情報学基礎 36－5
（1995. 2. 7）

結合編集ソフトウエアShotgunの評価

王忠清　村上康文　浴俊彦　菅原秀明
理化学研究所ライフサイエンス研究情報室

ゲノムプロジェクトにおいて、DNAシークエンサーから読まれてくる数多くのDNA断片を、それらの間の重複部分配列を利用して、ゲノムの塩基配列を再構成する結合編集システムは重要な役割を果たしている。ここでは、Shotgun（塩基配列フラグメントデータの結合編集システム）の評価方法とこの結果について報告する。この方法では43kbのフラグメントをランダムで500個の350bから450bまでのフラグメントに切り出して、そして、それらのデータをShotgunにかけ、結果を評価する。この評価によって、Shotgunシステムがヒトゲノム配列を高い精度で再構成できることが示されている。

A Method for the Test of Sequence Assembly Software

Zhongqing Wang, Yasufumi Murakami, Toshihiko Eki, Hideaki Sugawara
The Institute of Physical and Chemical Research(RIKEN), Hirozawa 2-1,
Wakoh, Saitama, 351-11, Japan.

We have developed a method for evaluating a sequence assembly software: Shotgun. In this method, a long sequence(43kb) was randomly divided into overlapping 500 fragments and the lengths of these fragments are between 350b to 450b. In addition, various degrees of error were randomly added to these fragments in order to simulate the actual experiment situations. This method is also applicable to the evaluation of other sequence assemble programs.

1. Introduction

One of the fundamental problems in human genome projects is the determination and interpretation of DNA sequence data. The determination of a sequence involves the assembling of several small, overlapping sequence fragments into a single, long contiguous sequence. In order to process these kinds of large DNA data, computer technology has played an important role. The program Shotgun was developed to assemble sequence data in a random sequencing method and has been selected as an assembling tool for the sequencing project of yeast chromosome VI. In a typical project to sequence cosmid DNA (30-40kb) several hundreds of sequence data(350b to 450b) are to be assembled to construct one consensus sequence in the shotgun sequencing method. In order to undertake an evaluation of the accuracy of Shotgun program, we chose an actual genomic sequence(43 kb fragment, the genomic sequence of human inter feron alpha receptor gene locus, accession number:X60459) and developed a testing program which cut this sequence randomly into 500 fragments and tried to assemble the model data sets with the program. The lengths of these fragments are ranged at random from 350 b to 450 b. In addition, some kinds of errors are added to these fragments in order to simulate actual experimental situations. The results obtained under various conditions indicated that Shotgun program is efficient enough to reconstruct human genomic sequences.

2. Structure of Shotgun system

Several stages are involved in the determination of a DNA sequence. First, the DNA sequence fragments obtained from the DNA sequencer in a few hundred bases are entered into the computer. Next, the vector insertion sites must be recognized and either deleted or marked so they are not included in the assembly. Then the fragments are aligned and formed into contigs. Finally, the overlapping fragments in the contig are examined and edited into final consensus sequence. The Shotgun program can connect these fragments by using the overlaps among them. In this process, the overlapping parts between 2 fragments are first searched out, and then if they satisfy the following two condition, 1)the over lapping base sequence is long enough and 2) the matching base ratio of the overlapping parts of the two fragments is high enough the two fragments are then connected into a new fragments. The above process is repeated until a whole sequence is reconstructed. Shotgun program consists of 4 parts: 1)Database creating program, 2) database handling program, 3) Base sequence editing program and 4) Fragments connecting program.

