Evaluation of the Sequence Data Assemble Software "Shotgun"

Zhong-qing Wang[1] Yasufumi Murakami[1] Toshihiko Eki[1] Akira Oyama[2]
Yukihiro Eguchi[2] Akinori Sarai[1] Hideaki Sugawara[1]

[1] Tsukuba life Science Center, The Institute of Physical and Chemical Research(RIKEN)
3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan

[2] Mitsui Knowledge Industry Co., Ltd., Research Institute
7-4, 3-chome Kojimachi, Chiyoda-ku, Tokyo 102, Japan

1 Introduction

The program Shotgun was developed to assemble sequence data in a random sequencing method and has been proved to be an essential 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 assembled to construct one consensus sequence in the Shotgun sequencing method. Although yeast chromosomes do not contain large numbers of repetitive elements, chromosomal DNA of higher eukaryotes are fully loaded with numbers of repetitive elements. To test whether the Shotgun program is able to overcome such repetitive elements, we chose an actual genomic sequence (43kb fragment, the genomic sequence of human interferon 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 450b. 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

The DNA sequence data obtained from the DNA sequencer are about a few hundred bases long. To determine long stretch of genomic DNA, target DNA (phage clone or cosmid clone, whose size ranges between 10kb and 100kb), is cut into small pieces (1-2kb) and these pieces are cloned into the vector. Then the inserts of those recombinant clones are sequenced and the sequence data have to be assembled into one consensus sequence. The Shotgun program can connect these fragments by using the overlaps among them. In this process, the overlaping parts between 2 fragments are first searched out, and then if they satisfy the following two conditions, 1) the overlaping base sequence is long enough and 2) the matching base ratio of the overlaping 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 software consists of 4 parts: 1) Database creating program, 2) Database handling program, 3) Base sequence editing program and 4) Fragments connecting program.

3 Evaluation results

In order to evaluate Shotgun software 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 350b-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

[1]王 忠清，村上 康文，浴浴 俊彦，菅原 秀明：理化学研究所ライフサイエンス筑波センター，〒305 茨城県つくば市高野台 3-1-1
[2]大山 彰，江口 至洋：三井情報開発株式会社，〒102 東京都千代田区麹町 3-7-4

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 43 kb sequence is cut randomly into 500 fragments. 400 fragments are selected randomly from the 500 fragments 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" or "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, 3% of "A"s, "C"s, "G"s or "T"s which are located at the final 100 bases of fragments are replaced by "N"s respectively.

It took more than 20 hours on SUN SS10 for evaluation program to generate the 500 fragments which may have every kind of errors described as above. It took another 20 hours for Shotgun program to connect these 500 fragments. The table 1 shows the evaluation results of Shotgun system. Here, the length of a fragment is between 350b-450b. The connection conditions are: 1) overlaping length is more than 50b, 2) matching ratio is over 80%. Error(%) is average errors per 100 bases. The Shotgun program was evaluated by using default connection conditions(M = 50, R = 80) where M means min. overlaping length and R means % ratio of match.

From table 1, it is indicated that if no error exists in the fragments, then 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).

Table 1

Case_No.	Fragments	Connection Conditions		Islands	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

4 Conclusion

In this study, we evaluated the Shotgun program for sequence assembly by simulating the situations which usually happen in the real experiments. From our evaluation results, it could be concluded that the fragments of 350-450 b. that have about 1% errors could be connected perfectly by the Shotgun program. The target sequence used in this study contained repetitive elements including alu-repeat. The density of alu repeat was approximately one repeat in 1.5kb. The result of the careful investigation of the consensus sequence showed that there was no erroneous connection. Therefore the Shotgun program overcame the repetitive elements like alu-repeat in the assembly process.

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,: *Humen Genome News*, 2, 8(1990).

[6] Asaine, H. Eguchi, Y.,: *Information Base*, 23-4, 9(1991).