Volume 14 Number 1 1986 Nucleic Acids Research

A computer program package for restriction map analysis and manipulation

Guenther Zehetner and Hans Lehrach

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, FRG

Received 6 June 1985

ABSTRACT

Programs for the calculation, storage and analysis of restriction maps derived from the analysis of partial digestion products from end labelled DNA (1,2,3) and their correlation with digestion - and hybridisation patterns in total digestions and Southern blot experiments are described. These programs allow direct input of gel patterns from partial or complete digestion experiments using a digitizer tablet, calculation of molecular weights and restriction maps, plotting of maps and actual or predicted fragment patterns and automated identification of overlapping cosmids from partial restriction mapping results. Programs are written in PASCAL and have been implemented on a VAX/VMS system, with a HP-7221T plotter and a digitizing tablet.

INTRODUCTION

Determination and analysis of restriction maps is an essential step in the analysis of cloned DNA sequences. In spite of constant improvements in the speed of sequence analysis, restriction mapping is the first and usually only technique to generate physical maps of long regions of DNA. We have developed an efficient protocol to derive restriction maps of sequences cloned in lambda (2) or cosmid vectors (3) by linearisation at the cos sequence, partial digestion with a restriction enzyme and identification of partial digestion products originating from either the right or the left end of the clone by hybridisation to chemically synthesised oligonucleotides complementary to the protruding end sequence. This approach is experimentally very fast, and well suited for

Programs will be provided for non-commercial use upon receipt of a self-addressed mailing label and a blank tape. A small charge will be requested to cover mailing and processing.

the application of computer techniques to acquire, store and analyse the resulting restriction maps. Ease of experimental protocol and maximal use of computer techniques in the analysis, will be essential for the determination of physical maps of large regions of genomes.

Analogous to the analysis of long DNA sequences, computer programs can greatly simplify the work encountered in manually analysing and interpreting restriction maps and in combining them with information on the distribution of repetitive and transcribed regions. In contrast to programs designed to handle DNA sequences, programs for manipulating and comparing restriction map data also have to take into account many more uncertainties and most importantly inaccurate distances between sites.

Steps in the analysis are:

1) Reading of gel data from partial (or complete) digestion experiments into the computer using a digitizer tablet, determination of molecular weights of digestion products from their mobilities relative to appropriate marker fragments (program FRASI)
2) Calculation of restriction maps by combining partial mapping results from the right and left ends of the clone (program FIMA)
3) Drawing of restriction maps (program DRAWMAP)
4) Comparison and linkup of restriction maps (program COMP)
5) Calculation of predicted single and double digestion patterns and comparison with observed patterns (programs MTRANS and DRAWGEL)
6) Modification of restriction maps by deleting, adding and replacing sites or regions of the maps (programs MTRANS and DRAWMAP)

The connection between the programs described here and their input and output files are shown in Fig. 1. A program called REMA manages the files and selects the used programs.

A) CALCULATION OF FRAGMENT SIZES

PROGRAM FRASI

The program uses the algorithm of Southern (4,5,6) to

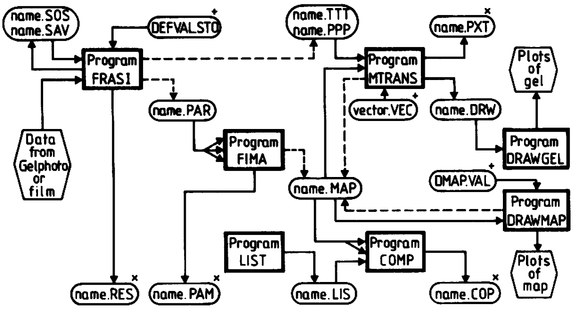


Figure 1. Interplay of used programs and files.
The shown file extensions are the default values added to file names entered without a extension by the programs. Broken lines indicate optional file creation.
"x" Files which are automatically printed and then deleted.
"+" Files are read by the program to get preset values.

calculate fragment lengths from mobilities. For fragments migrating faster than the smallest marker band a logarithmic equation is used. FRASI allows the use of multiple marker slots with different markers on the same gel and calculates sizes by interpolating between the two flanking marker lanes to correct for slight distortions during the run. To take into account the uncertainty in mobility measurement upper and lower limits are calculated for each size assuming a certain range around the measured migration. The range can be chosen according to the spread of each band. To avoid loss of data after an accidentally or intentional interuption all input is saved in a file with the extension name.SOS (renamed to name.SAV after normal termination). The progress of entering mobility data from a digitizer tablet can be followed at the terminal.

