Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 18 (2008) 595–602
Progress in Natural Science
www.elsevier.com/locate/pnsc

Special report

A DNA sequence alignment algorithm using quality information and a fuzzy inference method

Kwangbaek Kim [a,*], Minhwan Kim [b], Youngwoon Woo [c]

[a] Division of Computer and Information Engineering, Silla University, 617-736 San1-1 Gwaebop-Dong, Sasang-Gu, Busan, Republic of Korea
[b] School of Computer Science and Engineering, Pusan National University, Busan, Republic of Korea
[c] Department of Multimedia Engineering, Dong-Eui University, Busan, Republic of Korea

Received 17 December 2007; received in revised form 21 December 2007; accepted 21 December 2007

Abstract

DNA sequence alignment algorithms in computational molecular biology have been improved by diverse methods. In this paper, we propose a DNA sequence alignment that uses quality information and a fuzzy inference method developed based on the characteristics of DNA fragments and a fuzzy logic system in order to improve conventional DNA sequence alignment methods that uses DNA sequence quality information. In conventional algorithms, DNA sequence alignment scores are calculated by the global sequence alignment algorithm proposed by Needleman–Wunsch, which is established by using quality information of each DNA fragment. However, there may be errors in the process of calculating DNA sequence alignment scores when the quality of DNA fragment tips is low, because only the overall DNA sequence quality information are used. In our proposed method, an exact DNA sequence alignment can be achieved in spite of the low quality of DNA fragment tips by improvement of conventional algorithms using quality information. Mapping score parameters used to calculate DNA sequence alignment scores are dynamically adjusted by the fuzzy logic system utilizing lengths of DNA fragments and frequencies of low quality DNA bases in the fragments. From the experiments by applying real genome data of National Center for Biotechnology Information, we could see that the proposed method is more efficient than conventional algorithms.

Keywords: DNA sequence alignment; Quality information; Fuzzy inference

1. Introduction

In contig assembly process to acquire an overall DNA sequence of a genome, DNA sequence alignment is very important in molecular biology fields [1–4]. Recently, man power and processing time required for DNA sequence analyses can be reduced by automatic DNA sequence analyzers, but there is still no method to decode an overall DNA sequence of a very long specific genome by one execution of experiment. For a specific genome, each DNA fragment is analyzed after classification into several fragments and an overall DNA sequence should be reconstituted using these fragments information. This is called as contig assembly process [5]. Generally hundreds of nucleotide base pairs can be decoded by single pass of analysis in each experiment. Especially when DNA sequence analysis equipments are used, much sequencing results can be acquired in short time because many fragments are analyzed simultaneously, but low quality DNA bases in tips of DNA fragments are detected.

In conventional algorithms such as PHRED [6] used for DNA sequencing process, if low quality DNA bases exist in tips of DNA fragments, calculation errors in DNA sequencing score will occur. Therefore, we propose an algorithm with low quality information in DNA fragments which uses mapping score parameters to calculate DNA

* Corresponding author. Tel.: +82 51 999 5052; fax: +82 51 999 5657.
E-mail address: gbkim@silla.ac.kr (K. Kim).

1002-0071/$ - see front matter © 2007 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2007.12.011

sequencing scores to a fuzzy logic system. It improves conventional DNA sequence alignment algorithms.

2. Quality information

In DNA sequencing programs, a DNA sequence is created by reading trace data and quality information for every DNA base. In this study, we use quality information produced by PHRED, a well-known DNA sequencing program, because most DNA sequencing programs produce similar data. Trace data in PHRED are created by analyzing peaks of chromatogram in a sequencing machine. By these data, a file called "Fasta" having a DNA sequence and a file called "Quality" having DNA bases quality information are obtained. Fig. 1 shows examples of trace data.

The ideal trace data should have same intervals between every peak and no overlaps. Fig. 1(a) shows an example of almost ideal trace data. In this figure, intervals between each peak are almost uniform and the peak of a base is much higher than the peaks of other bases in a same position. But Fig. 1(b) shows unclear peaks for every base. This is because of errors in the experiment for producing primitive

