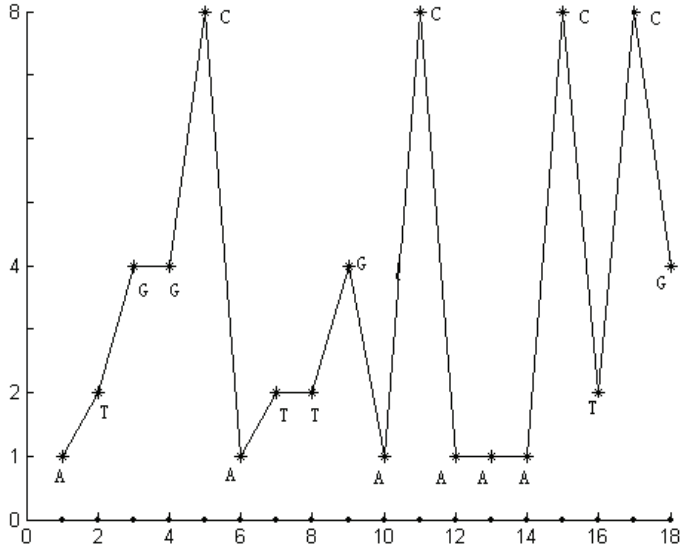

Fig .1.The representation of sequence ATGGCATTGACAAACTCCG

There is a distinct characteristic as follows: the corresponding representation is simple for a DNA sequence. In addition, we only can reconstruct a unique DNA sequence. There is no loss of information in the process of transforming from sequence to representation or from representation to sequence. The representation of sequence can viewed as a "signature" of the sequence.

Suppose two arbitrary sequences $a=a_1a_2\cdots a_n$ and $b=b_1b_2\cdots b_m$, where n and m are the lengths of sequences a and b, respectively. We can obtain two Theorems as flows:

Theorem1 If $\rho(a_i)-\rho(b_i)=(0,0)$, where $1\le i\le \min(n,m)$,then a_i should match b_i.

Proof. Let $\rho(a_i)=(i,\lambda),\rho(b_i)=(i,\mu),\lambda,\mu\in\{h_1,h_2,h_3,h_4\}$,if a_i doesn't match b_i ,then $\lambda\neq\mu$, so there is a contradiction with the equation $\rho(a_i)-\rho(b_i)=(0,0)$ .So the above proposition is true.

Theorem2 If the representation of the sequence $a=a_1a_2\cdots a_n$ shifts s(a positive integer) units paralleling to the x-axis, then we can obtain a new sequence $c=c_{1+s}c_{2+s}\cdots c_{n+s}$ and it satisfied $\rho(c_{i+s})-\rho(a_i)=(s,0)$,or $\rho(c_{i+s})=(s,0)+\rho(a_i)$, for any i, $1\le i\le n$.

Proof. Suppose the arbitrary base a_i corresponds to the dot (i,λ), $1\le i\le n$,if the dot

(i,λ) shifts s units paralleling to the x-axes, then it should arrive at the dot $(i+s,\lambda)$ corresponding to (i+s)th base of the new sequence, and there should be a equation $\rho(c_{i+s}) = (s,0) + \rho(a_i)$.

Corollary1 For any i, $1 \le i \le \min\{m,n\}$,if $\rho(b_i) - \rho(a_i) = (0,0)$, then sequence a is the subsequence of sequence b or sequence b is the subsequence of sequence a.

Corollary2 Suppose the representation of the sequence b moves s units paralleling to the x-axis to be transformed into a new representation corresponding to a new sequence c= $c_{1+s}c_{2+s}\cdots c_{m+s}$, where $c_i=b_{i-s}$, $1+s \le i \le m+s$.

(i) If $\rho(c_i) - \rho(a_i) = (0,0)$, where $\max(1,s+1) \le i \le \min(n,s+m)$, then a_i should match b_{i-s}.

(ii) If $\rho(c_i) - \rho(a_i) = (0,0)$, for any i, $\max(1,s+1) \le d1 \le i \le d2 \le \min(n,s+m)$,then the subsequence $a_{d1}a_{d1+1}\cdots a_{d2}$ of the sequence a should match the subsequence $b_{d1-s}b_{d1+1-s}\cdots b_{d2-s}$ of the sequence b.