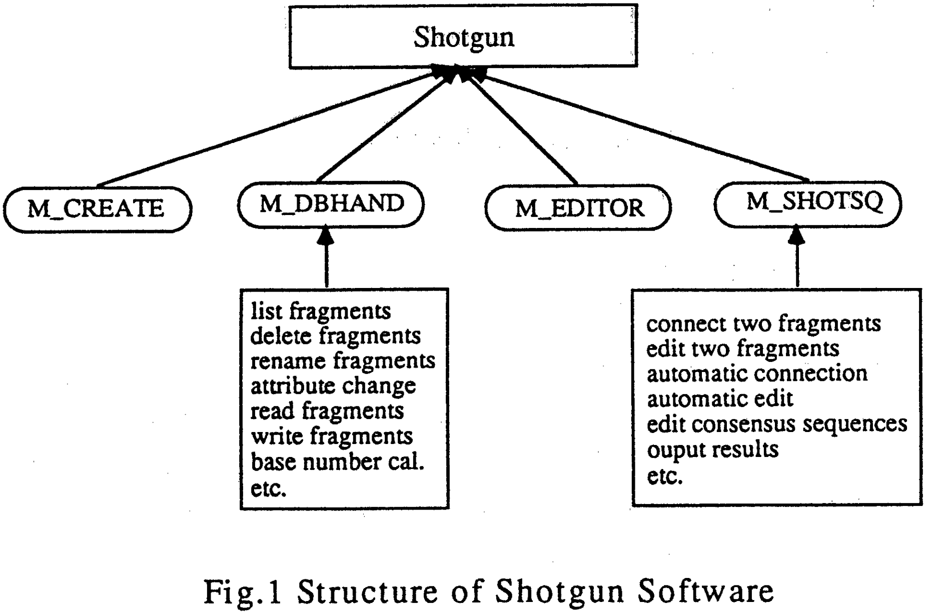

Fig.1 Structure of Shotgun Software

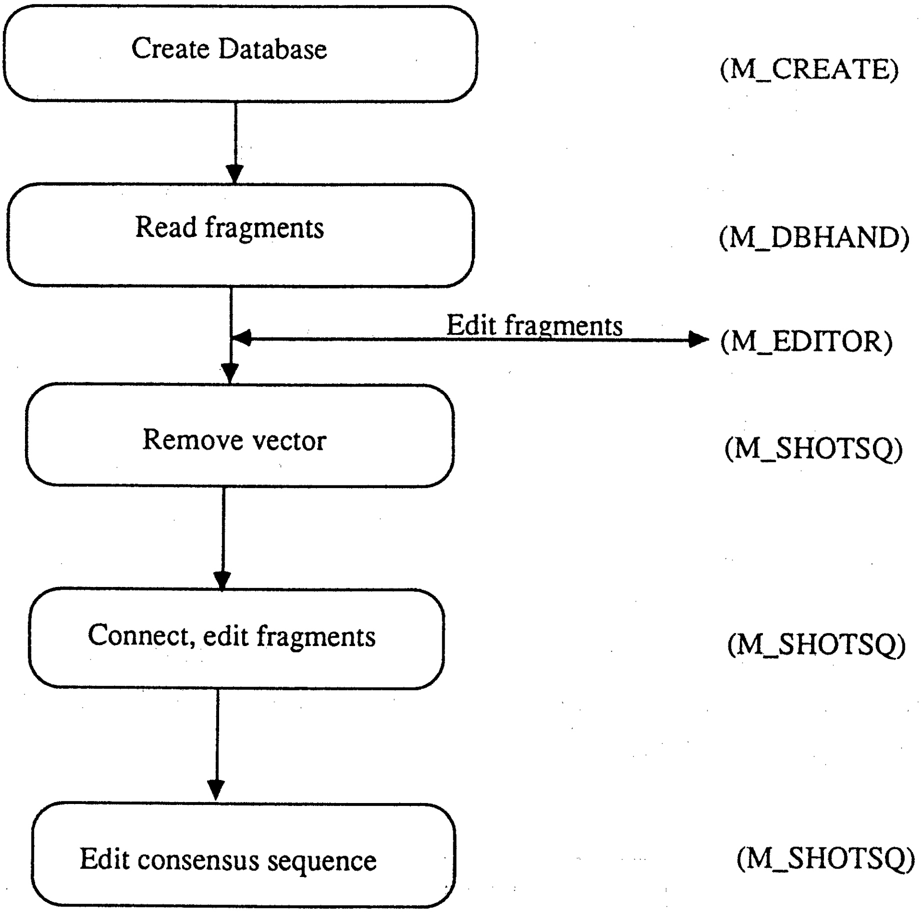

Fig.2 Flowchart of Shotgun program

3. Evaluation Results

In order to evaluate Shotgun system, a simple method is to cut a known genomic sequence into many fragments which are fed to the system to be connected into the whole fragment. We had developed a program which randomly cuts the 43kb sequence into 500 fragments. The lengths of these fragments are 350-450b and are randomly distributed. After the Shotgun system was run with these fragments, all of them were connected perfectly into one fragment.

