KOR. JOUR. MICROBIOL. February. 1986 p. 12-17 Vol. 24, No. 1

Rapid Plasmid Mapping Computer Program

Lee, Dong-Hun, Young-Joon Kim, Seung-Taek Lee, and Hyen-Sam Kang

Dept. of Microbiology, College of Natural Sciences, Seoul National University, Seoul 151, Korea

Plasmid의 제한효소 지도 작성을 위한 콤퓨터 프로그램

이동훈 · 김영준 · 이승택 · 강현삼

서울대학교 자연대 미생물학과

A new computer algorithm is described to order the restriction fragments of plasmid DNA which has been cleaved with several restriction endonucleases in single or double digestions rapidly with realistic error rates. The permutation and high weight on small fragments methods construct all logical circular map solutions. The program is written in Apple BASIC and run on an Apple II plus microcomputer with 64K memory. Several examples are presented which indicate the high efficiency of the program in constructing possible restriction map for YEp24.

Molecular biologists utilizing the tools of recombinant DNA technology frequently face the task of mapping DNA fragments using restriction endonucleases. But this is time-consuming, and it requires considerable effort to verify that the resulting map represents the unique solution. To avoid this, many computer programs and their associated algorithms have been developed (Polner et al., 1984; Durand et al., 1984; Nolan et al., 1984). The main advantages of using computer in DNA mapping are speed and accuracy within an imposed error range.

Stefik's program (Stefik, M., 1978) constructs the physical restriction map by permuting the fragments of the double digests. This program uses a rule system to evaluate the possible permutations. Pearson's program is described as a faster one (Peason, W.R., 1982). This program permutes the fragments of the single digestion and compares the actual double digested fragments with the hypothetical one.

In this paper, the program searches for a map in the similar way with the previous programs. But, in the previous programs, the error was calculated by giving more weight to large fragments. Therefore, in some cases, small fragments are ignored by errors of large fragments. To compensate this, we developed High Weight on Small Fragments Method (HWSFM), and errors are calculated by summing up the all absolute differences divided by actual size of double digested fragments.

Because most of programs are based on algorithms which have to establish the permutations of restriction fragments in order to deduce possible maps, they involve computation times which increase very quickly with the number of fragments to be mapped. Therefore, in this pro-

*This program is available on minifloppy disk. All requests should be addressed to Dr. Kang, H.S.

gram, it is useful to introduce some known point data to diminish unnecessary permutations from the beginning.

ALGORITHMS

The algorithm of this program is basically similar to that of Person's method, but differ in the methods of permutation and calculation of errors.

As figure 1 illustrates, some possible alignments can be made from the each single digested fragments. Then possible solutions are evaluated by calculating a hypothetical double enzyme digests and comparing this with the experimental data. The hypothetical and experimental data are compared by sorting the lengths in decreasing order. Then the squares of the differences of the largest fragment, the next largest and so on to the smallest double digested fragments are summed. This evaluation function has the advantage of giving more weight to the largest fragments. But, in the experiments the short fragments are measured more precisely than the long fragments. Thus, more weight on the small fragments is important to obtain more accurate maps where some error is unavoidable. In the ideal condition, above two methods give the same results, but in the practical experiments, data are obtained with some realistic errors.

For example, we assume that only the double digestion data have 10% errors. As figure 1 shows, Pearson's method has error range from 0 to 1.05 in the true alignment of the single digested fragments and error range from 0.64 to 5.45 in incorrect alignment. However, given more weight on the small fragments, the error range of incorrect alignment is 0.364-0.959 and that of correct alignment is 0-0.333.

The maximum error of the correct solution is less than the minimum error of the incorrect solution. Therefore, in laboratory, our method is more efficient.

Example 1.

