Volume 10 Number 1 1982 Nucleic Acids Research

Computer programs for nucleic acid sequence manipulation

R.M.Blumenthal, P.J.Rice and R.J.Roberts

Cold Spring Harbor Laboratory, P.O.Box 100, Cold Spring Harbor, NY 11724, USA

Received 1 October 1981

ABSTRACT

Computer programs are described which help during the collection and analysis of nucleic acid sequence data. They are written in FORTRAN and have been implemented on a PDP 11/60 computer.

INTRODUCTION

Following the advent of rapid methods for determining DNA base sequences (1, 2), it has become clear that automated (computer) methods are necessary to process nucleic acid sequence data (3 and references therein). As a result, several overlapping collections of useful computer programs have been assembled for the purpose of manipulating, analyzing, and comparing nucleic acid sequences. These collections fall into two broad categories. The first includes the large, integrated sets of programs together with extensive libraries of sequences that reside in a few, large mainframe computer systems. The second category includes the smaller sets of more specialized programs generally residing in local minicomputers.

We have developed a number of programs in our laboratory for use on a DEC PDP 11 minicomputer, and several of these are described below. All are written in FORTRAN and are available on tape (800 bpi) by request. Two additional programs which allows the entry of sequence data through the use of a digitizing tablet, are described in an accompanying paper (4).

THE PROGRAMS

The FORTRAN code used in the programs listed in Table 1, is DEC's FORTRAN IV-PLUS version 01C and has been implemented on a PDP 11/60 using the RSX11M version 3.1 operating system.

A total of fourteen programs are described, which cover four main aspects of a sequencing project. The first program helps to design cloning

TABLE 1

FUNCTION	PROGRAM(S)
Cloning Strategy	REVCUT (requires amino acid sequence of desired gene's product)
Bookkeeeping	M13 (stores clone data) SEQ (stores sequence reaction data) FIND (collates data for a clone)
Rearrangement/Annotation	REVSEQ (reformatting operations) SPREAD (finds short sequences) CLONE (inserts one sequence into another) INFO (creates information files) COMBINE (uses above to annotate sequences)
Analysis	PRIMER (designs sequencing strategies) RFRAME (displays translation frames) DSPLAY (displays translation frames) MAPPER - -CUTTER (finds restriction sites) -MUTATE (finds altered sites) -FRAGOS (orders fragments and tests sequencing strategies) REMAP (displays restriction maps)

strategies, for use early in a cloning/sequencing project. There are three programs used for bookkeeping operations during the course of the project. Five programs are involved in primary analyses of the data; rearranging, reformatting, and annotating the base sequences. Finally, there are five programs involved in more complex analyses of sequence data, including two that are used in designing strategies to complete the sequence determination of a partiallysequenced DNA species.

CLONING

REVCUT

This program operates upon the amino acid sequence of a polypeptide

and predicts which restriction enzyme recognition sites cannot be present in the DNA sequence which codes for the polypeptide. Such restriction enzymes can then be used to select, in vitro, against undesired clones, and thus allow enrichment of the desired clones. The program uses a simplified back-translation scheme in which no amino acid is represented by more than two codons (e.g., LEU = UUN and CUN). This liberal approach is possible since sites which do not occur are being sought. Because all but two type II restriction endonucleases (BglI and HgiE II) recognize sequences $\leq$9b (3 codons) in length, all other possible restriction sites will be generated by making just 8 sequence lists [2^3 = 2 codon types, 3 codon length]. In the first list, the first of two codon possibilities is always used (pattern=111); in the second list, the repeating pattern 112 is used; in the eighth, 222 is used. Any possible restriction site at any position will occur in at least one list. Once the eight lists are generated, they are searched in parallel by a modified version of our restriction enzyme program (see MAPPER, below), and restriction enzymes lacking sites are indicated. This program is most useful for cDNA cloning where complications due to intervening sequences (5, 6) can be avoided. The constant improvements in protein sequencing technology (e.g. 7) may make programs such as REVCUT increasingly useful in aiding the selection of clones carrying coding sequences.

BOOKKEEPING

A critical requirement in a major sequencing project is keeping track of the large number of clones, reactions, and gels that are generated. Towards that end, we have written 3 programs which are directed towards the book-keeping aspects of the M13 cloning sequencing strategy (11, 12).