Input files

DEFVAL.STO FRASI first reads values defining properties of the digitizer and terminal from this file

name.SOS/name.SAV data from a previously entered gel can be read from these files into the program and they can serve as startpoint to repeat any of the further analysis steps

Output files

name.SOS/name.SAV see input files of program FRASI name.RES contains results of size calculation

name.TTT/name.PPP if indicated a file with the extension name.TTT (for 'total digestion') or name.PPP (for 'partial mapping') is stored which is used to create a file for drawing the gel

name.PAR in case of 'partial mapping' data a file with this extension can be generated which is than used to derive a restriction map

B) CREATION OF A RESTRICTION MAP

PROGRAM FIMA

Usually not all restriction site positions can be determined from both DNA ends in 'partial mapping' experiments. Sometimes one of the two corresponding partial fragment bands, obtained by cutting at one site and labelling either the left or right end of the DNA, is missing from the film. For the single partner of such a fragment no counterpart can be found to fulfill equation [1] ('single' site). But in most cases both of the fragments are present and can be used to determine the site position ('double' site).

$$S[i] = LF[j] = TL - RF[k] \qquad [1]$$

S = site position

LF = length of fragment labelled on left DNA end

RF = length of fragment labelled on right DNA end

TL = total DNA length

EQ [1] requires also the total DNA length to assign the correct fragment pairs. Since the direct size determination of large fragments from a gel is inaccurate the program determines also the length of the total DNA. FIMA tests within a predetermined range all possible length values by establishing a restriction map with each value and evaluating it using equation [2] . If more than one neighbouring LF[j] and RF[k] fulfill EQ [1] the program searches for pairings which gave rise to the largest total number of 'double' sites regardless of their accuracy as long as they lie within defined limits.

$$G = \text{difsum} * (\text{single} / \text{double})^4 \qquad [2]$$

G = measure of reliability of the used total length

difsum = sum of all differences between site positions found from both DNA ends for one site

single = number of sites found only from one DNA end

double = number of sites found from both DNA ends

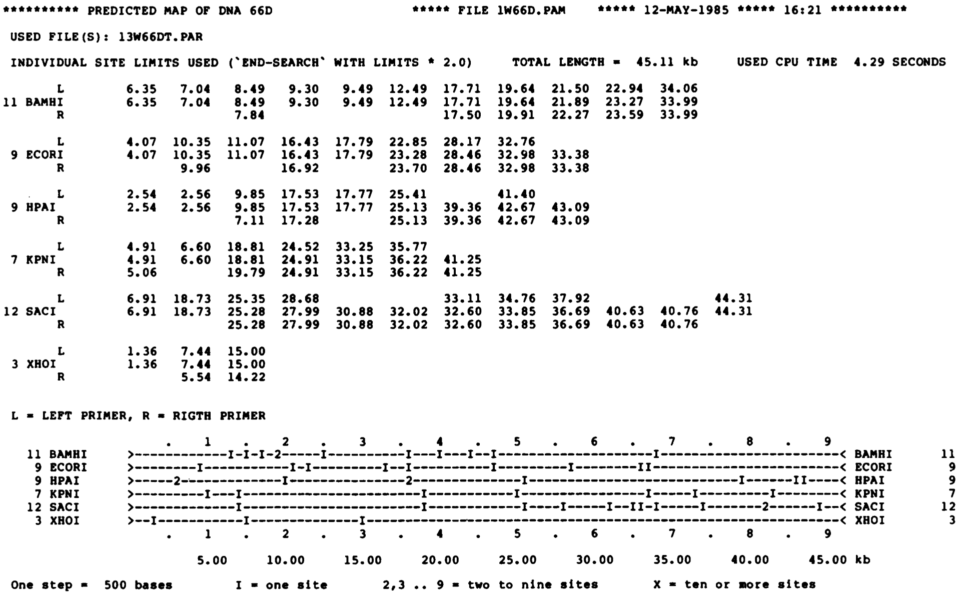
This is done by a stepwise increase in the total length until G passes a minimum. In a closer search around the first minimum the length is defined to an accuracy of 10 bases. FIMA shows the resulting optimal total length (which can be accepted or replaced by a newly entered value) which is then used to calculate the final map.

