Fragments Combination of DNA Sequence Alignment Using Fuzzy Reasoning Rule

Kwang Baek Kim[1], Dong Hui Yu[2] and Soyoung Hwang[2]*

[1]*Department of Computer Engineering, Silla University, Korea*
[2]*Department of Multimedia Engineering, Catholic University of Pusan, Korea*
gbkim@silla.ac.kr, dhyu@cup.ac.kr, soyoung@cup.ac.kr

Abstract

We proposed a method complementing failure of combining DNA fragments, defect of conventional contig assembly programs. In the proposed method, very long DNA sequence data are made into a prototype of fragment of about 700 bases that can be analyzed by automatic sequence analyzer at one time, and then matching ratio is calculated by comparing a standard prototype with 3 fragmented clones of about 700 bases generated by the PCR method. In this process, the time for calculation of matching ratio is reduced by Compute Agreement algorithm. Two candidates of combined fragments of every prototype are extracted by the degree of overlapping of calculated fragment pairs, and then degree of combination is decided using a fuzzy inference method that utilizes the matching ratios of each extracted fragment, and A, C, G, T membership degrees of each DNA sequence, and previous frequencies of each A, C, G, T. In this paper, DNA sequence combination is completed by the iteration of the process to combine decided optimal test fragments until no fragment remains. For the experiments, fragments of about 700 bases were generated from each sequence of 10,000 bases and 100,000 bases extracted from 'PCC6803', complete protein genome. From the experiments by applying random mutations on these fragments, we could see that the proposed method was faster than FAP program, and combination failure, defect of conventional contig assembly programs, did not occur. .

Keywords: *DNA fragments, PCR method, DNA sequence, fuzzy inference*

1. Introduction

In case of trying to determine a very long DNA sequence from a DNA sequencing project, at first the DNA sequence is converted to several DNA fragments and the sequences of the fragments is found out. And then the original long DNA sequence is reconstituted from the identified sequences of the fragments. The reconstitution problem occurred in this process is called `contig assembly problem'[1]. This problem requires high-speed computational ability of a computer because of inherent complexity and large amount of computation of the problem.

Nowadays, SEQAID[2], CAP[3] and FAP[4] are known as programs for assembling contig from sequences of DNA fragments. Four bases of A (Adenine), C (Cytosine), G (Guanine) and T (Thymine) can be used as input fragments in almost these programs and N is also used to represent an uncertain base.

In this paper, we proposed an algorithm to complement combination failure, defect of conventional contig assembly programs. In the proposed algorithm, we added a fuzzy reasoning method on a conventional sequencing method utilizing only matching ratio, and so combination failure did not occur and combination of uncertain bases was to be possible. In

order to acquire test data, we applied random mutations on 'Synechocystis PCC6803', complete protein genome. In the experiment results by these test data, we could see all original sequences were constituted and execution time was proportional to the number of fragments. And combination failure, defect of conventional contig assembly programs, did not occur

2. The Proposed Algorithm for DNA Sequence Analysis

For the process of DNA sequence analysis of this paper, 3 clones are generated by the PCR (Polymerase Chain Reaction) method [5] and a prototype of fragment of about 700 bases analyzed by a conventional automatic sequence analyzer. The 3 clones are fragmented to about 700 bases and are compared with a standard prototype in order to measure matching ratio. 2 candidate combination fragments for each prototype are extracted by degree of overlapping of fragment pairs.

Degree of combination is decided by a fuzzy reasoning method utilizing matching ratio of each extracted fragment, membership of A, C, G, T, and each previous frequency of A, C, G, T. Sequence combination is completed by the iteration of the process to combine decided optimal fragments with standard prototypes until no fragment remains.

2.1. The Proposed Compute Agreement Algorithm

The proposed Compute Agreement algorithm is as follows.

Step 1: A prototype fragment is located as row and a test fragment is located as column.

Step 2: The first column symbol (base) is compared with each symbol of row prototype symbols as shown in figure 1. If the first column symbol matches with a row symbol, it continues to search to the last column symbol in diagonal direction. If all the symbols are matched from the first column symbol to the last column symbol or the last row symbol, the length of matched symbol is set as a candidate matching ratio.

Step 3: The first row prototype symbol is compared with each symbol of column symbols as shown in Figure 1. If the first row symbol matches with a column symbol, it continues to search to the last row symbol in diagonal direction. If all the symbols are matched from the first row symbol to the last row symbol or the last column symbol, the length of matched symbols is set as another candidate matching ratio.

Step 4: The final matching ratio is set to the longer one between a candidate matching ratio of step 2 and a candidate matching ratio of step 3. If the symbols are matched to the last row symbol, the fragment is combined to left side. If the symbols are matched to the last column symbol, the fragment is combined to right side.

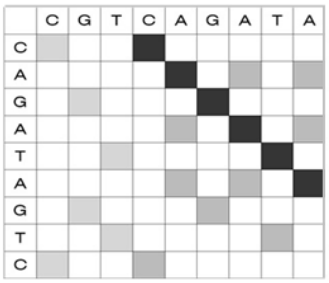

Figure 1. The Proposed Compute Agreement Algorithm

The best 2 fragments having the largest matching ratio are extracted from test fragments by each prototype fragment measured by the proposed algorithm, and then a test fragment is selected by the proposed fuzzy inference method.