M13

This program maintains a record of the various M13 clones made during the sequencing project. The information is stored in a file, M13CLONES.DAT, and includes the assigned clone number, date of entry, cloning vector, fragments inserted, and (when determined) the exact sequence location of the insert (see Fig. 1). Numbers are assigned automatically and date and vector are entered by default or by operator input. The file is easily updated or corrected by the program.

SEQ

SEQ keeps track of sequencing reactions in a very similar manner, assigning a reaction set number and storing the date of entry, primer, the clone used as template, and the results of the reactions in the file M13SEQ.DAT

```
                FROM PROGRAM M13:

NUMBER     DAY      VECTOR     FRAGMENTS                SEQUENCE
   1    17-OCT-80     mp7     MBO I / XHO I F          7497---->
   2    17-OCT-80     mp7     MBO I / XHO I F          7135---->
   8    17-OCT-80     mp7     MBO I / XHO I F          <----7500
   9    17-OCT-80     mp7     MBO I / XHO I F          deletion of mp7
```

```
                FROM PROGRAM SEQ:

 Rx #      Date       Primer     Template          Results
  363    6-JAN-81       CR       clone 390        very good
  364    6-JAN-81       CR       clone 391        fair
  365    6-JAN-81       CR       clone 392        very good
  366    6-JAN-81       CR       clone 393        fair
```

```
                FROM PROGRAM FIND:

         INFORMATION FOR CLONE NUMBER    32

  22-OCT-80    mp7     BCL I /BGL II #3          3589----> (right end)


  REACTION #    DATE        PRIMER    RESULTS

       9      17-OCT-80      fras      fair
      68      7-NOV-80       CR        very good
     126      12-DEC-80      B         failure
    1245      8-JUN-81       b-s       good

                The entry in the information file is :

    GATCTTTCATCCATG          <----------R---------- M13 clone #32
```

Fig. 1: Representative segments of the output file from the bookkeeping programs M13 (A), SEQ (B) and FIND (C). The last two columns in A and B derive solely from operator input -- all other columns have default entries if appropriate. The output from FIND is extracted from the pre-existing files M13CLONES.DAT, M13SEQ.DAT and M13.INF.

(see Fig. 1).

FIND

This program searches the two DAT files and an INF file (see INFO, below) to collate all available information concerning any given clone. Output is to a file, M13.OUT which can be printed (see Fig. 1) and also directly to the CRT terminal. We plan to integrate these programs soon to create a single master bookkeeping program.

REARRANGEMENT/ANNOTATION

REVSEQ

This program is used for reformatting a sequence (e.g. producing the complementary strand) and will also give versions suitable for printing. The sequences can be arranged in spaced blocks of ten bases, five blocks per line, with or without numbers, and with either the input sequence or its reverse complement as the upper line if double stranded format is requested.

SPREAD

This program searches an input file for all occurrences of a defined sequence (up to 30 bases in length), including occurrences of the reverse complement, and lists them together with 20 nucleotides of the flanking sequence and the base number of its occurrence. The output format is especially useful for comparing new sequence data with old data already stored in the computer.

CLONE

The program CLONE inserts a defined portion of one sequence file into a "vector" sequence at a defined point, and will do so in either orientation. This facilitates characterization of such clones, since analysis of the sequence pair with a restriction enzyme program (see MAPPER, below) often allows the insert orientation to be unambiguously determined through a single restriction digest.