Input files

name.PAR see program FRASI

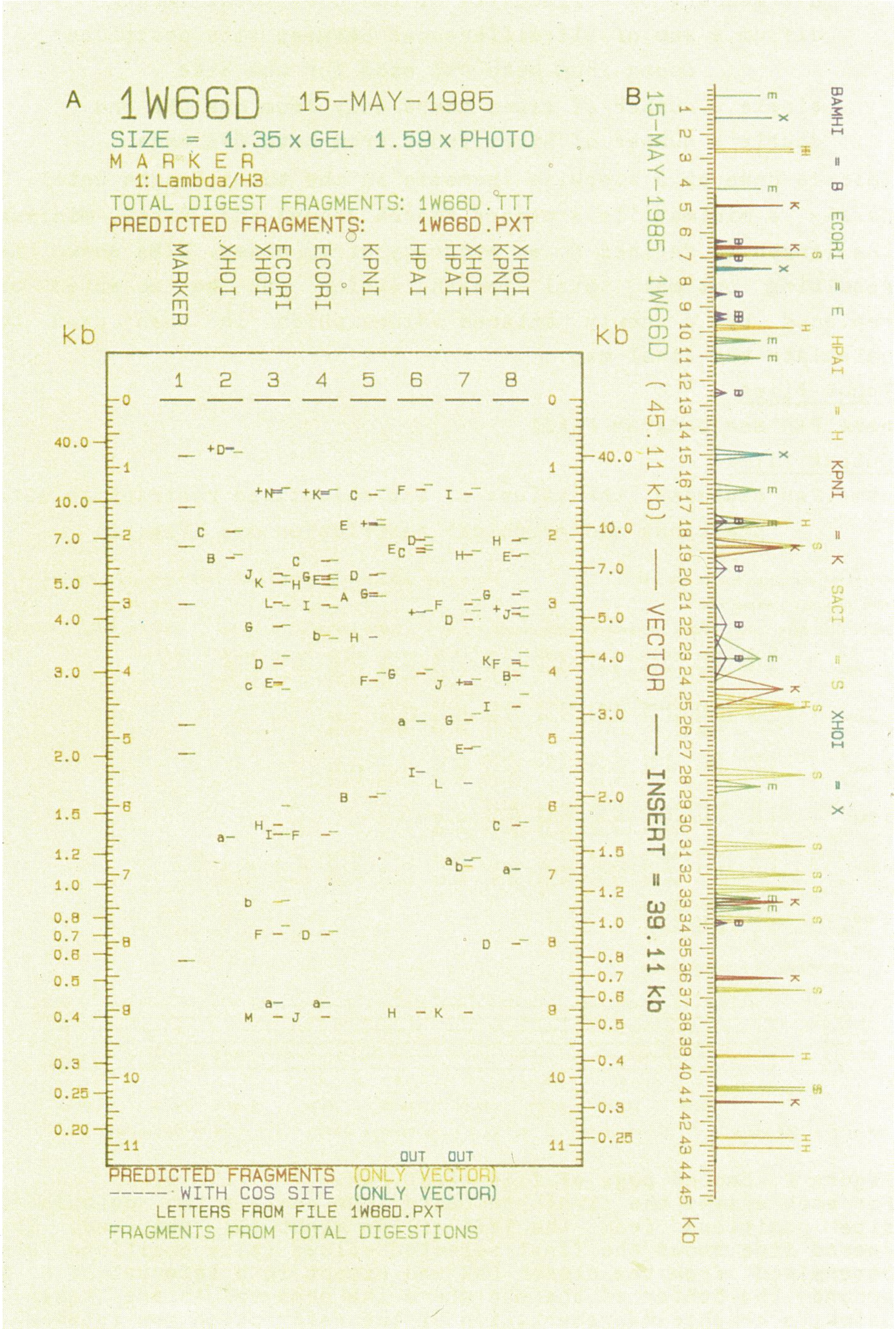
Output files

name.PAM contains the values of the calculated restriction site positions and graphical restriction map (Fig. 2)



```
********** PREDICTED MAP OF DNA 66D              ***** FILE 1W66D.PAM    ***** 12-MAY-1985 ***** 16:21 **********

 USED FILE(S): 13W66DT.PAR

 INDIVIDUAL SITE LIMITS USED (`END-SEARCH` WITH LIMITS * 2.0)     TOTAL LENGTH =  45.11 kb     USED CPU TIME  4.29 SECONDS

       L      6.35   7.04   8.49   9.30   9.49  12.49  17.71  19.64  21.50  22.94  34.06
11 BAMHI      6.35   7.04   8.49   9.30   9.49  12.49  17.71  19.64  21.89  23.27  33.99
       R                    7.84                        17.50  19.91  22.27  23.59  33.99

       L      4.07  10.35  11.07  16.43  17.79  22.85  28.17  32.76
 9 ECORI      4.07  10.35  11.07  16.43  17.79  23.28  28.46  32.98  33.38
       R             9.96         16.92         23.70  28.46  32.98  33.38

       L      2.54   2.56   9.85  17.53  17.77  25.41         41.40
 9 HPAI       2.54   2.56   9.85  17.53  17.77  25.13  39.36  42.67  43.09
       R                    7.11  17.28         25.13  39.36  42.67  43.09

       L      4.91   6.60  18.81  24.52  33.25  35.77
 7 KPNI       4.91   6.60  18.81  24.91  33.15  36.22  41.25
       R      5.06         19.79  24.91  33.15  36.22  41.25

       L      6.91  18.73  25.35  28.68                33.11  34.76  37.92                44.31
12 SACI       6.91  18.73  25.28  27.99  30.88  32.02  32.60  33.85  36.69  40.63  40.76  44.31
       R                   25.28  27.99  30.88  32.02  32.60  33.85  36.69  40.63  40.76

       L      1.36   7.44  15.00
 3 XHOI       1.36   7.44  15.00
       R             5.54  14.22

 L = LEFT PRIMER, R = RIGTH PRIMER

                      .    1    .    2    .    3    .    4    .    5    .    6    .    7    .    8    .    9
   11 BAMHI      >------------I-I-I-2-----I----------I---I---I--I--------------------I-----------------------< BAMHI      11
    9 ECORI      >--------I-----------I-I---------I--I----------I---------I--------II------------------------< ECORI       9
    9 HPAI       >-----2-------------I---------------2--------------I---------------------------I------II----< HPAI        9
    7 KPNI       >---------I---I-----------------------I-----------I----------------I-----I---------I--------< KPNI        7
   12 SACI       >-------------I-----------------------I------------I----I-----I--II-I-----I-------2------I--< SACI       12
    3 XHOI       >--I-----------I--------------I-------------------------------------------------------------< XHOI        3
                      .    1    .    2    .    3    .    4    .    5    .    6    .    7    .    8    .    9

                         5.00      10.00     15.00     20.00     25.00     30.00     35.00     40.00     45.00 kb

 One step =  500 bases        I = one site       2,3 .. 9 = two to nine sites       X = ten or more sites
```