2.2. Fragment Selection for Combination using the Proposed Fuzzy Inference Method

In conventional programs such as SEQAID, CAP, FAP only matching ratio is used for combination and so fragment having minimum error rate is combined. In this paper, combination is decided by fuzzy reasoning rules using memberships of bases of best 6 fragments selected by matching ratio calculated from the proposed fuzzy reasoning method and the memberships of previous frequencies. The fuzzy inference algorithm for selecting a fragment for combination is as follows.

Step 1: Membership grades are calculated for each A, C, G, T of the best fragment's bases. Equation (1) shows fuzzy input values.

$$\text{fuzzy value of each } A, C, G, T = \frac{\text{number of each } A, C, G, T}{\text{number of table bases}} \quad (1)$$

Step 2: Membership grades of previous frequencies are calculated. Membership value of the first base is calculated by equation (2) because there is no previous base of the first base. Membership values except the first base are calculated by equation (1).

$$\text{fuzzy value of the first base} = \frac{\sum \text{matching ratios of the best fragments}}{\text{total matching ratio}} \quad (2)$$

Step 3: Combination is decide by the proposed fuzzy reasoning rules using membership grades of each DNA base, A, C, G, T and previous frequencies.

The optimal test fragments decided by above steps are combined and DNA sequence combination is completed by the iteration of the process until no fragment remains.

Membership grades for each base, A, C, G, T of the best fragment are calculated using membership function shown in figure 2. In figure 2, Low interval represents low membership grade for prototype fragment and High interval represents high membership grade for the prototype fragment.

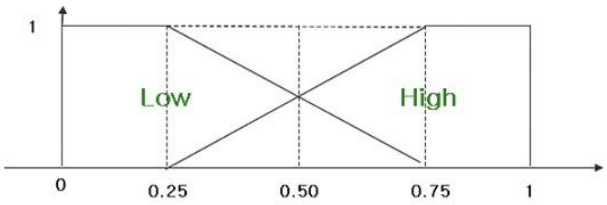

Figure 2. Membership Function for each A, C, G, T Base

Membership grades for previous frequencies of each base, A, C, G, T of the best fragment are also calculated using membership function shown in Figure 2.

μ (A), μ (G), μ (C), μ (T) are membership grades for each base and μ (Y_a), μ (Y_g), μ (Y_c), μ (Y_t) are membership grades for previous frequencies of each base. μ (W) is a

membership grade for deciding combination of each final base. Rules to reason μ (W) for four kinds of bases, A, C, G, T are as follows.

If μ(A) is L, μ(Y_a) is L then μ(W) is F	If μ(A) is L, μ(Y_a) is H then μ(W) is F
If μ(A) is H, μ(Y_a) is L then μ(W) is F	If μ(A) is H, μ(Y_a) is H then μ(W) is T
If μ(G) is L, μ(Y_g) is L then μ(W) is F	If μ(G) is L, μ(Y_g) is H then μ(W) is F
If μ(G) is H, μ(Y_g) is L then μ(W) is F	If μ(G) is H, μ(Y_g) is H then μ(W) is T
If μ(C) is L, μ(Y_c) is L then μ(W) is F	If μ(C) is L, μ(Y_c) is H then μ(W) is F
If μ(C) is H, μ(Y_c) is L then μ(W) is F	If μ(C) is H, μ(Y_c) is H then μ(W) is T
If μ(T) is L, μ(Y_t) is L then μ(W) is F	If μ(T) is L, μ(Y_t) is H then μ(W) is F
If μ(T) is H, μ(Y_t) is L then μ(W) is F	If μ(T) is H, μ(Y_t) is H then μ(W) is T

3. Experimental Results and Analysis

The main title (on the first page) should begin 1 3/16 inches (7 picas) from the top edge of the page, centered, and in Times New Roman 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Please initially capitalize only the first word in other titles, including section titles and first, second, and third-order headings (for example, “Titles and headings” — as in these guidelines). Leave two blank lines after the title.

The program for experiment is implemented by Visual Studio 6.0. 'Synechocystis PCC6803', a complete protein genome is applied as experimental data in order to acquire test data. For the experiment, fragments of about 700 bases are generated from each sequence of 10,000 bases and 100,000 bases extracted from the protein because the length of the protein sequence is about 3.5 million bases. And then random mutations are applied to these fragments.

A captured image of the program interface for the proposed DNA sequence combination is presented in Figure 3.