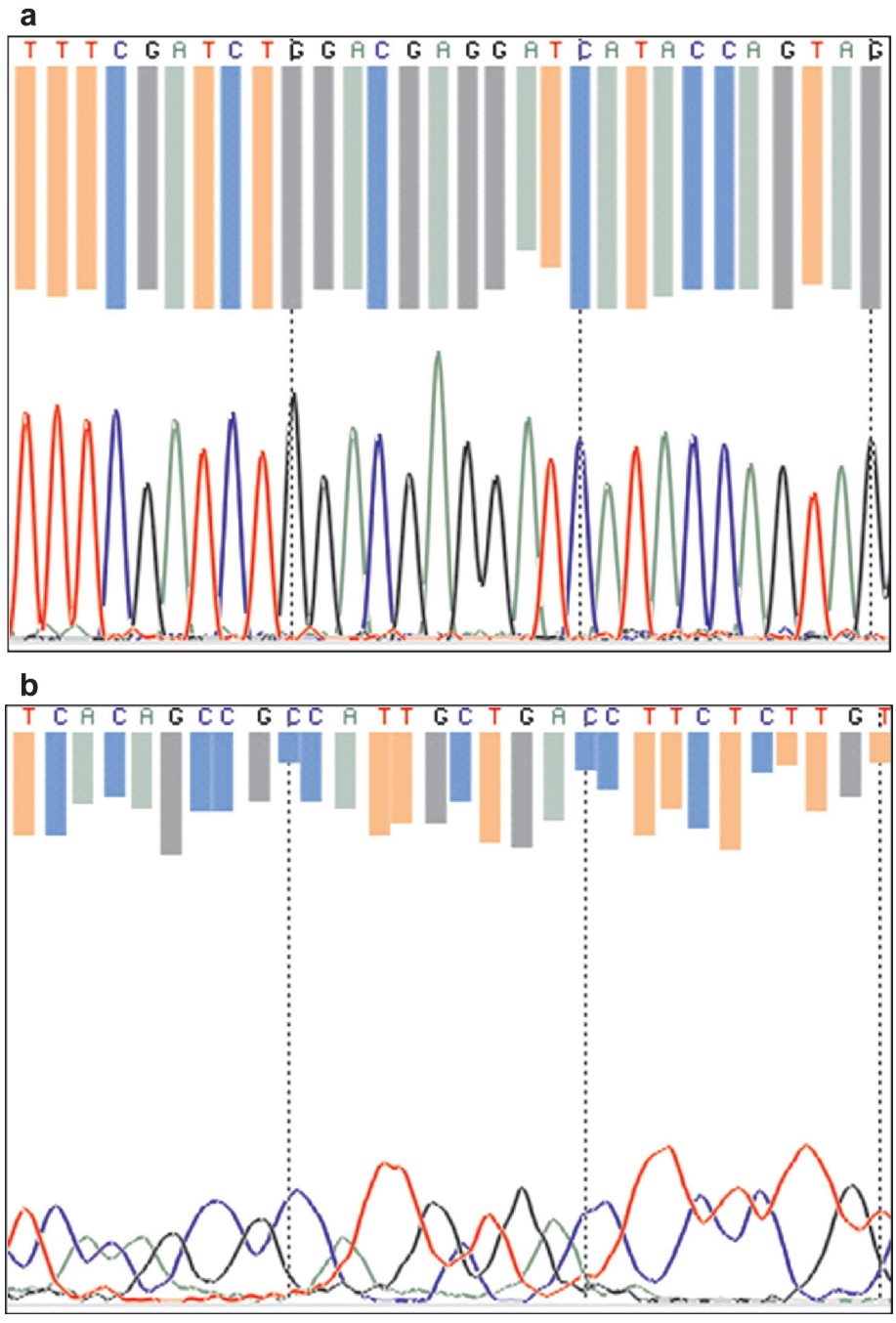

Fig. 1. Examples of trace data. (a) High quality trace data, (b) low quality trace data.

Table 1
Error probabilities in accordance with quality scores

Quality score (Q)	Error probability (P)	Quality score (Q)	Error probability (P)
10	10^{-1}	60	10^{-6}
20	10^{-2}	70	10^{-7}
30	10^{-3}	80	10^{-8}
40	10^{-4}	90	10^{-9}
50	10^{-5}	100	10^{-10}

data for trace data. The characteristics of low quality trace data are as follows: (i) Intervals between each peak are not uniform; (ii) two or more peaks have similar height; (iii) all peaks of four bases are very low in a same position.

Low quality trace data are acquired when DNA bases cannot be decided precisely. Symbols on the top of Fig. 1 are a part of a DNA sequence produced by trace data of PHRED and is called Fasta. In PHRED, base-specific quality scores are one of the most innovative features. Quality scores range from 0 to 99 and this is related with the probability of difference from true DNA bases. If quality score is denoted as Q and error probability of a DNA base is denoted as P, we can get $Q = -10 \times \log_{10} P$. Table 1 shows error probabilities in accordance with quality scores.

A process for trimming a DNA sequence is executed using quality scores after creating a DNA sequence and quality scores in PHRED. Generally, tips of trace data contain errors because of the limitations on biological experiments. So quality scores for tips of DNA fragments are very low by the errors. Tips of DNA fragments having lots of errors have unwanted influence on experiments, so these tips are eliminated. Therefore, the ratio of DNA bases having quality score less than 10 is about 2–5% in a DNA sequence.