Corollary3 If the subsequence $a_{i+1}\cdots a_{i+d}$ would match the subsequence $b_{j+1}\cdots b_{j+d}$, and if there would not exist s (an integer)to make the new sequence c= $c_{1+s}c_{2+s}\cdots c_{m+s}$ satisfied that $\rho(c_k) - \rho(a_k) = (0,0)$,for any k, $\max(1,s+1) \le d1 \le k \le d2 \le \min(n,s+m)$,and $d2 - d1 > d$, where $c_i=b_{i-s}$, $1+s \le i \le m+s$, then the sequences a and b should have the longest common (matching) subsequence $a_{i+1}\ldots a_{i+d}$ (or $b_{j+1}\ldots b_{j+d}$).

Alignment of DNA sequences

According to Corollary1, Corollary2 and Corollary3, we can draw a conclusion that both arbitrary sequences *a* and *b* have the longest common subsequence. We can obtain the longest common subsequence by moving the representation of the sequence *b* along the horizontal direction for the sequence *a*. That is to say, for arbitrary sequences $a=a_1a_2\cdots a_n$ and $b= b_1b_2\cdots b_m$, where n and m are the lengths of sequences *a* and *b* respectively, we should obtain a new sequence $c = c_{1+s}c_{2+s}\cdots c_{m+s}$ by moving sequence *b* which satisfies $\rho(c_i) - \rho(a_i) = (0,0)$, for any *i*,

$\max(1,s+1) \le d1 \le i \le d2 \le \min(n,s+m)$. Then sequences a and b should have the longest common subsequence. We suppose that the longest common subsequence is $a_{d1}a_{d1+1}\ldots a_{d2}$(or $b_{d1-s}b_{d1+1-s}\ldots b_{d2-s}$). So that sequence $a=a_1a_2\ldots a_{d1}a_{d1+1}\ldots a_{d2}a_{d2+1}\ldots a_n$, $b=b_1b_2\ldots b_{d1-s}b_{d1+1-s}\ldots b_{d2-s}b_{d2+1-s}\ldots b_m$, where $a_j =b_{j-s}, d_1 \le j \le d_2$.

Next, we introduce an approach to search optimal alignment based on the

proposed representation of DNA sequences. Assume sequence $a=a_1a_2\cdots a_n$ and sequence $b=b_1b_2\cdots b_m$, where n and m are the lengths of sequences a and b respectively. First, according to Corollary1, Corollary2 and Corollary3, we can obtain the longest common subsequence t by moving the representation of the sequence b along the horizontal direction, right and left. Let $t=a_{i+1}\cdots a_{i+d}=b_{j+1}\cdots b_{j+d}$. So the sequence a is generally divided into three parts $A1=a_1\ldots a_i$, $A2=a_{i+1}\ldots a_{i+d}$ and $A3=a_{i+d+1}\ldots a_n$, and the sequence b corresponds to the next three parts $B1=b_1\ldots b_j$,$B2=b_{j+1}\ldots b_{j+d}$ and $B3=b_{j+d+1}\ldots b_m$. The subsequence A1 will align with the subsequence B1 and the subsequence A3 with the subsequence B3 in the following. Second, in the same way we can obtain the two longest common subsequences between the representation of A1 and the representation of B1, and between the representation of A3 and the representation of B3. Don't terminate this process until the two corresponding short subsequences match completely or don't match at all. In the result, we obtain the optimal alignment.

For example, let sequence a= CTCTACTTGGAAACGACATC and sequence b= CTGATGCTTGGAATTCAT. The corresponding representations are $\{(1,h_4),(2,h_2),(3,h_4),(4,h_2),(5,h_1),(6,h_4),(7,h_2),(8,h_2),(9,h_3),(10,h_3),(11,h_1),(12,h_1),(13,h_1),(14,h_4),(15,h_3),(16,h_1),(17,h_4),(18,h_1),(19,h_2),(20,h_4)\}$ and $\{(1,h_4),(2,h_2),(3,h_3),(4,h_1),(5,h_2),(6,h_3),(7,h_4),(8,h_2),(9,h_2),(10,h_3),(11,h_3),(12,h_1),(13,h_1),(14,h_2),(15,h_2),(16,h_4),(17,h_1),(18,h_2)\}$,respectively. We depict their corresponding curves respectively in Fig.2.