INFO/COMBINE

The detailed annotation of a sequence can be performed by the programs INFO and COMBINE, which together constitute a mini-data base management system able to correlate a variety of biological information with specific sequence data. INFO creates information files consisting of an identifying sequence up to 20 bases in length, and an associated label of up to 50 characters in length. A variety of these files have been created. For example, M13CLONES.INF carries data from the sequenced M13 clones of Adenovirus 2 DNA (Fig. 1, bottom); while PROMOTER.INF carries data concerning the known (sequenced) E. coli transcription promoters. COMBINE prints out an input sequence in the double-stranded, numbered format, and wherever an occurrence of an identifying sequence in the chosen INF file is found, the associated label is printed above or below the sequence according to the appropriate strand (Fig. 2). It also places an asterisk above the base corresponding to the first base of the sequence identifier. One distinct advantage of this program is that any renumbering of key sequence locations necessitated by a correction in the sequence is done

```
      THIS FILE COMBINES :
 ADENO.                ADENOVIRUS 0-38% (SEPT 5,81)

 R.INF                 ANNOTATION FOR ADENO ... RESTRICTION SITES, ETC


3.84% M13 clone #1203 ---------------------->
    *   1410          1420          1430          1440          1450
ACACCCGGTG    GTCCCGCTGT    GCCCCATTAA    ACCAGTTGCC    GTGAGAGTTG
TGTGGGCCAC    CAGGGCGACA    CGGGGTAATT    TGGTCAACGG    CACTCTCAAC



       1460          1470          1480          1490          1500
GTGGGCGTCG    CCAGGCTGTG    GAATGTATCG    AGGACTTGCT    TAACGAGTCT
CACCCGCAGC    GGTCCGACAC    CTTACATAGC    TCCTGAACGA    ATTGCTCAGA
                                        *
                    <---------------------- M13 clone #1196   4.05%


                                4.21% Terminator UAA for early Ia polypeptides
       1510          1520          1530          *             1550
GGGCAACCTT    TGGACTTGAG    CTGTAAACGC    CCCAGGCCAT    AAGGTGTAAA
CCCGTTGGAA    ACCTGAACTC    GACATTTGCG    GGGTCCGGTA    TTCCACATTT


             4.29% Hpa I
       1560          *             1580          1590          1600
CCTGTGATTG    CGTGTGTGGT    TAACGCCTTT    GTTTGCTGAA    TGAGTTGATG
GGACACTAAC    GCACACACCA    ATTGCGGAAA    CAAACGACTT    ACTCAACTAC



4.39% AAUAAA for early Ia mRNAs
             4.45% Poly-A addition site for early Ia mRNAs
    *             1620          *             1640          1650
TAAGTTTAAT    AAAGGGTGAG    ATAATGTTTA    ACTTGCATGG    CGTGTTAAAT
ATTCAAATTA    TTTCCCACTC    TATTACAAAT    TGAACGTACC    GCACAATTTA
```

Fig. 2: Typical output from the program COMBINE. The segment shown is an annotated portion of the Ad2 sequence.

automatically.

ANALYSIS

This can be the most complex facet of processing sequence data, and the phase in which it is often most advantageous to use the larger systems referred to in the Introduction. Nevertheless, some types of sequence analysis can and should be performed on smaller, resident systems since

they really represent part of the sequencing process, e.g. testing the quality of the sequence data, and aiding in the sequencing strategy.

PRIMER

This program finds suitable restriction fragments that could be used as primers in sequencing a defined region. The input sequence file, often containing large unsequenced segments ("N-blocks"), is first analyzed for the existence of restriction sites (see MAPPER, below). PRIMER then examines all predicted fragments adjacent to a given region of interest for potential primers of appropriate length. Primers on either side of a defined region are sought.

RFRAME/DSPLAY

These programs are used to reveal open translational reading frames in an input sequence (Fig. 3), but in complementary formats. RFRAME shows all occurrences of ATG and of nonsense codons in each of the 6 reading frames and is designed for output on a printer. The vertical display is useful in sequence troubleshooting where some expectation exists concerning the size of the open reading frame: shifts of the open frame caused by an error in the sequence are easily spotted. DSPLAY, in comparison, produces a much more compact output, designed for viewing on a CRT screen, and shows open reading frames longer than a user-defined limit. The user also determines whether all open reading frames are shown, or only those beginning with ATG. The relative compactness of the DSPLAY output is not a drawback, however, since the program allows one to progressively "zoom in" on any defined region of the sequence.

MAPPER

This set of four programs finds and displays the locations of restriction sites. The first program, CUTTER compares an input sequence to the list of restriction enzyme recognition sites (RENZYMES). The central algorithm, developed by J. Milazzo of SUNY at Stony Brook, limits the search of this list by arranging the RENZYMES substrate sites in alphabetical order (by base) and doing the appropriate limited scan. Output from CUTTER shows the position of the first base of the site, the cleavage site (where known), and whether a 5' overhang, 3' overhang, or blunt end is generated (see Fig. 4, top). Consecutive Ns (undefined bases) are ignored, but non-consecutive Ns can define a site (indicated by "????" in place of "IS @"). Closely related to CUTTER is MUTATE, which finds all sequences which are one base away from being a restriction site.