3. The proposed DNA sequence alignment algorithm

There is no need to assemble multiple fragments simultaneously in decoding a DNA sequence except a few experiments such as shotgun sequencing. PCR or direct sequencing methods are often used to decode a DNA sequence, and an experimenter can sequence fragments in order through designing an appropriate primer. There are global alignment and local alignment in DNA sequence alignment and various algorithms have been developed for searching optimal alignment of a DNA sequence. In this paper, we propose an algorithm which adjusts mapping score parameters to calculate DNA sequence alignment scores dynamically by applying low quality information in DNA fragments to a fuzzy logic system, and this algorithm improves conventional DNA sequence alignment algorithms.

3.1. The proposed DNA sequence alignment algorithm using quality information

In this section, the proposed DNA sequence alignment algorithm using quality information is explained. First, the definition and characteristics of a DNA sequence using quality information are explained and then the method to align a DNA sequence having quality information and the method to calculate alignment scores are proposed.

3.1.1. A DNA sequence

A DNA sequence means a sequence of DNA bases belonged to Σ. In this paper, we deal with the sequences of DNA bases, so Σ is defined as $\{a, g, c, t\}$ and space is defined as $\Delta \notin \Sigma$. For ith base of a DNA sequence, A is defined as A_i and a partial DNA sequence $A_iA_{i+1}\cdots A_j$ is defined as $A[i..j]$. When two DNA fragments of $A = A_1A_2\cdots A_m$ and $B = B_1B_2\cdots B_n$ with each length m and n are provided, the alignment of two DNA fragments are $A^* = A_1^*A_2^*\cdots A_m^*$ and $B^* = B_1^*B_2^*\cdots B_n^*$ $(n, m \leqslant l)$. A_i^* and B_i^* are classified into one of the three kinds of mapping by a DNA base type. The three kinds of mapping are as follows:

Match: $A_i^* \neq \Delta$, $B_i^* \neq \Delta$, and $A_i^* = B_i^*$
Mismatch: $A_i^* \neq \Delta$, $B_i^* \neq \Delta$, and $A_i^* \neq B_i^*$
Gap: A_i^* or B_i^* is Δ

$A_i^* = B_i^* = \Delta$ is not allowed. Each mapping has its score and the score is γ in case of match or δ in case of mismatch or μ in case of gap. These γ, δ, μ are called mapping score parameters and these parameters have diverse values depending on applications. γ is a positive number, δ and μ are the negative numbers in general.

3.1.2. The proposed DNA sequence alignment algorithm

In a DNA sequence $A = A_1A_2\cdots A_m$ having quality information, each A_i is one of DNA bases in Σ and each Q_{A_i} is the quality score for each A_i. Quality score Q_{A_i} means that error probability of A_i is $10^{-Q_A/10}$. We define a DNA sequence having quality information as "a quality DNA sequence" and a DNA sequence having no quality information as "an ordinary DNA sequence". When we look into the meaning of the quality scores of DNA bases in detail, if a DNA base is $x \in \Sigma$ in a certain position and Q_x is the quality score of the DNA base, x appears in the position by the probability of $1 - 10^{-Q_x/10}$. Other DNA bases appear in the position by the probability of $10^{-Q_x/10}$ or a space appears in the position. A space caused by the absence of a DNA base is denoted as '-' in a quality DNA sequence and made: x is a typical base in the position. In this paper, the following assumptions are made:

Assumption 1. Generally, the probability of a typical DNA base at position i in a quality DNA sequence is higher than 0.9.

Assumption 2. The probabilities of other DNA bases except for a typical base and a space (-) are all the same. If the quality score of the typical DNA base is 10, then the error probability of the position is 0.1 and the probability of the typical DNA base is 0.9. So the probabilities of other DNA bases and a space are 0.025.

When two DNA fragments of $A = A_1A_2\cdots A_m$ and $B = B_1\ B_2\cdots B_n$ with each global alignment length m and n are provided, the alignment of two quality DNA fragments is $A^* = A_1^*A_2^*\cdots A_m^*$ and $B^* = B_1^*B_2^*\cdots B_n^*$ $(n, m \leqslant l)$. A_i^* and B_i^* are aligned by inserting 0 or more spaces (Δ) into gaps in DNA bases of A and B. There is no difference between the alignment by inserting spaces (Δ) and the alignment of an ordinary DNA sequence. The meaning of inserting spaces (Δ) into alignment is that any symbol does not exist, so the probability of a DNA base $x \in \Sigma$ in the position is 0 and the probability of a space (-) is 1. A DNA base pair, A_i^* and B_i^*, is classified into one of the following three kinds of mapping by the typical DNA base:

Regular-match: $A_i^* \neq \Delta$, $B_i^* \neq \Delta$, and $A_i^* = B_i^*$.
Regular-mismatch: $A_i^* \neq \Delta$, $B_i^* \neq \Delta$, and $A_i^* \neq B_i^*$.
Regular-gap: A_i^* or B_i^* is Δ.