Figure 2. Second page of file 1W66D.PAM.
For each enzyme the first and third line contain the calculated site positions from the left (L) or right (R) DNA end. The second line shows the final selected values (site positions are determined from the closer DNA end except in a interval of 5 % around the center of the map where the mean values are taken). Below, a graphic representation of the restriction map is shown.


A
1W66D 15-MAY-1985
SIZE = 1.35 x GEL 1.59 x PHOTO
MARKER
1: Lambda/H3
TOTAL DIGEST FRAGMENTS: 1W66D.TTT
PREDICTED FRAGMENTS: 1W66D.PXT
MARKER
XHOI
XHOI ECORI
ECORI
KPNI
HPAI
HPAI XHOI
KPNI XHOI
kb
1 2 3 4 5 6 7 8
40.0
10.0
7.0
5.0
4.0
3.0
2.0
1.5
1.2
1.0
0.8
0.7
0.6
0.5
0.4
0.3
0.25
0.20
OUT OUT
PREDICTED FRAGMENTS (ONLY VECTOR)
------ WITH COS SITE (ONLY VECTOR)
LETTERS FROM FILE 1W66D.PXT
FRAGMENTS FROM TOTAL DIGESTIONS
B
15-MAY-1985 1W66D (45.11 kb) — VECTOR — INSERT = 39.11 kb
BAMHI = B ECORI = E HPAI = H KPNI = K SACI = S XHOI = X
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 kb

name.MAP contains commentary lines, total DNA length, length of vector arms, restriction site positions with upper and lower limits, and a code number for the used enzyme

C) DRAWING OF MAPS AND GELS

Three programs are involved in this step, MTRANS, DRAWMAP and DRAWGEL.

PROGRAM MTRANS

The program MTRANS serves to prepare predicted and real gel data for plotting. Restriction maps can be modified by deleting single restriction sites or all sites of an enzyme at once or by adding sites to the map. The program can also replace the region of the restriction map corresponding to the vector part by the theoretical map derived from the vector DNA sequence. Any enzyme in the restriction map can be used to predict fragments of single or double digestions. A 'theoretical' gel can be created assigning these digestions to specified slots or copying automatically the slot configuration of a real gel. A letter is assigned to each fragment marking its position within the map (Fig. 4). The fragments are then arranged by decreasing size (Fig. 5). Using the gel parameters of a real gel

Figure 3. (A) Plot of a gel with program DRAWGEL showing original and predicted fragment patterns of digestions of clone 1W66Dn.
The drawn gel was derived by program MTRANS from an original gel with fifteen slots and the map shown in Fig. 3 B. Slot one contains lambda DNA digested with *Hind* III as size marker. Green bands on the right half of the slots represent original fragments. Predicted fragments (on the left half of the slots) contain either insert DNA (red bands) or vector DNA only (brown bands). Violet (insert DNA) and blue (vector DNA only) colour marks fragments containing the cos site used for linearization in the partial mapping procedure. Only one band (corresponding to the biggest violet or blue predicted fragment) can be present in the original digest pattern since the cosmids are circular molecules and the first and last fragment in the map are combined over the cos site (see also Fig. 5). The letters on the left side of each predicted fragment refer to its map position as shown in Fig. 4. 'OUT' at the bottom of a slot shows that one or more predicted fragments have a mobility greater than the length of the gel.
(B) Map of clone 1W66D drawn with program DRAWMAP showing also the limits of the ranges of site positions.
The middle line indicates the calculated site position.

(determined by program FRASI) the theoretical mobilities of the predicted fragments on that gel can be calculated using the same algorithms as in FRASI but rewritten for unknown mobilities and known sizes. Observed and predicted fragment patterns can be stored for drawing by the program DRAWGEL.