However the above method is too simple to reflect the real situation. From the viewpoint of experiment experts, the actual fragments read from a DNA sequencer usually have some kinds of errors. For example, a "A", "G", "T", or "C" base may be read incorrectly as "N", and "AAA", "GGG", "TTT", "CCC" may be read incorrectly as "AAAA", "GGGG", "TTTT", "CCCC" or "AA", "GG", "TT", "CC". The error rate is usually about 1%, that is 1% of the bases of a fragment may be misread.

In the revised program, the 43kb sequence is cut randomly into 500 fragments. From these 500 fragments, 400 fragments are selected randomly and are divided into 8 groups. Each group consists of 50 fragments. In group1, the "AAA"s in fragments are changed intentionally to "AAAA"s, in group2 "CCC"s to "CCCC"s, in group3 "GGG"s to "GGGG"s, in group4, "TTT"s to "TTTT"s, in group5 "AAA"s to "AA"s, In group6, "CCC"s to "CC"s, in group7, "GGG"s to "GG"s and in group8 "TTT"s to "TT"s.

According to experiment experiences, errors like "A->N", "C->N", "G->N", "T->N" usually occur at the final 100 bases of fragments. These kinds of errors are simulated in the program as follows. The remaining 100 fragments are divided into 4 groups. Each group consists of 25 fragments. In each group, "A"s, "C"s, "G"s and "T"s which are located at the final 100 bases of the fragments are replaced respectively by "N"s at certain percentage.

It takes more than 20 hours on SUN SS10 for the evaluation program to generate the 500 fragments which may have every kind of errors described as above. It takes another 20 hours for Shotgun program to connect these 500 fragments. Table1 shows the evaluation results of shotgun system. Here, the length of a fragments is between 350b to 450b. The connection conditions are: overlapping length is more than 50 b, 2) matching ratio is over 80%, Error(%) is the average error bases per 100 bases. The Shotgun program was evaluated by using default connection conditions (N = 50, R = 80) where M means min. overlapping length and R means % ratio of match.

From table1, it is indicated that if no error exists in these fragments, Shotgun can connect the 500 fragments completely. And if the error rate is below 2%, it can also connect them completely(no isolated fragment and no extra island). Even though the error rate is as high as 3%, the fragments can also be connected into one island(in this case 3 fragments did not take part in the assembly).

Case_No	Fragments	Connection		Island	Isolated Fragments	Error(%)
		M	R			
1	500	50	80	1	0	0
2	500	50	80	1	0	1
3	500	50	80	1	0	2
4	500	50	80	1	3	3

Table1 Evaluation Results

4. Conclusion

In this study, we evaluated the Shotgun program for sequence assemble by simulating the situations which usually happen in the real experiments. From out evaluation results, it could be concluded that the fragments of 350-450b that have about 1% errors could be connected perfectly by the Shotgun program. The techniques described here are applicable to other sequence assembly programs that run on any type of platform. This method also can be used to check the robustness of an assembly by splitting a known sequence into random fragments, adding different kinds of errors and reassembling. By comparing the error level at which the assembly fails with the error level of the original fragments, the level of confidence in an assemble software can be gained.

References

1) Dumas, J-P., Ninio, J.: Nucleic Acids Research, 10, 197(1982).

2) Korn, L. J., Queen, C.L., Wegman, K.N.: Proceedings of National academic Science of USA, 74, 4401(1977)

3) Feng, D.F., Doolittle, R.F.: Journal of Molecular Evolution, 25, 351(1987)

4) Goth, O.: Journal of Molecular Biology. 162, 705(1982)

5) Mavournin, K. H., Mansfield, B.K,: human Genome News, 2. 8 (1990)

6) Ahern, K. Sequencher 1.0. Biotech. Software 9, 8-11. (1992)

7) Ahern, K. DNAstar/LaserGene. Biotech. Software 10, 6-12.(1993).

8) Ahern, K. GeneWorks 2.2.1: Biotch.software 10, 4-10(1993).

9) Ahern, K. MacVector 4.0 Biotech. Software 10, 12-19, (1993).

10) Berezney, R. J. Cell Biochem 47, 109-23.(1991).

11) Bowling, J. M., Bruner, K.L., Cmarik, J.L. and Tibbetts,

C. Nucleic Acids Res 19, 3089-97(1991).
12) Koop, B.F., Rowan, L., et al, Biotechniques 14, 441-7.(1993).
13) Pearson, W. R. and Miller, W. Methods Enzymol 210, 575-601(1992).