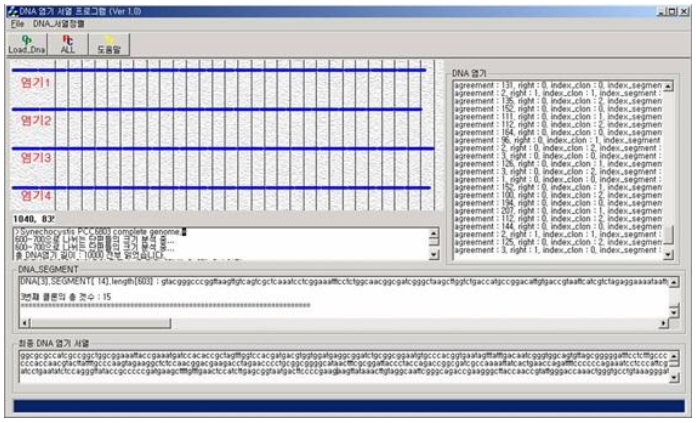

Figure 3. The DNA Sequence Combination Program Interface

The proposed method was experimented with fragments generated from each sequence of 10,000 bases and 100,000 bases having random mutations. In the results, all fragments were

combined to the original sequence and the processing time was proportional to the number of bases.

In Table 1, combination for all fragments in FAP method sometimes did not complete because only matching ratios were used. But combination failure did not occur in the proposed method of this paper, because a test fragment having the largest membership grade was combined. The portion of measuring matching ratios of fragment pairs consumed most processing time for combining sequence. The processing time for combining fragments in the proposed method was less than FAP method, because the proposed Compute Agreement algorithm was applied in order to reduce processing time for measuring matching ratios.

Table 1. Combination Time by the Number of Extracted Bases

Number of bases	FAP	Synechocystic PCC 6803
10,000	26 sec	24 sec
100,000	252 sec	243 sec

4. Conclusions

In this paper, very long DNA sequence data were made into a prototype of fragment of about 700 bases that can be analyzed by automatic sequence analyzer at one time, and then matching ratio was calculated by comparing standard prototypes with 3 fragmented clones of about 700 bases generated by PCR method. In this process, the time for calculation of matching ratio was reduced by the proposed Compute Agreement algorithm. Two candidates of combined fragments of every prototype were extracted by the degree of overlapping of calculated fragment pairs, and then degree of combination was decided using a fuzzy reasoning method that utilizes the matching ratios of each extracted fragment, and A, C, G, T membership grades of each DNA sequence, and previous frequencies of each A, C, G, T. In the proposed method, DNA sequence combination was completed by the iteration of the process to combine decided optimal test fragments until no fragment remains.

'Synechocystis PCC6803', a complete protein genome was applied as experimental data in order to acquire test data. For the experiment, fragments of about 700 bases were generated from each sequence of 10,000 bases and 100,000 bases extracted from the protein because the length of the protein sequence was about 3.5 million bases. And then random mutations were applied to these fragments. In the experimental results using these fragments, the processing time of the proposed method reduced in comparison with FAP program, and combination failure, defect of conventional contig assembly programs, did not occur. The proposed method in this paper improved in comparison with previous researches because all the fragments finally combined were assembled to the original sequence.

References

[1] R. Staden, “A new computer method for the storage and manipulation of DNA gel reading data”, Nucl. Acids. Res., 8, 16 **(1980)**.

[2] H. Peltola, H. Söderlund and E. Ukkonen, “SEQAID: a DNA sequence assembling program based on a mathematical model”, Nucl. Acids. Res., 12, 1 **(1984)**.

[3] X. Huang, “A contig assembly program based on sensitive detection of fragment overlaps”, Genomics, 14, 1 **(1992)**.

[4] B. U. Lee, K. J. Park, W. Park and Y. H. Park, “Development of a program for fragment assembly from DNA sequence data”, Korea Journal of Microbiology and Biotechnology, 25, 6 **(1997)**.

[5] F. Sanger, S. Nicklen and A.R. Coulson, “DNA Sequencing with chain-terminator inhibitors”, Proc. Natal. Acad. Sci. USA, 74, 12 **(1977)**.

Authors

Kwang Baek Kim received his M.S. and the Ph.D. degrees in Department of Computer Science from Pusan National University, Busan, Korea, in 1993 and 1999, respectively. From 1997 to present, he is a professor, Department of Computer Engineering, and Silla University in Korea. He is currently an associate editor for Journal of The Korea Society of Computer and Information, and The Open Artificial Intelligence Journal (USA). His research interests include fuzzy neural network and application, bioinformatics and image processing.

Dong Hui Yu received the B.S., the M.S., and the Ph.D. degrees in Computer Science from Pusan National University, Busan, Korea in 1992, 1994, 2001 respectively. From 1994 to 1997, she was a researcher in ETRI (Electronics and Telecommunications Research Institute), Daejeon, Korea. From 2003, she has been a faculty of Department of Multimedia Engineering at Catholic University of Pusan, Korea. Her research interests are time synchronization and fuzzy neural network.

Soyoung Hwang received the B.S., the M.S., and the Ph.D. degrees in Computer Science from Pusan National University, Busan, Korea in 1999, 2001 and 2006 respectively. From 2006 to 2010, she was a senior researcher in ETRI, Daejeon, Korea. Since 2010, she has been a professor of Department of Multimedia Engineering at Catholic University of Pusan, Korea. Her research interests include embedded systems and fuzzy neural network.