Input files

name.MAP see programs FIMA, MTRANS, DRAWMAP

name.TTT/name.PPP see program FRASI

Output files

name.PXT contains map where the fragments are marked by letters (Fig. 4,5) (created only if map data are entered)

name.DRW contains data of real digestion fragments and/or predicted fragments (created only if gel data are entered)

name.MAP new, modified version of entered name.MAP file

PROGRAM DRAWMAP

This program draws the calculated restriction map in normal or inverted form. Vector arms and insert are drawn in different colours. The restriction site positions are indicated by lines. Each enzyme has a characteristic height and colour of the line, and a one letter abbreviation. These are stored in a file called DMAP.VAL. Maps showing the limits of the calculated positions of the restriction sites can also be drawn (Fig. 3 A). The position of the map on the paper can be chosen freely to allow the alignment of maps of overlapping clones (Fig. 6 B). For the analysis and comparison of different restriction maps a map can be modified before drawing by 1) deletion, 2) insertion, 3) inversion, 4) normal or inverted duplication of specified DNA regions. An unlimited number of these modifications can be applied consecutively. Positions for modifications can refer either to the original or the modified map.

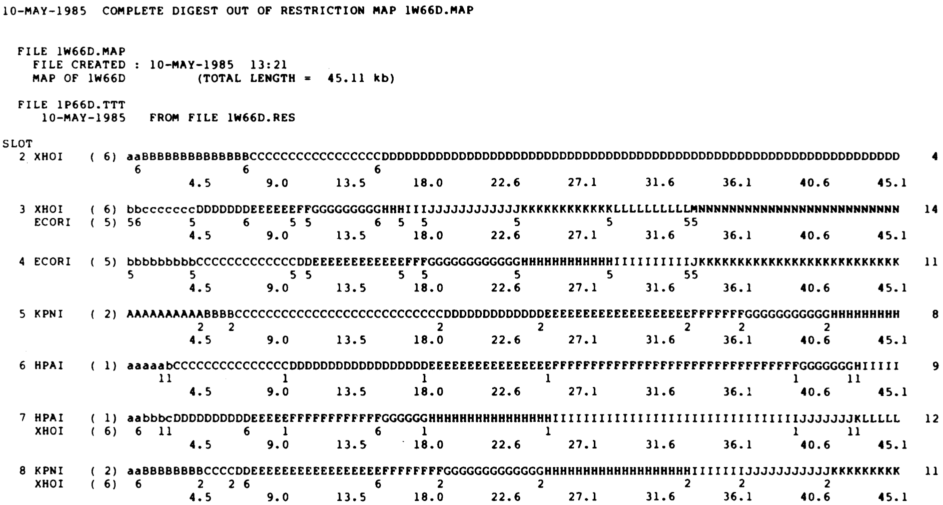
Input files

DMAP.VAL contains values for length and colour of lines symbolising restriction sites and one letter abbrevation for each enzyme

name.MAP see programs FIMA, MTRANS, DRAWMAP

Output file

name.MAP file in which a modified map is stored

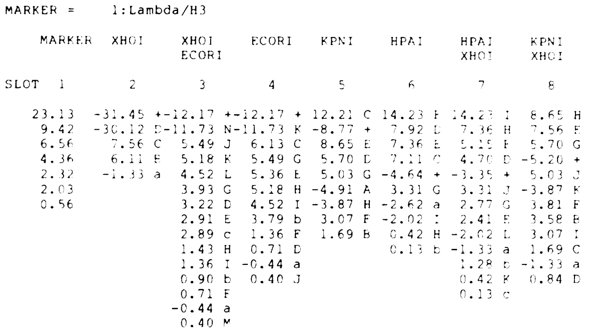


```
10-MAY-1985  COMPLETE DIGEST OUT OF RESTRICTION MAP 1W66D.MAP

  FILE 1W66D.MAP
    FILE CREATED : 10-MAY-1985  13:21
    MAP OF 1W66D          (TOTAL LENGTH =  45.11 kb)

  FILE 1P66D.TTT
     10-MAY-1985   FROM FILE 1W66D.RES

SLOT
  2 XHOI   ( 6) aaBBBBBBBBBBBBBBCCCCCCCCCCCCCCCCCDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD    4
                 6             6                6
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  3 XHOI   ( 6) bbcccccccDDDDDDDEEEEEEFFGGGGGGGGGHHHIIJJJJJJJJJJJJKKKKKKKKKKKKKKLLLLLLLLLMNNNNNNNNNNNNNNNNNNNNNNNNNN   14
    ECORI  ( 5) 56      5      6     5 5        6  5  5           5           5         55
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  4 ECORI  ( 5) bbbbbbbbbCCCCCCCCCCCCCDDEEEEEEEEEEEEEFFGGGGGGGGGGGGGHHHHHHHHHHHHIIIIIIIIIIJKKKKKKKKKKKKKKKKKKKKKKKKK   11
                5       5            5 5           5  5           5           5         55
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  5 KPNI   ( 2) AAAAAAAAAABBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCDDDDDDDDDDDDDEEEEEEEEEEEEEEEEEEEEFFFFFFFGGGGGGGGGGGHHHHHHHH    8
                         2   2                          2            2                  2      2          2
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  6 HPAI   ( 1) aaaaabCCCCCCCCCCCCCCCDDDDDDDDDDDDDDDDDEEEEEEEEEEEEEEEEEFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFGGGGGGGHIIIII    9
                    11              1                 1               1                               1      11
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  7 HPAI   ( 1) aabbbcDDDDDDDDDDEEEEEEFFFFFFFFFFFFGGGGGGHHHHHHHHHHHHHHHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIJJJJJJKLLLLL   12
    XHOI   ( 6)  6  11         6    1           6     1               1                               1      11
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1

  8 KPNI   ( 2) aaBBBBBBBBCCCCDDEEEEEEEEEEEEEEEEEFFFFFFFFGGGGGGGGGGGGGHHHHHHHHHHHHHHHHHHHIIIIIIJJJJJJJJJJJJJKKKKKKKK   11
    XHOI   ( 6)  6       2   2 6                6       2            2                  2      2          2
                        4.5       9.0      13.5      18.0      22.6      27.1      31.6      36.1      40.6      45.1
```