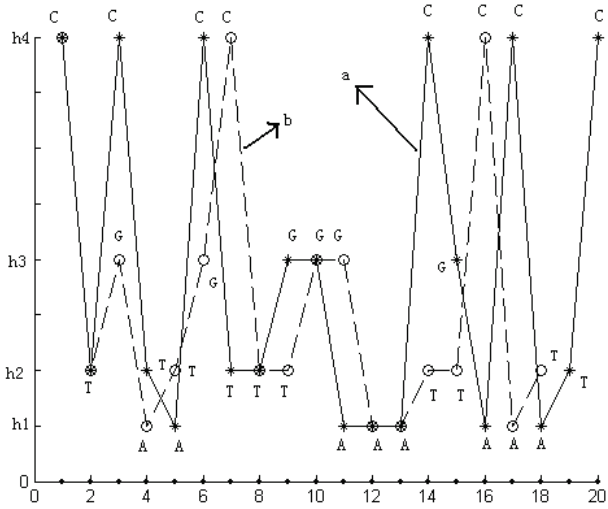

Fig.2. The corresponding curves of a and b

First of all, moving the representation of the sequence b along the horizontal direction, right and left ,we can obtain the longest common representation which corresponds to subsequence S=CTTGGAA, so that the sequence $a=a_1a_2a_3a_4a_5Sa_{13}a_{14}a_{15}a_{16}a_{17}a_{18}a_{19}a_{20}$ and the sequence b= $b_1b_2b_3b_4b_5b_6Sb_{14}b_{15}b_{16}b_{17}b_{18}$,

where $a_1a_2a_3a_4a_5$ =CTCTA $a_{13}a_{14}a_{15}a_{16}a_{17}a_{18}a_{19}a_{20}$=ACGACATC, $b_1b_2b_3b_4b_5b_6$=CTGATG and $b_{14}b_{15}b_{16}b_{17}b_{18}$=TTCAT. Next, in the same way , we can obtain the longest common representation of $a_1a_2a_3a_4a_5$ and $b_1b_2b_3b_4b_5b_6$ by moving the corresponding representation of $b_1b_2b_3b_4b_5b_6$ along the horizontal direction, and the longest common representation corresponds to the subsequence S1=CT , and $a_1a_2a_3a_4a_5$ =S1$a_3a_4a_5$,$b_1b_2b_3b_4b_5b_6$=S1$b_3b_4b_5b_6$. Also we can obtain the longest common subsequence S2 of $b_{14}b_{15}b_{16}b_{17}b_{18}$ and $a_{13}a_{14}a_{15}a_{16}a_{17}a_{18}a_{19}a_{20}$, where S2=CAT, $a_{13}a_{14}a_{15}a_{16}a_{17}a_{18}a_{19}a_{20}$= $a_{13}a_{14}a_{15}a_{16}$S2 a_{20}, and $b_{14}b_{15}b_{16}b_{17}b_{18}$= $b_{14}b_{15}$S2 . Obviously, $a_{13}a_{14}a_{15}a_{16}$ does not match with $b_{14}b_{15}$ at all, while for a_{20} there does not exist alignment. Moving the representation of $b_3b_4b_5b_6$, we can obtain the longest common representation S3 of $a_3a_4a_5$ and $b_3b_4b_5b_6$, where S3=T ,$a_3a_4a_5$ =a_3S3a_5, and $b_3b_4b_5b_6$= b_3 b_4S3b_6. However, b_3 b_4 does not match a_3, and b_6 doesn't match a_5. In the result, a=S1a_3S3a_5S$a_{13}a_{14}a_{15}a_{16}$S2a_{20}, ,b= S1b_3b_4S3b_6S $b_{14}b_{15}$S2.Obviously, one of b_3 and b_4 should correspond to an insertion with respect to the sequence a, and the bases of a_{13},a_{14},a_{15}, a_{16} ,a_{20} should correspond to deletions with respect to the sequence b. There are $C_2^1 \times C_4^2$ different alignments in all. In Table 1, we list six different alignments.

(i)

Sequence a	C	T	-	C	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	T	T	-	-	C	A	T	-

(ii)

Sequence a	C	T	C	-	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	T	T	-	-	C	A	T	-

(iii)

Sequence a	C	T	C	-	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	T	-	T	-	C	A	T	-

(iv)

Sequence a	C	T	C	-	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	T	-	-	T	C	A	T	-

(v)

Sequence a	C	T	C	-	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	-	-	T	T	C	A	T	-

（vi）

Sequence a	C	T	-	C	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	-	T	T	-	C	A	T	-

(vii)

Sequence a	-	C	T	C	T	A	C	T	T	G	G	A	A	A	C	G	A	C	A	T	C
Sequence b	C	T	G	A	T	G	C	T	T	G	G	A	A	T	T	C	A	T	-	-	-

Table 1. (i)-(vi) :the result of alignment by our approach .(vii): the result of alignment using programming Clustalx(1.83).