A. Permutation generation
X I---I I--------I 1kb, 12kb
Y I-------I I----I 10kb, 3kb

B. Calculation of hypothetical double digest fragments
X/Y I---I I---------I I----I 1kb, 9kb, 3kb

C. Comparison of hypothetical and actual double digest fragments
Hypothetical: I------I I-----I I---I 9kb, 3kb, 1kb
Acutal : I-------I I----I I---I 10 ± 1kb, 2 ± 0.2kb, 1 ± 0.1kb

D. Error calculation
(1) Pearson's method
i) minimum error: $(9\text{-}9)^2 + (3\text{-}2.2)^2 + (1\text{-}1)^2 = 0.64$
ii) maximum error: $(9\text{-}11)^2 + (3\text{-}1.8)^2 + (1\text{-}0.9)^2 = 5.45$
(2) High weight on small fragments method
i) minimum error: $\frac{|9\text{-}9|}{9} + \frac{|2\text{-}2.2|}{2.2} + \frac{|1\text{-}1|}{1} = 0.364$
ii) maximum error: $\frac{|9\text{-}11|}{11} + \frac{|3\text{-}1.8|}{1.8} + \frac{|1\text{-}0.9|}{0.9} = 0.959$

Example 2.

A. X I---------I I---I 12kb, 1kb
Y I--------I I-----I 10kb, 3kb

B. X/Y I--------I I----I I---I 10kb, 2kb, 1kb

C. Hypothetical: I-------I I-----I I---I 10kb, 2kb, 1kb
Actual : I-------I I-----I I---I 10 ± 1kb, 2 ± 0.2kb, 1 ± 0.1b

D. (1) Pearson's method
i) minimum error: $(10\text{-}10)^2 + (2\text{-}2)^2 + (1\text{-}1)^2 = 0$
ii) maximum error: $(10\text{-}9)^2 + (2\text{-}1.8)^2 + (1\text{-}0.9)^2 = 1.05$
(2) High weight on small fragments method
i) minimum error: $\frac{|10\text{-}10|}{10} + \frac{|2\text{-}2|}{2} + \frac{|1\text{-}1|}{1} = 0$
ii) maximum error: $\frac{|10\text{-}9|}{9} + \frac{|2\text{-}1.8|}{1.8} + \frac{|1\text{-}0.9|}{0.9} = 0.333$

Fig. 1. An algorithm for mapping a circular plasmid. A simple example of a linear molecule is shown. The unknown DNA has been digested with two enzymes X and Y, and with X and Y together. (A) The two possible permutations of the Y digest are compared with the X digest. Only one permutation of the X digest is considered since the second is just the reverse of the first. (B) Hypothetical X and Y double digests are calculated for each of the possible maps. (C) The double digested fragments are sorted from the largest to the smallest and compared with the double digested data. The actual digestion data are considered that they have 10% error. (D-1) Pearson's method: The squared errors are calculated. (D-2) High weight on small fragment method (HWSFM): The errors are calculated by summing up the all absolute differences divided by the actual size of double digestion fragments. This error calculation emphasizes on the position of small fragment. The solution of example 2 has a lower error and is judged correct.

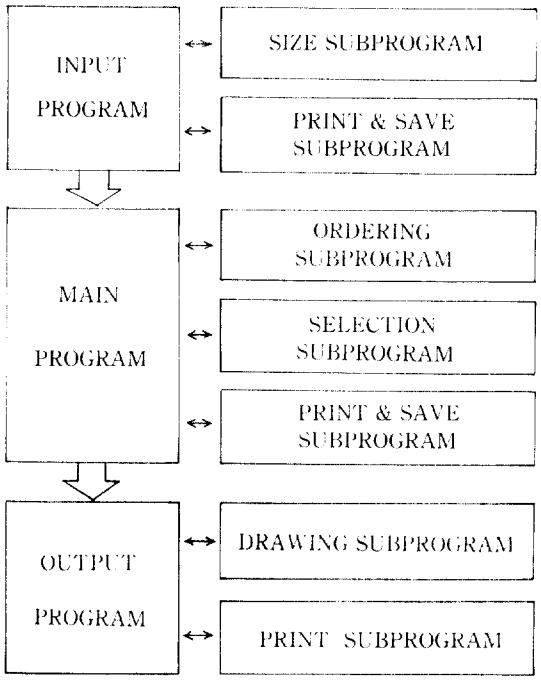


Fig. 2. Block diagram of rapid plasmid mapping program. The program is composed of 3 parts; input program, main program and output program.

PROGRAM DESCRIPTION

This program has three major parts; (1) input program (2) main program (3) output program. Each program has several subprograms that have various functions. The block diagram of the program is shown in figure 2. Although the program is divided into three parts, user can utilize the program as a whole. In that case, each following program is called automatically by the preceeding program and the results are saved in the disk file. Therefore the user can save time and efforts to run the every programs.

