Volume 10 Number 1 1982 Nucleic Acids Research

Apple II software for M13 shotgun DNA sequencing

Roger Larson and Joachim Messing

Department of Biochemistry, University of Minnesota, St. Paul, MN 55108, USA

Received 18 September 1981

ABSTRACT

A set of programs is presented for the reconstruction of a DNA sequence from data generated by the M13 shotgun sequencing technique. Once the sequence has been established and stored other programs are used for its analysis. The programs have been written for the Apple II microcomputer. A minimum investment is required for the hardware and the software is easily interchangable between the growing number of interested researchers. Copies are available in ready to use form.

INTRODUCTION

In recent years computers have become the logical choice for storage and analysis of DNA sequence data. Time sharing systems will accomplish this task with little trouble but there are some disadvantages to time sharing. The biggest disadvantage is the expense of the computer time, file storage and remote terminal leases. Down time must also be considered as well as the occasions that the users out number the available terminals. In just a year or two the expenses could add up to the purchase price of a microcomputer system.

The Apple II microcomputer has some distinct advantages over the time sharing system. The Apple can be used as a "smart" remote terminal and with the right software it can be just as efficient as a time sharing system. In addition the Apple also allows dynamic file attachment. This feature enables the operator to open disk files while the program is running and is essential in the analysis of sequences derived from the M13 shotgun sequencing method (1). Dynamic file attachment also makes it easier for people with little or no computer experience to work with the computer. Many of the older time sharing systems do not have this feature. Therefore, a set of programs has been written in Pascal and Assembly language for the Apple II. They can be used to record gel readings and reconstruct a DNA sequence from overlapping and complementary M13 templates. A standard set of programs is included

with features such as restriction cleavage mapping and translation to amino acid codons.

HARDWARE

The hardware we are using consists of the following:

1.	Apple II Plus 48k memory	$ 1100
2.	Apple Language System (Pascal)	320
3.	2- Apple II Disk Drives	1000
4.	12" Sanyo monitor	250
5.	M & R Super Term. 80 column board	300
6.	Printer	500 and up
7.	D.C. Hayes Micromodem for Apple II	250
8.	Communications software for Micromodem	50
		$ 3770

The micromodem, which is optional, can be used to make the Apple into a remote terminal. The communications software can be purchased from the Micro Computer Group at the University Computer Center (PHONE: 37M-ICRO). With this software the Apple can "talk" to almost any other computer. Files may be transferred from the Apple to another computer or vice versa.

SOFTWARE

The sequence analysis software for the Apple II is written in Pascal with the exception of the comparison and search subroutines. These were written in Assembly language to decrease execution time. The programs will not run on an Apple without the Apple language system. A basic knowledge of the Apple Pascal operating system is necessary so that new diskettes may be formatted. File deletion and diskette krunching are handled by the Apple Pascal operating system "FILER", so knowledge of this branch of the system is also necessary. The comparison, edit and join programs were written to help construct a master sequence from sequences derived from the M13 shotgun sequencing method. They were designed to help the operator learn from the computer by revealing the mistakes that might have been recorded on the sequence file. The programs are all linked together by using a group of subroutines in the SYSTEM LIBRARY called CHAINSTUFF. A file named SYSTEM. STARTUP is used to start the programs as soon as the Apple is turned on.

Program Description

To start the programs insert the disk into the main drive and turn the

Apple on. Fig. 1 will appear on the screen. All programs may be executed from this main option level. After the completion of a program, control will return to this option level. When using this set of programs, all the members of the same contig (2) should be on the same disk. A contig is a set of gel readings which are related to one another by overlaps of their sequences. Storing a contig on one disk will prevent having to switch disks repeatedly and allow the compare program to search the entire contig for overlaps at one time. All programs are designed so that N is recognized as A, C, T or G.

When the programs ask for sequence file names they must be typed in this form "diskname:filename". A volume number may be substituted for the diskname (ex. '#5:filename'). The volume numbers are as follows: DRIVE 1 = #4, DRIVE 2 = #5, and DRIVE 3 = #11. The diskname or volume number must always be separated from the file name by a colon, unless the file name and disk name are asked for separately. Program execution may be interrupted by typing control (CTRL) and S at the same time, this will enable the operator to view different sections of the results without putting them on the printer. Execution may be resumed by typing another control (CTRL) S.

Option 1. List of Options

No. 1 will print the list of options just as in fig. 1.