$A_i^* = B_i^* = \Delta$ is not allowed. The above three mappings are quality mappings and match, mismatch, and gap of an ordinary DNA sequence are ordinary mappings. A mapping score $S(A_i^*, B_i^*)$ of a DNA base pair, A_i^* and B_i^*, is defined as an expectation value of an ordinary mapping score. Quality mapping becomes one of match, mismatch, and gap of ordinary mapping in accordance with the actual DNA bases. Table 2 shows the result that A_i^* and B_i^* are analyzed into ordinary mapping in accordance with the actual DNA bases.

M: Actual DNA bases match with each other. Match score is γ in this case.
N: Two positions are not spaces and the actual DNA bases do not match with each other. Mismatch score is δ in this case.
G: One position has Σ and the other position is space (-). This is considered as gap mapping and score μ is provided in this case.
E: All of A_i^* and B_i^* are spaces. Score 0 is provided in this case because it can be considered as no mapping on the alignment.

Therefore, if we define the probability of match mapping as $P_m(A_i^*, B_i^*)$, the probability of mismatch mapping as $P_n(A_i^*, B_i^*)$ and the probability of gap mapping as $P_g(A_i^*, B_i^*)$, we can get Eq. (1)

$$S(A_i^*, B_i^*) = \gamma \times P_m(A_i^*, B_i^*) + \delta \times P_n(A_i^*, B_i^*) + \mu \times P_g(A_i^*, B_i^*) \tag{1}$$

The detailed method to calculate each quality mapping is as follows: the probability that A_i^* will be $x \in \Sigma$ is defined as α_x, the probability that A_i^* will be a space is defined as α_-, the probability that B_i^* will be $x \in \Sigma$ is defined as β_x and the probability that B_i^* will be a space is defined as β_-.

3.1.2.1. In the case of regular-match. If we define typical DNA bases of A_i^* and B_i^* as a, Eq. (2) can be derived from Assumption 2.

$$\begin{aligned} \frac{1-\alpha_a}{4} &= \alpha_c = \alpha_g = \alpha_t = \alpha_- \\ \frac{1-\beta_a}{4} &= \beta_c = \beta_g = \beta_t = \beta_- \end{aligned} \tag{2}$$

Table 3 shows the probabilities of regular-match.

Probabilities of each DNA base shown in Table 3 are calculated by Eq. (3)

$$a = \alpha_a, \quad \{c, g, t\} = \frac{1-\alpha_a}{4} \tag{3}$$

X, Y, and Z shown in Table 3 are explained by Eq. (4)

$$X = \frac{(1-\alpha_a)\beta_a}{4}, \quad Y = \frac{\alpha_a(1-\beta_a)}{4}, \quad Z = \frac{(1-\alpha_a)(1-\beta_a)}{16} \tag{4}$$

Eq. (5) can be derived from the probabilities shown in Table 3. Z is ignored in (5) because Z is a very small value of 0.0056 although quality scores of α_a and β_a are all 1.

$$\begin{aligned} P_m(A_i^*, B_i^*) &= \alpha_a\beta_a + 3Z \approx \alpha_a\beta_a \\ P_n(A_i^*, B_i^*) &= 3X + 3Y + 6Z \approx \frac{\alpha_a + \beta_a - 2\alpha_a\beta_a}{4} \times 3 \\ P_g(A_i^*, B_i^*) &= X + Y + 6Z \approx \frac{\alpha_a + \beta_a - 2\alpha_a\beta_a}{4} \end{aligned} \tag{5}$$

Therefore, mapping score of regular-match can be expressed as Eq. (6)

$$S(A_i^*, B_i^*) = \gamma \times \alpha_a\beta_a + \delta \times \frac{\alpha_a + \beta_a - 2\alpha_a\beta_a}{4} \times 3 + \mu \times \frac{\alpha_a + \beta_a - 2\alpha_a\beta_a}{4} \tag{6}$$

Table 2
Ordinary mapping in accordance with the actual DNA bases

	A_i^*				
B_i^*	a	c	t	g	–
a	M	N	N	N	G
c	N	M	N	N	G
t	N	N	M	N	G
g	N	N	N	M	G
–	G	G	G	G	E

Table 3
Probability of regular-match

	A_i^*				
B_i^*	a	c	t	g	–
a	$\alpha_a\beta_a$	X	X	X	X
c	Y	Z	Z	Z	Z
t	Y	Z	Z	Z	Z
g	Y	Z	Z	Z	Z
–	Y	Z	Z	Z	Z