For the two same sequences a and b, we use a tool (a aligning programming that is called Clustalx(1.83)) to obtain the alignment shown in Table 1.(vii) . In order to compare two methods each other, we define a function of score as follows:

$$\phi(sqe1_i, sqe2_i) = \begin{cases} 1 & sqe1_i = '\_' or sqe2_i = '\_' \\ 2 & sqe1_i = sqe2_i \\ -1 & sqe1_i \neq sqe2_i \end{cases} \quad (2)$$

where $sqe1_i$ and $sqe2_i$ denote the ith base in seq1 and seq2, respectively. The total score of alignment between the two sequences are computed by the following equation:

$$\text{Score} = \sum \phi(sqe1_i, sqe2_i) \quad (3)$$

The total score of alignment by using the aligning programming is 22, while the score is 26 by using our method.

Mutation analysis based on the presented representation

Biological mutations can be generally classified into four types: substitutions, transpositions, insertions and deletions. The first two types of mutations result in errors and the last two types of mutations change the lengths of input DNA (RNA) sequences [22]. Liao [20] proposed an approach of mutation analysis based on the graphical representation, where a sequence is depicted as three different curves in the three patterns {A,T},{A,G} and {A,C} respectively. It can help judge the mutations but the computation is complicated.

Next, we introduce an approach to judge the mutations between sequences based on the new representation. In the following, we discuss mutations in the case of only considering substitutions, insertions and deletions between corresponding positions in sequences a and b respectively. We suppose that $h_2 - h_1 \neq h_1 - h_2 \neq h_3 - h_1 \neq h_1 - h_3 \neq h_4 - h_1 \neq h_1 - h_4 \neq h_2 - h_3 \neq h_3 - h_2 \neq h_2 - h_4 \neq h_4 - h_2 \neq h_4 - h_3 \neq h_3 - h_4$, and let sequence $a=a_1a_2\cdots a_n$, $b=b_1b_2\cdots b_m$, where n and m are lengths of the sequences a and b ,respectively. Obviously, there are some conclusions as follows:

Theorem3 (i) If $\rho(b_i) - \rho(a_i) = (0, h_2 - h_1)$, then the ith base A of the sequence a should be replaced by T;

(ii) If $\rho(b_i) - \rho(a_i) = (0, h_1 - h_2)$, then the ith base T of the sequence a should be replaced by A;

Theorem4 (i) If $\rho(b_i) - \rho(a_i) = (0, h_3 - h_2)$,then the ith base T of the sequence a should be replaced by G;

(ii) If $\rho(b_i) - \rho(a_i) = (0, h_2 - h_3)$, then the ith base G of the sequence a should be replaced by T;

Theorem5 (i) If $\rho(b_i) - \rho(a_i) = (0, h_3 - h_1)$, then the ith base A of the sequence a

should be substituted by G;

(ii) If $\rho(b_i)-\rho(a_i)=(0,h_1-h_3)$, then the ith base G of the sequence a should be substituted by A;

Theorem6 (i) If $\rho(b_i)-\rho(a_i)=(0,h_4-h_3)$, then the ith base G of the sequence a should be substituted by C;

(ii) If $\rho(b_i)-\rho(a_i)=(0,h_3-h_4)$, then the ith base C of the sequence a should be substituted by G;

Theorem7 (i) If $\rho(b_i)-\rho(a_i)=(0,h_4-h_2)$,then the ith base T of the sequence a should be substituted by C;

(ii)If $\rho(b_i)-\rho(a_i)=(0,h_2-h_4)$,then the ith base C of the sequence a should be substituted by T;

Theorem8 (i) If $\rho(b_i)-\rho(a_i)=(0,h_4-h_1)$,then the ith base A of the sequence a should be substituted by C;

(ii)If $\rho(b_i)-\rho(a_i)=(0,h_1-h_4)$, then the ith base C of the sequence a should be substituted by A;

Theorem9 If $\rho(b_i)-\rho(a_i)=(0,\infty)$,then the ith base of the sequence a should be insertion or deletion .

(i) If $\rho(a_i)=(i,h_1)$ ($(i,h_2),(i,h_3)$ or (i,h_4)), then the corresponding base A(T,G or C)should be deleted.

(ii) If $\rho(b_i)=(i,h_1)$ ($(i,h_2),(i,h_3)$ or (i,h_4)),then the corresponding base A (T,G or C) should be inserted.

