Analysis of DNA Sequence Alignment Using Fuzzy Membership Degree

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

[1] Department of Computer Engineering, Silla University
617-736 Busan, Korea
gbkim@silla.ac.kr
[2] Department of Multimedia Engineering, Catholic University of Pusan
609-757 Busan, Korea
{dhyu, 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. 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] and CAP[3] 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.

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[4] 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.

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.

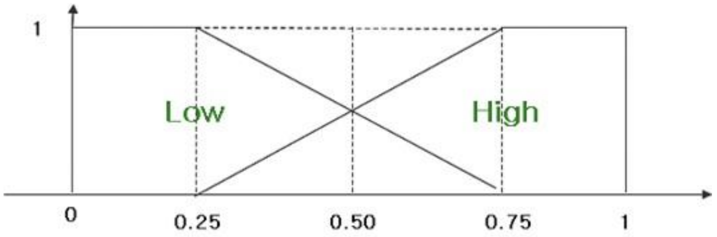

Fig. 1. Membership function for each A, C, G, T base

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 1. In figure 1, Low interval represents low membership grade for prototype fragment and High interval represents high membership grade for the prototype fragment.

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 1. Rules to reason $\mu(W)$ for four kinds of bases, A, C, G, T are applied to Max-Min inference rule.

3 Experimental Results and Analysis

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.

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

'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, 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. F. Sanger, S. Nicklen, A.R. Coulson, DNA Sequencing with chain-terminator inhibitors, Proc. Natal. Acad. Sci. USA, 74, 12 (1977)