Table 4
Probability of regular-mismatch

B_i^*	A_i^* a	c	t	g	–
a	X	$\alpha_c\beta_a$	X	X	X
c	Z	Y	Z	Z	Z
t	Z	Y	Z	Z	Z
g	Z	Y	Z	Z	Z
–	Z	Y	Z	Z	Z

3.1.2.2. In the case of regular-mismatch. If we define a typical DNA base of A_i^* as c and a typical DNA base of B_i^* as a, Eq. (7) can be derived from Assumption 2.

$$\frac{1-\alpha_c}{4} = \alpha_a = \alpha_g = \alpha_t = \alpha_- \\ \frac{1-\beta_a}{4} = \beta_c = \beta_g = \beta_t = \beta_- \tag{7}$$

Table 4 shows the probabilities of regular-mismatch.

Probabilities of each DNA base shown in Table 4 are calculated by Eq. (8)

$$c = \alpha_c, \quad \{a,g,t\} = \frac{1-\alpha_c}{4} \tag{8}$$

X, Y, and Z shown in Table 4 are, respectively,

$$X = \frac{(1-\alpha_c)\beta_a}{4}, \quad Y = \frac{\alpha_c(1-\beta_a)}{4}, \\ Z = \frac{(1-\alpha_c)(1-\beta_a)}{16} \tag{9}$$

Eq. (10) can be derived from probabilities shown in Table 4. Z is ignored in (10) because Z is a very small value of 0.0056 although the quality scores of α_a and β_a are 1.

$$P_m(A_i^*, B_i^*) = X + Y + 2Z \approx \frac{\alpha_c + \beta_a - 2\alpha_c\beta_a}{4} \\ P_n(A_i^*, B_i^*) = \alpha_c\beta_a + 2X + 2Y + 7Z \approx \frac{\alpha_c + \beta_a}{2} \\ P_g(A_i^*, B_i^*) = X + Y + 6Z \approx \frac{\alpha_c + \beta_a - 2\alpha_c\beta_a}{4} \tag{10}$$

Therefore, mapping score of regular-mismatch can be expressed by Eq. (11)

$$S(A_i^*, B_i^*) = \gamma \times \frac{\alpha_c + \beta_a - 2\alpha_c\beta_a}{4} + \delta \times \frac{\alpha_c + \beta_a}{2} + \mu \\ \times \frac{\alpha_c + \beta_a - 2\alpha_c\beta_a}{4} \tag{11}$$

3.1.2.3. In the case of regular-gap. If we define a typical DNA base of A_i^* as a and a typical DNA base of B_i^* as space (Δ), $\beta_- = 1$ because B_i^* becomes space (Δ). Mapping score of regular-gap is expressed by Eq. (12).

$$S(A_i^*, B_i^*) = \mu \times (1 - \alpha_-) = \mu \times \frac{3 + \alpha_a}{4} \tag{12}$$

Alignment score $S(A_i^*, B_i^*)$ of a quality DNA sequence is defined as the summation of mapping scores of all DNA base pairs for alignment like an ordinary DNA sequence. Alignment score $S(A_i^*, B_i^*)$ is defined as Eq. (13).

$$S(A^*, B^*) = \sum_{i=1}^{l} S(A_i^*, B_i^*) \tag{13}$$

In this proposed algorithm, when two DNA fragments with length m and length n are provided, and if an alignment score of the DNA fragments has the highest value, the alignment is considered as an optimal alignment and the optimal alignment is searched. If we define $H_{i,j}$ as the optimal alignment score of $A[1\ldots i]$ and $B[1\ldots j]$, $H_{i,j}$ can be calculated by Dynamic Programming method and this method has a same structure as conventional Needleman–Wunsch algorithm. So memory of $O(mn)$ and time of $O(mn)$ are required in this method.

3.2. Adjustment of mapping score parameters using fuzzy inference rules

Tips of a DNA fragment having quality score less than 20 can be occurred in the process of decoding a DNA sequence because of experimental limitations. In this paper, we define such a DNA fragment as a low quality DNA fragment. In a conventional algorithm [7] using quality score, errors often occur in calculating sequence mapping scores in case of large difference of length between DNA fragments and of low quality of tips of DNA fragments because optimal fragments are selected by mapping score parameters of user input. In order to solve these problems, we improved the conventional algorithm by applying lengths of DNA fragments and low quality information in DNA fragments to a fuzzy logic system. In the improved algorithm, mapping score parameters are adjusted dynamically utilizing a fuzzy logic system. Inputs of the fuzzy logic system are the lengths of DNA fragments and the frequencies of low quality bases in the fragments, outputs of the fuzzy logic system are mismatch mapping score parameters. A fuzzy logic system is composed of fuzzification of input signals, fuzzy inference by fuzzy rules based on expert knowledge and defuzzification of outputs. In this study, we use a mini-max operator for the inference of fuzzy rules and the center of gravity method [8] such as Eq. (14) for the defuzzification.