Proof of Theorem3 : we suppose $a_i\neq A,T,a_i\in\{A,C,G,T\},b_i\neq A,T,b_i\in\{A,C,G,T\}$, then $a_i=G$ or C, $b_i=C$ *or* G , $\rho(b_i)-\rho(a_i)\neq(0,h_2-h_1)$ and $\rho(b_i)-\rho(a_i)\neq(0,h_1-h_2)$,obviously, there is a contradiction with the condition $\rho(b_i)-\rho(a_i)=(0,h_2-h_1)$ or $\rho(b_i)-\rho(a_i)=(0,h_1-h_2)$, so the conclusion is true.

Using a similar method, we can prove the other theorems. Based on the above mentioned theorems, we can quickly judge mutations by moving different vectors (such as $(0,h_2-h_1)$, $(0,h_1-h_2),(0,h_2-h_4)$ and $(0,h_2-h_3)$ and $(0,h_4-h_3)$ and so on) along the horizontal line between two representations of the sequences.

For example, let sequences a = GTTCGACGGT and b = GCTCAATAT, and we suppose $h_1 = 1, h_2 = 2$, $h_3 = 4, h_4 = 8$. The corresponding curves are shown in Fig.3. Moving the vector (0,6) on the two representations, we can conclude that T at the 2th position of the sequence a should be substituted by C ;Moving the vector (0,-2), we can obtain that the 9th base G of sequence a should be substituted by T; Moving the vector (0,-3), we can obtain that the 5th and the 8th base G should be replaced by A; Moving the vector (0,-6), we can obtain that the 7th base C should be replaced by T.

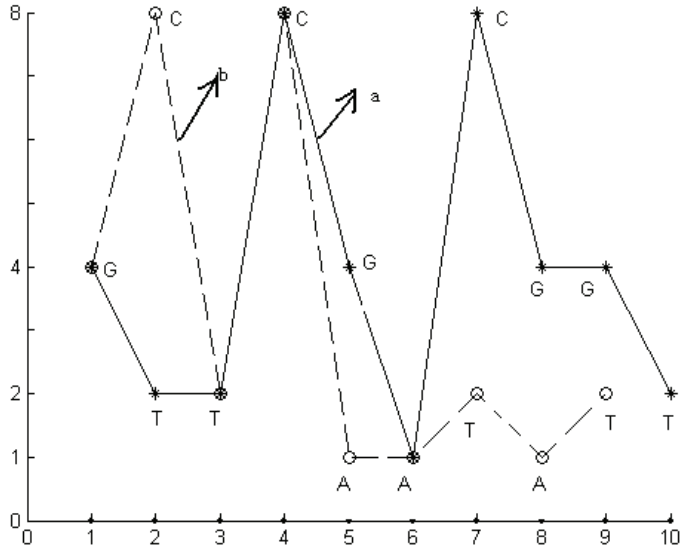

Fig.3. The corresponding curves of sequences a and b.

Conclusion

In this article, we outlined an approach to search optimal alignment based on the new representation of DNA sequence .Compared to other alignment algorithms, the advantage of our method is that it is simple and efficient. By separating the four lines to be not unit distance intervals, we offered an approach to judge mutations quickly between bases in sequences.

Acknowledgment This work is supported in part by the National Nature Science Foundation of China (Grant 10571019) and the National Natural Science Foundation of Hunan University. The authors thank the anonymous referees for many valuable suggestions, which have improved this manuscript.

References

[1] A. Torres, J. J. Nieto, Fuzzy logic in medicine and bioinformatics, J. Biomed. Biotech. **2006** (2006) 1-7.

[2] E. Hamori, J. Ruskin, A novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. **258** (1983) 1318-1327.

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

[4] M. A. Gates, Simpler DNA sequence representations, Nature **316** (1985) 219-219.

[5] M. A. Gates, A simple way to look at DNA, J. Theor. Biol. **119** (1986) 319-328.

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

[7] 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.

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

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

[10] 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.

[11] D. Bielińska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. **442** (2007) 140-144.

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

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

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

[15] 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.

[16] M. Randić, A. T. Balaban, On a four-dimension representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. **43** (2003) 532-539.

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

[18] B. Liao, R.F. Li, W. Zhu, On the similarity of DNA primary sequences based on 5-D representation, J. Math. Chem. **42** (2007) 47-57.

[19] M. Randić, A. F. Kleiner, L. M. DeAlba, Distance/distance matrices, J. Chem. Inf. Comput. Sci. **34** (1994) 277-286.

[20] B. Liao, K. Q. Ding, Graphical approach to analyzing DNA sequences**,** J.

Comput. Chem. **26** (2005)1519-1523.

[21] 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.

[22] G. Hu, S. Shen, J. Ruan**,** SGA: A grammar-based alignment algorithm, Comp. Meth. Progr. Biomed. **86** (2007) 17-20.