Input program

Data obtained from the agarose gel electrophoresis are entered this program to construct the possible restriction map. In order to deduce the possible restriction map, each single digestion data needs at least one related double digestion data . To construct the possible restriction map with n enzymes, this program needs n single digestion data and at least n-1 double digestion data.

Input program operates in two ways : (1) The user can input the length of the fragments, or (2) The user can input the mobility of the fragments on the gel electrophoresis. In the case (2), user must input the standard data for the calculation of the lengths. There are equations and program to determine the length of the fragments of a DNA molecule digested with a restriction endonucleases with mobilities on the argarose gel (Southern, E., 1979; Maina et al., 1984). This is done in the size subprogram and it changes the entered mobility to the size of the fragments.

To save and edit data, the user use save & print subprogram. This subprogram can print the input data as well as save them in the applicable format to main program. Another helpful thing is that the user can enter the special known data. These data are informations of the known region of the plasmid used for experiments ; such as enzyme names and sites that user knew already. The special known data are very important because they affect the speed of main program. The user can save the special known data of the common vector and can use them by only entering the vector name.

All data entered in the program are saved in disk files and user can insert, delete and change the old data file and save the edited data as a new file name.

Main program

This program searches a possible restriction map. The process is shown below;

(1) If the special known data are introduced to this program, the vector region is fixed. Then the program compares the data come from the input program with special known data and selects the fragments that have the minimum difference with the known region. The more the special known data, the less the number of permutation. Because the number of permutation increase as factorial, the speed of main program is dramatically increased by the special known data. If there are no special known data, the program assumes that the site of the first entered enzyme is 0, and the next site of the other enzyme is determinded by using the size of the largest double digested fragment

that doesn't appear in single digestion data.

(2) Then, this program permute the rest fragments to construct the possible restriction map. In the case that the number of single digested fragments of enzyme A is Na and that of enzyme B is Nb, the number of permutation with the data of enzyme A and enzyme B is Na! × Nb!. Program doesn't use the matrix, but the subprogram permutes the order of fragments rapidly. Because this method doesn't need much memories for matrix, this program is available with microcomputer.

(3) By checking consistantly with the known data, computer abandon the permutation result when the combination of fragments with this permutation order is opposed to the known data and go to the next permutation order.

(4) By repeating process (2) and (3), computer make and save 5 possible maps according to the order of error size.

(5) During the process, each step can be monitored. The dependent process is more useful for rapid searching. In this process, some enzyme sites determined during the ealry process are used directly in the later processing. Besides this, restriction map can be obtained independently to avoid confusion between each double digestion data.

Output program

The user runs the output program to confirm the results of the main program. The output program shows all possible restriction map and their errors. The user can construct total restriction map by entering each figure numbers and then map is appeared on the screen. The user can add any figures to map to combine or delete any figures from the combined map. Therefore, the user can make any combination of figures of the figure catalog. This program can display the results in circular form as showed in other programs (Nolan et al., 1984; Stone et al., 1984). The ability making any combination of figures and the circular graphic result make it easy for user to find the possible restriction map. The circular graphic output can be displayed with maximum 8 enzymes data and the vector region is indicated with thick curve at the most inside circle.

PROGRAM USE

Sample run was done for ascertaining a normal run of this program with YEp24 (Tho-e, A., 1982).

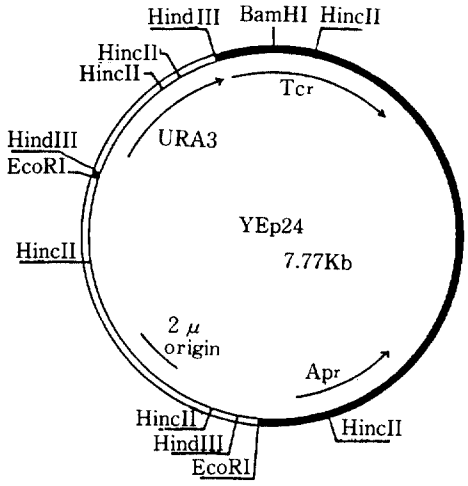
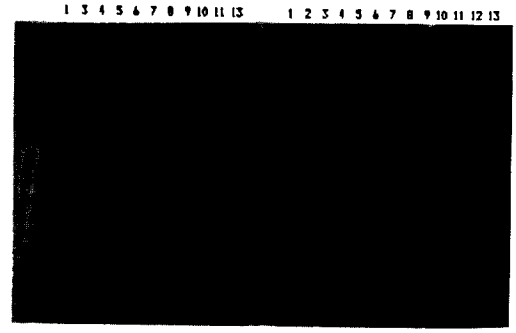