$$y^* = \frac{\sum \mu(y_i)x_i}{\sum \mu(y_i)} \tag{14}$$

3.2.1. Membership functions for the length of a DNA fragment

We designed membership functions for the length of a DNA fragment as shown in Fig. 2(a), and calculated membership grades for the lengths of DNA fragments using the functions. Low section means the length of a DNA fragment is short, middle section means the length of a DNA

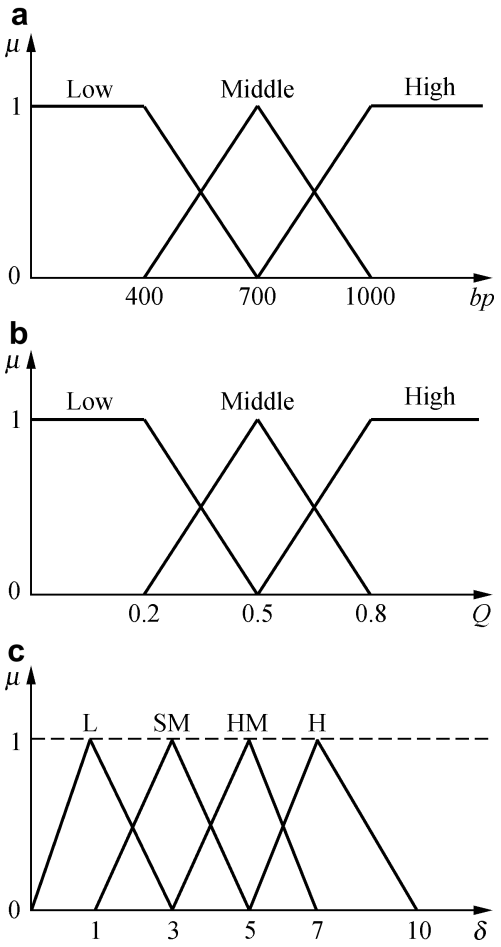

Fig. 2. Membership functions. (a) For the length of a DNA fragment; (b) for the frequency of low quality DNA bases; (c) for a mismatch mapping.

fragment is medium, and high section means the length of a DNA fragment is long.

Frequency of low quality (quality score is less than 20) DNA bases in a fragment, Q, is calculated by Eq. (15)

$$Q = \frac{\#\text{ of low quality DNA bases}}{\text{Total length of a fragment}} \tag{15}$$

Fig. 2(b) shows membership functions for the frequency of low quality DNA bases. In the figure, low section means the frequency of low quality DNA bases is low, middle section means the frequency of low quality DNA bases is medium, and high section means the frequency of low quality DNA bases is high.

3.2.2. Output membership functions for mismatch mapping score parameters

The final mismatch mapping scores are calculated by defuzzification using the center of gravity method after inferring the fuzzy inference rules (Table 5) based on membership grades of the lengths of DNA fragments and the frequencies of low quality DNA bases in the fragments. Fig. 2(c) shows output membership functions for a mismatch mapping score. Match mapping scores and gap

Table 5
Fuzzy inference rules for a mismatch mapping score

	Q		
	Low	Middle	High
Low	L	SM	SM
Middle	SM	SM	HM
High	SM	HM	H

mapping scores are dynamically adjusted by difference of ±1 from mismatch mapping scores derived from the fuzzy logic system for every replacement of fragments.

4. Experimental results and analyses

4.1. Experimental environments

The experimental environments are implemented by Microsoft Visual C++ 6.0 and installed on a Samsung laptop computer having single CPU of 1.3 GHz and main memory of 512 MB. We used genome data received from NCBI (http://www.ncbi.nlm.nih.gov/traces/trace.fcgi), which is "gnl|ti|1147316796", and from influenza A virus. Each genome is composed of 166 DNA sequences having Fasta and Quality files generated by PHRED. The length of each fragment is between 311 and 872 bp. Fig. 3(a) shows the main window of the implemented program by the proposed algorithm for optimal sequence alignment.

In the main window of the implemented program, the average quality of the fragments and the number of DNA bases in each fragment can be checked after loading Fasta and Quality files received from NCBI. Quality of each DNA base can be distinguished by color expression. Fig. 4 shows the information of a DNA fragment and color expression for each quality class. Table 6 shows the range of quality scores for each quality class.