Option 2. Join

This option is used after it has been determined that an overlap exists between two sequences. The only limit on this option is that the finished sequence may be no more than 10000 bases. The program will ask for the names of the two sequence files and the position where each sequence will be cut. Next the overlap region will be displayed. If the positioning is

```
WHEN ASKED FOR THE NEXT OPTION PLEASE RESPOND WITH THE NUMBER
OF THE OPTION YOU WISH TO EXECUTE.ALL RESPONSES MUST BE FOLLOWED
       BY A RETURN (<CR>)

 1  :  PRINT THIS LIST OF OPTIONS
 2  :  JOIN TWO SEQUENCES
 3  :  SEARCH FOR RESTRICTION SITES OR STRINGS LESS THEN 25 BASES
 4  :  COMPARE TWO SEQUENCES FOR HOMOLOGY
 5  :  PRINT A SEQUENCE(SINGLE OR DOUBLE STRANDED)
 6  :  LIST OF FILES ON THE DISKETTE IN DRIVE NO. 2
 7  :  TO CREATE A NEW SEQUENCE FILE ON ANY DISKETTE
 8  :  TRANSLATE A SEQUENCE INTO 3 LETTER AMINO ACID CODE
 9  :  EDIT OR CHANGE A SEQUENCE

 NEXT OPTION PLEASE
```

Figure 1

correct the two sequences are put together and stored under a new name on the disk the operator chooses. The operator has the choice of looking at the new sequence on the screen or printer.

Option 3. Search

This program is similar to R. Staden's search program (3) in that it searches a sequence for occurrences of restriction enzymes or short strings. There is no limit to the length of the sequence that is to be searched. Only matches of 100% homology are reported. The operator has a choice of four suboptions: ALL, STRINGS, NAMES and MAP. ALL, NAMES and MAP use a file of restriction enzymes called RENZYMES.TEXT. Enzyme names and cleavage sites may be added or deleted from this file by using the EDITOR that comes with the Pascal operating system. Fig. 2 is an example of the way the RENZYMES file must look. ALL searches the sequence for all occurrences of each of the restriction enzymes in the file. NAMES asks the operator for a name from the restriction enzyme file and searches for each occurrence of it. STRINGS asks the operator for a sequence of less than 25 bases and searches for it. Strings with unknown characters may be searched for by inserting N in place of the unknowns. The output for these three options includes the position number, the distance between adjacent positions, and the site name or the string. Enzymes in the RENZYMES file that are not found are also printed. MAP is the same as ALL except for the output. A restriction enzyme map (fig. 3) is printed with the enzyme names above the corresponding cleavage site. The first letter of the enzyme name lines up with the second base in the cleavage site.

When the program starts the operator is asked whether the results go on the screen or printer. After choosing a suboption, the operator types the name of the sequence to be searched. If NAMES or STRINGS was chosen then the operator types a short sequence or an enzyme name (fig. 4). The sequence

```
'ACC I A' 'GTCGAC' 'ACC I B' 'GTCTAC' 'ACC I C' 'GTATAC' 'ACC I D' 'GTAGAC'
'ACY I A' 'GGCGCC' 'ACY I B' 'GGCGTC' 'ACY I C' 'GACGCC' 'ACY I D' 'GACGTC'
'ALU I' 'AGCT' 'ASU I' 'GGGCC' 'ASU II' 'TTCGAA'
'AVA I A' 'CCCGGG' 'AVA I B' 'CCCGAG' 'AVA I C' 'CTCGGG' 'AVA I D' 'CTCGAG'
'AVA II A' 'GGACC' 'AVA II B' 'GGTCC' 'AVA III' 'ATGCAT'
'AVR II' 'CCTAGG' 'BAL I' 'TGGCCA' 'BAM HI' 'GGATCC' 'BBV I A' 'GCAGC'
'BBV I B' 'GCTGC' 'BCL I' 'TGATCA' 'BGL I' 'GCCNNNNNGGC' 'BGL II' 'AGATCT'
'BSTE II A' 'GGTAACC''BSTE II B' 'GGTCACC''BSTE II C' 'GGTGACC'
'BSTE II D' 'GGTTACC''CLA I' 'ATCGAT''DDE I A' 'CTAAG''DDE I B' 'CTCAG'
'DDE I C' 'CTGAG''DDE I D' 'CTTAG''ECO RI' 'GAATTC''ECO RII A' 'CCAGG'
'ECO RII B' 'CCTGG''FNUD II' 'CGCG''FNU4H I A' 'GCAGC''FNU4H I B' 'GCCGC'
```

Figure 2

```
                  20                  40                  60                  80
       HAE I B          MNL I A                        BBV I B
        HAE III                                        FNU4H I D

CAATGGCAGGCCAAAATATTTTGCCTCCTTATGCTCCTTGGTCTTTCTGCAAGTGCTGCTACGGCGACCATTTTCCCGCA
                 100                 120                 140                 160
         ALU I     ALU I    MBO II B RSA I  HPH I B     MBO II B               M
                                     MNL I A

ATGCTCACAAGCTCCTATAGCTTCCCTTCTTCCCCCGTACCTCTCACCAGCGGTGTCTTCGGTATGTGAAAACCCAATTC
                 180                 200                 220                 240
BO II B      BAM HI              ALU ISFAN I B      HPH I B
              MBO I             BBV I A                          MNL I A
             XHO II A          FPVU II A
TTCAACCCTACAGGATCCAACAGGCAATCGCAGCTGGCATCTTACCTTTATCACCCTTGTTCCTCCAACAATCATCAGCC
                 260                 280                 300                 320


CTATTACAGCAGTTACCTTTGGTGCATTTATTGGCACAAAACATCAGGGCACAACAACTACAACAACTTGTGCTAGCAAA
```