Figure 4. First page of file 1W66D.PXT showing the map of clone 1W66D with letters assigned to each predicted fragment indicating its position within the map.
Lower case letters stand for fragments only containing vector DNA, upper case letters for those containing also insert DNA. Numbers beneath show enzyme and site positions.

PROGRAM DRAWGEL

This program draws patterns of predicted and original restriction digest fragments of gels in different sizes. To simplify the assignment of labelled bands detected on autoradiograms to the corresponding bands on stained gels, the fragment pattern previously entered from a photograph of the gel to the program FRASI is drawn in the size of the gel. Superposition of the X-ray film and the gel drawing identifies the hybridizing bands. Visual inspection of the gel plot showing the original together with the fragments predicted from the restriction map (including the letters pointing to the positions of the fragments in the map) allows the identification of the two corresponding fragments (Fig. 3 A). Using the indicating letters the map position of the original fragments can be easily determined in the corresponding restriction map created by the program MTRANS (Fig. 4). In addition all bands on a gel drawing can be labelled with their size. Mobilities can be drawn also with error limits determined either from the



```
MARKER =      1:Lambda/H3

    MARKER  XHOI      XHOI     ECORI    KPNI      HPAI     HPAI     KPNI
                      ECORI                                XHOI     XHOI

SLOT  1        2         3        4        5         6        7        8

   23.13  -31.45 +-12.17 +-12.17 + 12.21 C 14.23 F 14.23 I  8.65 H
    9.42  -30.12 D-11.73 N-11.73 K -8.77 +  7.92 D  7.36 H  7.56 E
    6.56    7.56 C  5.49 J  6.13 C  8.65 E  7.36 E  5.15 F  5.70 G
    4.36    6.11 B  5.18 K  5.49 G  5.70 D  7.11 C  4.70 D -5.20 +
    2.32   -1.33 a  4.52 L  5.36 E  5.03 G -4.64 + -3.35 +  5.03 J
    2.03            3.93 G  5.18 H -4.91 A  3.31 G  3.31 J -3.87 K
    0.56            3.22 D  4.52 I -3.87 H -2.62 a  2.77 G  3.81 F
                    2.91 E  3.79 b  3.07 F -2.02 I  2.41 E  3.58 B
                    2.89 c  1.36 F  1.69 B  0.42 H -2.02 L  3.07 I
                    1.43 H  0.71 D          0.13 b -1.33 a  1.69 C
                    1.36 I -0.44 a                  1.28 b -1.33 a
                    0.90 b  0.40 J                  0.42 K  0.84 D
                    0.71 F                          0.13 c
                   -0.44 a
                    0.40 M
```

Figure 5. Second page of file 1W66D.PXT showing the sizes of the predicted fragments in the order in which they appear on the gel composed in program MTRANS.
Three fragments in each digestion (the first and the last in the map and a new one combining both) are preceded by a minus sign. The newly generated fragment replaces the two smaller ones if the DNA has not been linearized at the cos-site. It is marked by a '+' sign if it contains insert DNA otherwise by a '-' sign.

uncertainties of the positions of the sites in the restriction map (predicted fragments) or from the indicated broadness of the entered bands (original fragments).

Input file

name.DRW see program MTRANS

D) SEARCH FOR OVERLAPPING DNA CLONES

PROGRAM COMP

Program COMP searches for overlapping regions within all indicated maps in normal and inverted orientation. It allows selection of a specific region of a map which is used for the comparison. If not otherwise specified the vector arms are disregarded. The selected region of the second map (map B) is shifted relative to the first map (map A) in steps of hundred bases (starting from -length of map B to +length of map A) and each time the number of matching and nonmatching sites and the percentage of matching sites of map B within the overlapping region are counted. The result is written in a file as histogram of the shift of map B relative to map A showing [1] the number of matching sites or/and [2] percentage of matches in map B. To avoid unnecessarily long printouts only those shifts can be selected in which one or both of these criteria exceed specified values (Fig. 6 A). [1] and [2] can be connected either by logical 'and' or by 'or'.

Input files

name.MAP see programs FIMA, MTRANS, DRAWMAP

name.LIS generated by program LIST, contains a list of names of name.MAP files. Only the name of map B can be replaced by such a file of file names (if also map A is replaced by such a list program SCOMP has to be used)

Output file

name.COP contains the diagram showing the quality of match for each shift of map B relative to map A (Fig. 6 A)

Program LIST This program creates a file with a list of specified files from a directory containing restriction map data which can be used in the program COMP as file of file names.