4.2. Analyses of experimental results

We experimented with real 166 DNA sequences received from NCBI in order to compare the proposed DNA sequence alignment algorithm with the conventional DNA sequence algorithm using quality information [7]. Low quality fragments appeared in all of the 166 DNA sequences applied to the experiment. In Fig. 3(b), red colored symbols in the tips of two DNA fragments indicate low quality DNA bases.

Table 7 shows the numbers of matched and mismatched optimal DNA sequence alignments implemented by the proposed algorithm and the conventional algorithm [7]. In Table 7, cases of matched optimal alignments in both algorithms occur when the average quality score of fragments is not less than 41, and cases of mismatched optimal alignments in both algorithms are occurred when the average quality score of fragments is no more than 20. Left part of Fig. 5 shows a part of the optimal alignment of a DNA sequence by the conventional algorithm and the right part

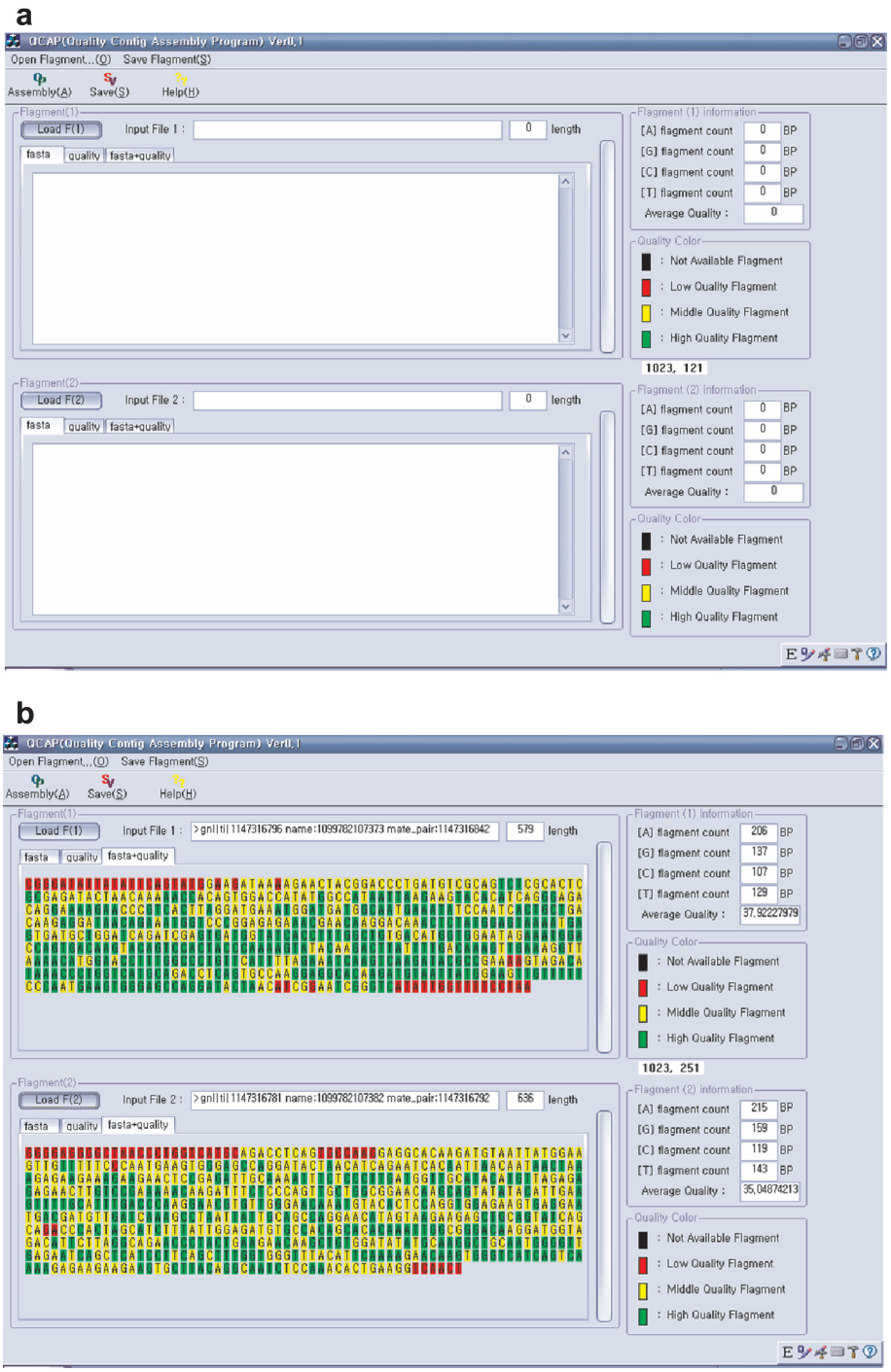

Fig. 3. The main window of the implemented program for optimal sequence alignment (a) and a result window for the information of two DNA fragments (b).

of Fig. 5 shows a part of the optimal alignment of a DNA sequence alignments by the proposed algorithm. Table 8 shows the results of optimal sequence alignments by the conventional algorithm and the proposed algorithm. The number of match in the proposed algorithm is more than the number of match in the conventional algorithm and the number of mismatch in the proposed algorithm is less than the number of mismatch in the conventional algorithm. So we verified the improvement of DNA sequence alignment in the proposed algorithm by less error rate than the conventional algorithm from the experiments.