The FRAGOS program rearranges the CUTTER output file by grouping

```
                                FROM RFRAME:

        E.coli dam methylase gene

         FRAME ONE     FRAME TWO     FRAME THREE                 FRAME FOUR    FRAME FIVE    FRAME SIX
BASE #   TERM     AUG  TERM     AUG  TERM     AUG                TERM     AUG  TERM     AUG  TERM     AUG         BASE #

   19    I             I             I UAA          I            I             I             I             I      1561
   52    I             I             I              I            I             I UAG         I             I      1528
   79    I             I             I UGA          I            I             I             I             I      1501
   85    V             V UAA         V              V            ^             ^             ^             ^      1495
   88    I             I             I              I            I             I UGA         I             I      1492
   94    I             I             I              I            I             I UAA         I             I      1486
  106    I             I        AUG  I              I            I             I             I             I      1474
  112    V             V             V              V            ^ UAA         ^             ^             ^      1468
  130    I             I             I UGA          I            I             I             I             I      1450
  142    I             I        AUG  I              I            I             I             I             I      1438
  178    I             I UAA         I              I            I             I             I             I      1402
  181    V             V             V UAG          V            ^             ^ UAA         ^             ^      1399

------------------------------------------------------------------------------------------
--------------------------------------------------

    FROM DSPLAY:

          RESULTS OF TRANSLATING THE FILE :
 ECO DAM (MELD OF REGIONS 01/28/81)

     LENGTH OF SHORTEST POLYPEPTIDE DISPLAYED IS    20 AMINO ACIDS
     EACH - REPRESENTS    12 NUCLEOTIDES

--------------->                 ----------->             ----------------->                                      -------------->
------> ------>
                ------------------------------------------------------------------------------->
*************|*************|*************|*************|*************|*************|*************|*************|*************|*************|
           157          314          471          628          785          942         1099         1256         1413         1570
        10.06%  .    20.06%       30.06%       40.06%       50.06%       60.06%       70.06%       80.06%       90.06%      100.00%
*************|*************|*************|*************|*************|*************|*************|*************|*************|*************|
                                                                                                               <-----------------
                                                                    <------------          <-------------------------------------    <--------
                <-------                                    <----------------------------          <-------      <---------          <-----
```

Fig. 3: Typical output from the programs RFRAME (A) and DSPLAY (B) showing open translational reading frames. In (A) the base number refers to the base in Frame 1 on the input strand, with frames 2 and 3 being +1 and +2 bases respectively. The corresponding numbers from the 5'-end of the complementary strand are printed in the last column. In (B) both the base numbers and their corresponding percentage coordinates are printed, with either the entire sequence treated as 100% or by recognition of the input sequence file name and automatic assignment of a particular value for 100%. The output shown is from the 130 column output print file, the version on the CRT screen is condensed to 80 columns.

```
      ECO DAM (MELD OF REGIONS 01/28/81)                    LENGTH=  1579--1
    BamH I    IS @      1       CUTS@      1 G^GATCC              5EXT
    Xho II    IS @      1       CUTS@      1 G^GATCC              5EXT
    Mbo I     IS @      2       CUTS@      1 ^GATC                5EXT
    Hha I     IS @     16       CUTS@     18 GCG^C                3EXT
    FnuD II   IS @     17       CUTS@     18 CG^CG                BLNT
    Fnu4H I   IS @     27       CUTS@     28 GC^NGC               5EXT
```

```
    FROM FRAGS:

  ECO DAM (MELD OF REGIONS 01/28/81)                LENGTH= 1579; CIRCULARITY=1

MULTIPLE DIGEST WITH THE ENZYMES
 BamH I
  Dde I
   *
    Hae III
          TOTAL NUMBER OF SITES =    8
  **     1...   419 BASES...B     |     601...   601 BASES...A
  *    420...    80 BASES...E     |       1...   419 BASES...B **
  *    500...    56 BASES...G     |    1202...   220 BASES...C
  *    556...    45 BASES...H     |    1498...    82 BASES...D *
       601...   601 BASES...A     |     420...    80 BASES...E *
      1202...   220 BASES...C     |    1422...    76 BASES...F *
  *   1422...    76 BASES...F     |     500...    56 BASES...G *
  *   1498...    82 BASES...D     |     556...    45 BASES...H *
   * = ASSYMMETRIC, AND ** = SYMMETRIC LABELING.
    sites are at the BEGINNING of the lengths shown
```

```
    FROM REMAP:

ECO DAM (MELD OF REGIONS 01/28/81)


SCALE BEGINS AT  1%
```