Program SCOMP If two sets of restriction maps are compared to find overlapping DNA's, program SCOMP has to be used instead of COMP. It runs in batch mode and allows the entering of two files of file names where each map named in the first file is compared with each of those in the second one by calling succesively the program COMP.

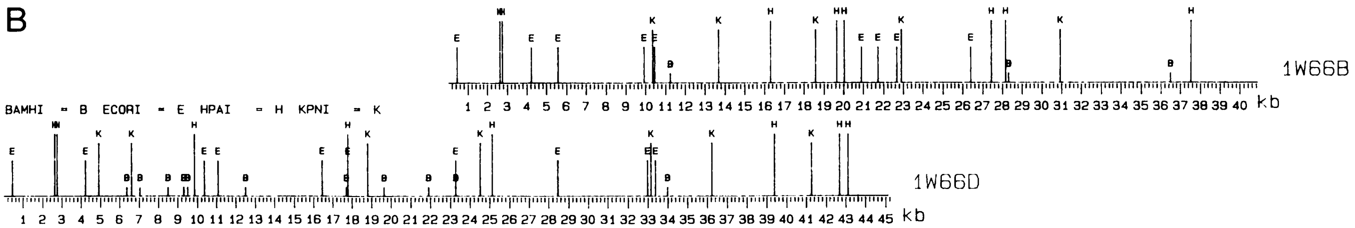
RESULTS

Clone 1W66D, a member of a cluster of cosmid clones that have been analysed in more detail (3), was used here to demonstrate the results from the described programs. We have already shown that maps calculated by 'partial mapping' experiments correlate well with restriction maps generated from sequence data (2).

After calculation of partial fragment sizes from the 'partial mapping' experiment by program FRASI a restriction map was calculated by program FIMA. Fig. 2 shows the resulting total length and site positions determined from either one or both DNA ends. The start and end points of the map is the cos-site position of the vector pCOS2 (7) according to the used mapping

A

```
COMPARISON OF MAP A 1W66B.MAP       AND MAP B 1W66D.MAP

12-MAY-1985   13:15       USED CPU-TIME :  6.13 SECONDS

FILE 1W66B.MAP       (FROM POSITION   4600 TO POSITION   39200)  NUMBER OF USED SITES :  21
         FILE CREATED : 20-DEC-1984  19:59:04
         TOTAL DNA OF  1W66B.PAM      (LENGTH =  40.85 kb)

FILE 1W66D.MAP       (FROM POSITION   4500 TO POSITION   43600)  NUMBER OF USED SITES :  32
         FILE CREATED : 10-MAY-1985  13:21
         MAP OF 1W66D         (TOTAL LENGTH =  45.11 kb)

MINIMUM OF MATCHES TO SHOW SHIFT :  7   O R    30.0 % OF MAP B SITES MATCHING         USED TOLERANCE :  CALCULATED VALUES

 1 = KPNI       2 = BAMHI      3 = HPAI       4 = ECORI

`+` MAX GAP > 2    `=` MAX GAP = 1 OR 2    `*` NO GAP    `:` % SITES OF MAP B MATCHING    SUM = # MATCHES    MIS = # MISMATCHES

1/ 2/ 3/ 4/SUM/MIS/ kb SHIFT OF MAP B
0| 0| 2| 1|  3/14 -23.4 |+++:::::::::::::::::  30%
0| 0| 3| 1|  4/12 -23.3 |====:::::::::::::::::::::::  40%
0| 0| 3| 1|  4/12 -23.2 |====:::::::::::::::::::::::  40%
1| 0| 3| 3|  7/ 6 -23.1 |=======::::::::::::::::::::::::::::::::::::::::  70%
2| 0| 3| 3|  8/ 4 -23.0 |========:::::::::::::::::::::::::::::::::::::::::::::  80%
2| 1| 3| 3|  9/ 2 -22.9 |=========:::::::::::::::::::::::::::::::::::::::::::::::::::  90%
2| 1| 2| 2|  7/ 6 -22.8 |+++++++::::::::::::::::::::::::::::::::::::::::  70%
3| 1| 2| 2|  8/ 5 -22.7 |========:::::::::::::::::::::::::::::::::::::::::::::  80%
3| 1| 1| 2|  7/ 7 -22.6 |=======::::::::::::::::::::::::::::::::::::::::  70%
2| 1| 1| 1|  5/11 -22.5 |=====::::::::::::::::::::::::::::  50%
2| 0| 1| 1|  4/13 -22.4 |++++:::::::::::::::::::::::  40%
2| 0| 1| 1|  4/13 -22.3 |++++:::::::::::::::::::::::  40%
0| 1| 1| 4|  6/27 -12.0 |++++++::::::::::::::  30%
0| 1| 1| 4|  6/27 -11.9 |++++++::::::::::::::  30%
2| 1| 2| 2|  7/31  -5.6 |+++++++::::::::::::  29%
```


B
BAMHI = B ECORI = E HPAI = H KPNI = K
1W66B
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 kb
1W66D
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 kb
E
H
K
B

procedure in which the circular cosmid is cleaved and labelled at the cohesive end site of lambda (3).