5. Conclusions

In this paper, we propose an algorithm for optimal DNA sequence alignment utilizing quality information generated from a DNA sequencing program such as PHRED used for sequencing process of biological DNA

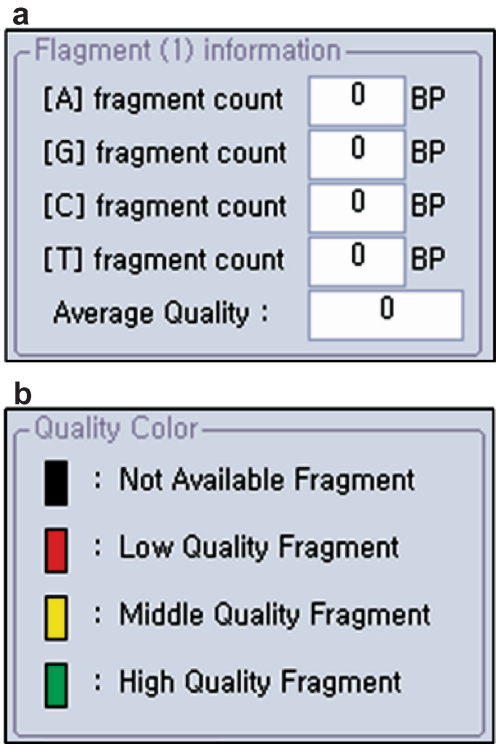

Fig. 4. Information of a DNA fragment and color expression for each quality class. (a) Information of a DNA fragment; (b) color expression for each quality class.

Table 6
Range of quality scores

Quality class	Range of quality scores
Nonavailable fragment	0
Low quality fragment	1–20
Middle quality fragment	21–40
High quality fragment	41–99

Table 7
The numbers of matched and mismatched optimal DNA sequence alignments

No. of matched alignments in both algorithms	No. of mismatched alignments in both algorithms
12/166	144/166

sequences. In the proposed algorithm, mapping score parameters are not inputted by a user but dynamically adjusted by applying lengths of DNA fragments and frequencies of low quality DNA bases in the fragments to fuzzy inference rules. Through the experiments, we verified the improvement of DNA sequence alignment by the proposed algorithm which has lower error rate than the conventional global alignment algorithm using only quality information in spite of low quality of DNA fragment tips. We applied quality information and mapping score parameters of DNA fragments to a global alignment algorithm and a fuzzy logic system in this study. Further research should involve application of a local alignment algorithm in order to improve accuracy of DNA sequence alignment.

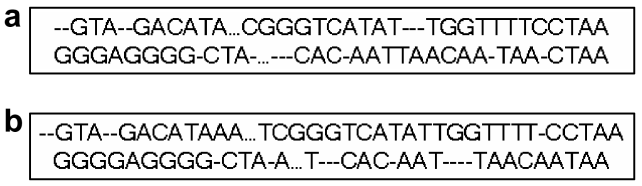

Fig. 5. Parts of the optimal alignment by the conventional algorithm and the proposed algorithm.

Table 8
Results of optimal sequence alignments

	The conventional algorithm [7]	The proposed algorithm
No. match	10/29	14/29
No. mismatch	14/29	10/29
No. gap	15/29	15/29

References

[1] Waterman MS. Introduction to computational biology. Oxford: Chapman and Hall; 1995.
[2] Gusfield D. Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge: Cambridge University Press; 1997.
[3] Apostolico A, Giancarlo R. Sequence alignment in molecular biology. J Comput Biol 1998;5(2):173–96.
[4] Pevzner P. Computational molecular biology: an algorithmic approach. MA: MIT Press; 2000.
[5] Staden R. A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res 1980;8:3673–94.
[6] Ewing B, Hillier L, Wend MC, et al. Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res 1998;8(3):175–85.
[7] Na JC, Noh KH, Park KS. DNA sequencing algorithm using quality information. Korea Inf Sci Soc 2005;32(11):578–86.
[8] George JK, Bo Y. Fuzzy sets and fuzzy logic theory and applications. NJ: Prentice Hall; 1995.