Fig. 3. The circular map of YEp24.
YEp24 is a yeast vector, constructed by inserting URA 3 fragment into HindII site and 2μ origin into EcoRI site of pBR322. It has 3 sites for HindIII, 2 sites for EcoRI, 6 sites for HincII and one site for BamHI.



I; 0.8% Agarose Gel II; 1.5% Agarose Gel

Fig. 4. Digestion pattern of YEp24.
YEp24 was single and double digested with BamHI, EcoRI, HindIII and HincII and electrophoresed on 0.8% and 1.5% agarose gels.
Lane 1, 13: λ-HindIII, EcoRI. Lane 2, 12: pBR322-HinfI. Lane 3: YEp24-BamHI Lane 4: BamHI, EcoRI Lane 5: EcoRI Lane 6: EcoRI, HindIII Lane 7: HindIII Lane 8: HincII Lane 9: HincII, BamHI Lane 10: HincII, EcoRI Lane 11: HincII, HindIII

Table 1. *The sizes of each fragments were calculated by the input program on the basis of the mobility of each fragments comapared with that of reference molecule. The real valeus were shown in parenthesis These calculated values are plausible within 5% error.*

Enzyme	Fragment Size	Total Size	Error (%)
BamH I	7800 (7772)	7800	0.36
EcoR I	5410 (5529), 2173 (2243)	7583	2.43
Hind III	4500 (4436), 2059 (2169), 1113 (1167)	7672	1.29
Hinc II	3200 (3256), 1423 (1469), 1266 (1319), 918 (921), 666 (671), 139 (136)	7612	2.06
BamH I / EcoR I	3889 (3935), 2173 (2243), 1500 (1544)	7512	3.35
BamH I / Hinc II	3200 (3256), 1423 (1469), 1266 (1319), 670 (671), 642 (646), 280 (275), 139 (136)	7620	1.96
EcoR I / Hind III	4157 (4331), 2045 (2138), 1113 (1167), 108 (105), 30 (31)	7453	4.10
EcoR I / Hinc II	3200 (3256), 1423 (1469), 918 (921), 760 (762), 560 (557), 460 (454), 212 (217), 139 (136)	7672	1.29
Hinc II / Hind III	3200 (3256), 1423 (1469), 741 (731), 630 (621), 595 (588), 566 (559), 285 (300), 139 (136), 115 (112)	7694	1.00

YEp24 of which restriction map is shown to figure 3 is a yeast vector, constructed by inserting URA 3 fragment into HindIII site and 2μ origin fragment into EcoRI site of pBR322. It has 3 sites for HindIII, 2 sites for EcoRI, 6 sites for HincII and a single site for BamHI.

After YEp24 was single and double digested with various restriction endonucleases, their digestion mixtures and reference fragments were electrophoresed on the agarose gels to obtain the mobility and pictured by Polaroid camera (Figure 4).

The mobility of each lanes was measured and the sizes of each fragments were calculated by the input program on the basis of the mobility and size of the reference fragments.

These results (Table 1) are plausible within 5% error range when compared with the published data. The sizes calculated by the input program transfered to the main program. Data from two single digested and one double digested results were analysed and drawn to circular map in the main program. Figure 5 is the YEp24 circular map, combined each circular map of two restriction endonucleases which has the lowest error in output program. This result was proved to be the same as figure 3, the published map of YEp24.