Fig.3 B shows the drawing of this map after program MTRANS has replaced the found vector sites by sites determined from the sequence. In this example the option to draw also the borders of the ranges of the site positions was used. The ranges of sites in the insert decreases from the middle of the map toward both ends because the site positions are always calculated from the nearest end and the variance of size calculations due to mobility inaccuracy is much less for small fragments than for large ones. The size of these ranges is also influenced by the broadness entered for each fragment in the program FRASI.

In the next step, the location of fragments generated by single and double digestions within the restriction map was determined. A gel with complete digestions was run, the fragment pattern was entered from a gel photograph using program FRASI and then combined with predicted fragments calculated from the map (Fig. 4) by program MTRANS. Fig. 5 shows the predicted fragments labelled with their size and a letter indicating their map position. To assign the original and predicted fragments a composed gel containing both patterns was drawn (Fig.3 A). In most cases it is easy to identify actual bands corresponding to

Figure 6. (A) File 1W66D.COP created by program COMP after comparison of clone 1W66D and 1W66B.
Shift positions of map 1W66D relative to map 1W66B have been written into the file only if they result in the match of more than 7 site pairs or more than 30 % of sites of map 1W66D compared. Each line of the histogram represents a different shift. The first columns show the number of matches for the respective enzymes (only enzymes found in both maps), 'SUM' is the total sum of matches, 'MIS' is the number of not matching sites, 'kb SHIFT' the relative position of map B to map A. The histogram represents the number of matches (indicated by a '*' if no not matching site interrupt the match, by a '=' if a gap of one or two sites exists, otherwise by a '+') and the percentage of matching sites in map B (indicated by ':') and the percentage value. For each shift all numbers take only the compared region in account.
(B) Maps of 1W66D and 1W66B drawn by program DRAWMAP giving the calculated optimal overlap if one map is shifted 22.9 kilobases relative to the second one.
Note the start of the insert in 1W66B at 4.6 kilobases and the end of the insert in 1W66D at 43.6 kilobases.

predicted fragments and to determine their position in the map (Fig. 4). The assignment is only ambiguous if several fragments exist with very similar mobilities. Additional information is given by the colour of the predicted fragments which distinguishes between fragments generated from the vector, the insert and fragments containing the cos-site sequence. Most of the fragments smaller than 500 bases were not visible on the the gel photograph but might be observed in Southern blot experiments.

To identify additional cosmid clones overlapping with cosmid 1W66D the program COMP was used. Fig. 5 A shows the result of the alignment of the cosmids 1W66D and 1W66B. A strong peak appears if map 1W66D is shifted 22.9 kilobases to the left relative to map 1W66B. This alignment allows 90 % of the compared sites in 1W66B to match with sites in 1W66D. The overlapping regions are obvious from a drawing of both maps with program DRAWMAP (Fig.5 B). The homology of these and other overlapping cosmids was confirmed by heteroduplex analysis and by comparing the distribution of repetitive and non repetitive regions identified by hybridization with total labelled mouse DNA. In this analysis DRAWGEL was used to assign fragments on a X-ray film to bands on a gel drawing and to determine then their positions within the map (3).

The programs described above greatly increase the speed and ease of analysing and combining results generated mainly by 'partial mapping' and hybridization experiments and allow the linkup of overlapping clones. This is the first step in the computerized analysis of large DNA regions. With the increasing number of groups of overlapping genomic clones the use of computers for the storage and interpretion of all known data is essential. Simultaneous access to all available information (restriction maps, locations of repetitive and transcribed DNA regions, genetic markers, partial sequences) will provide the basis for efficient reconstruction and characterisation of genetically meaningful stretches of DNA from cloned segments.

REFERENCES

1 . Smith,H.O. and Birnstiel,M.L. (1976) Nucl. Acids Res. 3, 2387 - 2400.

2. Rackwitz,H-R.,Zehetner,G.,Frischauf,A-M. and Lehrach,H. (1984) Gene 30, 195 - 200.
3. Rackwitz,H-R.,Zehetner,G.,Murialdo,H.,Delius,H.,Chai,J.H., Poustka,A.,Frischauf,A-M.,Lehrach,H. submitted.
4. Southern,E.M. (1979) Anal. Biochem. 100, 319 - 323.
5. Elder,K.J.,Amos,A.,Southern,E.M., and Shippey,G.A. (1983) Anal. Biochem. 128, 223 - 226.
6. Elder,K.J., and Southern,E.M. (1983) 128, 227 - 231.
7. Poustka,A.,Rackwitz,H-R.,Frischauf,A-M.,Hohn,B., and Lehrach,H. (1984) Proc. Nat. Acad. Sci. USA 81, 4129 - 4133.