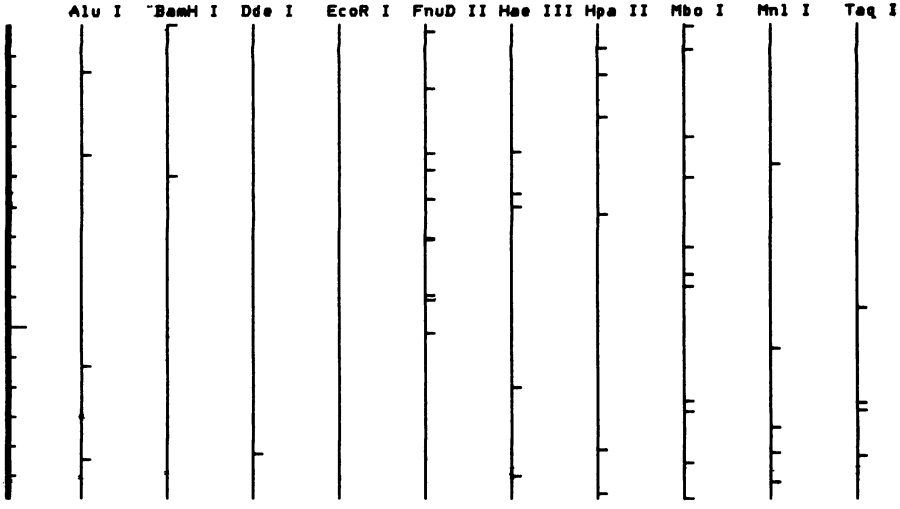

Fig. 4: Representative output from the MAPPER programs. A) Numerically ordered list of cleavage sites. B) Restriction fragments ordered by position (left) and size (right). C) Restriction enzyme maps printed using the plot mode of a Printronix line-printer.

together all like sites, calculating fragment lengths, and putting the output both in order of map position and in order of fragment size. FRAGOS also allows multiple digests. In the example shown in Fig. 4 (middle), the asterisk denotes an end-labeling prior to recleavage with HaeIII, and the labeling pattern of the resulting fragments is shown. FRAGOS has been useful in dry-run tests of strategies for Maxam-Gilbert sequencing where part of the sequence has already been determined.

The program REMAP also uses the output from CUTTER and produces up to 14 conventional restriction maps side-by-side: The resolution (bases per dot) is defined by the user. The program display, shown in Fig. 4 (bottom), depends on the availability of a high resolution plotter or dot-matrix printer (such as the PRINTRONIX P300 in our lab). This program presently runs as a separate module, because improvements are still being made in the graphic output.

CONCLUDING REMARKS

For all of our programs, the current emphasis is on improving the graphic nature of the displays. For programs such as DSPLAY and REMAP, it has been necessary to generate duplicate outputs: one high resolution output for the matrix plotter, and one lower resolution output for the CRT terminal. This is a situation we hope will change as less expensive terminals are made available with more sophisticated graphics capabilities.

ACKNOWLEDGEMENTS

We thank T.R. Gingeras and J. Milazzo for many helpful suggestions during the development of these programs. We also thank N. D'Anna for help in the preparation of this manuscript. This work was supported by a grant from the National Cancer Institute (CA-27275).

REFERENCES

1. Sanger, F., Nicklen, S. and Coulson, A.R. Proc. Natl. Acad. Sci. USA 74, 5463-5467 (1977).
2. Maxam, A.M. and Gilbert, W. Proc. Natl. Acad. Sci. USA, 74, 560-564 (1977).
3. Gingeras, T.R. and Roberts, R.J. Science, 209, 1322-1328 (1980).
4. Gingeras, T.R., Rice, P.I. and Roberts R.J. This issue.
5. Chow, L.T., Gelinas, R.E., Broker, T. and Roberts, R.J. Cell 12, 1-8 (1977).
6. Berget, S.M., Moore, C. and Sharp, P.A. Proc. Natl. Acad. Sci. USA 74, 3171-3175 (1977).

7. Hood, L. Hunkapiller, M., Hewich, R., Griffin, C. and Dreyer, W. J. Supramolecular Structure. In press.
8. Messing, J., Crea, R. and Seeburg, P.H. Nucleic Acids Res. 9, 309-321 (1981).
9. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. J. Mol. Biol. 143, 161-178 (1980).