***** RESTRICTION MAP OF YEp24 *****

VECTOR=pBR322 RANGE= -346/3985
FIG. NO.= 1,3,4,9,12,
(1) BamHI SITE= 1
(1) 0
(2) EcoRI SITE= 2
(2) 3985 (2) 6158
(3) HindIII SITE= 3
(1) -345 (2) 4154 (3) 6213
(4) HincII SITE= 6
(1) 275 (2) 3531 (3) 4197 (4) 5620 (5) 6886 (6) 7025

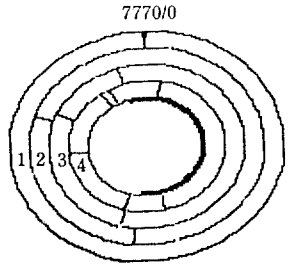


Fig. 5. YEp24 circular map constructed by computer program. BamHI site was given as 0 point and the bases were counted clockwise. The known point, that is, vector range was indicated as thick curve at the most inside circle. Each enzyme sites was indicated as bars in the circle and its position was shown as number of bases from BamHI site.

DISCUSSION

Some problems arise in using Pearson's progrm to construct a restriction map, because the researcher is usually unfamiliar with computers. Thus, it is necessary to develope the program which can be used easily in any labor atories with only a few equipments. For this reason, this program was written in Apple BASIC and run on

microcomputer AppleII plus wiht 64K memory.

A limitation in using micrcomputer is small memory capacity, but this is not concerned in this program because computer permutes the fragments one by one and saves only some valuable combinations in the memory without saving full matrix. Beside this, known point values can be entered directly to the input program and also cut down the number of permutations, thus main program needs only a few minutes of running time and small memory capacity.

This program is emphasized on accuracy as well as speed. By giving more weight on the small fragments, we could construct the restriction maps more precisely from many experimental data. In experiments, sometimes small fragments are not detected on agarose gel and this often lead to counting the single or double digested fragments less than the real numbers. This input data are not accepted by input program and "number of fragments is mismatched" appears on the screen. Therefore all fragments must be identified before applying to the computer analysis.

This program can be applied to the restriction map of linear molecule with a small modification. We hope to continue to develope the programs for the molecular genetic researchers. Currently programs are being developed for complete anlysis of nucleic acids.

적 요

Plasmid의 원형 제한효소 지도를 작성하는 computer program을 개발하였다. 실험시 나타나는 오차를 극복할 수 있으며, 모든 경우의 수를 비교하면서도 제한효소 절단시 생기는 작은 절편에 중점을 두어 정확한 제한효소 지도를 작성하였다. 이 program은 Apple BASIC으로 작성되었으며, 64K memory를 갖는 Apple Ⅱ plus 소형 computer로 작동된다. YEp 24를 사용하여 program을 작동시켜 본 결과 매우 효율적이면서도 정확한 원형 제한효소 지도를 얻을 수 있었다.

REFERENCES

1. Durand, R., and F. Bregegere, 1984. An efficient program to construct restriction maps from experimental data with realistic error levels, *Nucl. Acids Res.*, **12** 703-716
2. Maina, C.V., G.P. Nolan, and A.A. Szalay, 1984. Molecular weight determination program, *Nucl. Acids Res.*, **12**, 695-702
3. Nolan, G.P., C.V. Maina, and A.A. Szalay, 1984. Plasmid mapping computer program, *Nucl. Acids Res.*, **12**, 717-729
4. Pearson, W.R., 1982. Automatic construction of restriction site maps. *Nucl. acids Res.*, **10**, 217-227
5. Polner, G., L. Dorgai, and L. Orosz, 1984. PMAP, PMAPS: DNA physical map construc-tig program, *Nucl. Acids Res.*, **12**, 227-236
6. Southern, E., 1979. Gel electrophoresis of restriction fragments *in* Methods in Enzymology, Edited by Wu, R., 1979. **68**, 152-176, Academic Press, New York.
7. Stefik, M., 1978. *Artificial Intelligence*, **11**, 85-144.
8. Stone, B.N., G.L. Griesinger, and J.L. Modelevsky, 1984. PLASMAP: an interactive computational tool for storage, retrieval and device-independent graphic display of conventional restriction maps, *Nucl Acids Res.*, **12**, 465-471
9. 東江 昭夫 1982, 酵母宿主 Vector改良 in 遺伝子組換実用化技術, vol. 2, 124－142.

(Received Nov. 11, 1985)