Figure 3

```
PICK ONE OF FOUR OPTIONS FOR THE SEARCH
  A  FOR A SEARCH OF ALL RESTRICTION SITES
  N  FOR SPECIFIC NAMED RESTRICTION SITES
  M  FOR A MAP OF ALL RESTRICTION SITES
  S  FOR STRINGS ENTERED FROM THE KEYBOARD
CHOOSE ONE ,TYPE THE CORRESPONDING LETTER AND THEN <CR>
N
IF YOU WOULD LIKE THE RESULTS PRINTED ON THE PRINTER TYPE P <CR>
TYPE S <CR>  IF YOU WANT THEM ON THE SCREEN.
S
NAME OF SEQUENCE FILE ?#5:SEQA30
AFTER THE PROMPT (TYPE SITE NAME OR STRING ?)
TYPE THE RESTRICTION SITE NAME OR STRING IN SINGLE QUOTES THEN <CR>
WHEN YOU WANT TO STOP TYPE STOP IN SINGLE QUOTES THEN <CR>
TYPE SITE NAME OR STRING ?
'ALU I'
READING THE SEQUENCE FILE.......792 BASES READ

SEARCH FOR (AGCT) ALU I
            POSITION          DIST. BETWEEN SITES

           90                   90
           99                   9
           192                   93
           369                   177
           422                   53
           524                   102
           528                   4
SITE NAME ?
'STOP'

 NEXT OPTION PLEASE
```

Figure 4

is searched and the results printed. The program asks for another string or enzyme name and then searches again. This will continue until STOP in single quotes is typed.

Option 4. Comparison

This program compares two sequences for homology or checks for overlaps. Gel readings from a shotgun sequencing project may have some mistakes in them either from false gel reading or recording errors. This program prints out the comparison in such a way that these errors may be seen. Then the operator can use the editor to correct these mistakes.

When the program starts the operator is asked to type in the name of the sequence file to compare or search for. We will call this sequence A. There is a 500 base limit to this sequence. If a longer sequence is used the program will not terminate, but only the first 500 bases will be used. Then the operator is asked for the name of the sequence file to compare to sequence A. Logically these two sequences should be members of the same contig. There is no limit to the number of bases in this sequence. At this point the program gives the operator the option of typing the sequence file name or a question mark. A question mark will allow the operator to compare to every sequence in the contig without typing in each individual file name.

If the operator types the name of a sequence file, the program goes on and asks the operator to set the parameters. The first parameter is the minimum percentage of bases that must be homologous for a match to be recorded. If a position is found with a percentage of homology above this minimum then that position is recorded. The second parameter is the number of bases in sequence A that the operator wants to use to compare. Here the operator may

```
 A COMPARISON OF Z4ALU4 AND Z4ALU5
 SEARCHING FOR DIRECT HOMOLOGIES

THE UPPER SEQUENCE IS  Z4ALU5
THE LOWER SEQUENCE IS Z4ALU4

       130        140        150        160        170        180        190        200
GCACCGATGGATAATTCTCTCACCAGCGGTGTCTTCGGTATGTGAAAACCCAATTCTTCAACCCTACAGGATCCAACAG
**** ****** ****     *              **  **    *   * *    *    **   * **  **
GCACGGATGGAGAATTTCACGATGGAATCGAATGGTCATAGCTGTTTCCTGTAGTGGAAATTGTTATCCGCTCACACTTC
        10         20         30         40         50         60         70         80

       462        472        482        492        502        512        522        532
TCACGGACGGAGGATACCCCCAGCAATTTCTTCCATTCAACCAACTGGCAGCATTGAACTCTCCTGCTTATTTACAGCA
 ****** **** **  * * *   * *       *   * *   *  * * * ** *   *  *      * ***
GCACGGATGGAGAATTTCACGATGGAATCGAATGGTCATAGCTGTTTCCTGTAGTGGAAATTGTTATCCGCTCACACTTC
        10         20         30         40         50         60         70         80
```

Figure 5

choose a region from the left or the right end of sequence A or all of sequence A. The last parameter is the choice of comparing with the complement, inverse or inverse complement of sequence A. After the parameters are set the operator chooses to have the results printed on the screen or printer. Then the two sequences are compared. The number of matches is reported and the operator is asked how many should be printed. Matches are printed in order from the most homology to least. At the end of every comparison the operator is given the choice of comparing the same two sequences with different parameters, comparing sequence A with another sequence or editing either sequence.

If a question mark is typed the operator can compare to all of the sequences in the contig or just to some of them. If ALL is typed the program asks if the operator wants the parameters the same for all the comparisons. If the operator types "Y" then the parameters are set in the same manner as above and the comparison begins. Every sequence on the disk in DRIVE 2 is compared to sequence A but only five or fewer matches will be printed for each comparison. Again, those positions with the most homology will be printed. If the operator doesn't want all the parameters the same, the program asks for new parameters before comparing each sequence. If SOME is typed, each sequence file on the disk in DRIVE 2 is listed on the screen in this form 'Compare to SEQUENCE 2 ?'. The operator responds with "Y" or "N" after each. After all the files on the diskette have been listed the program proceeds just as it does after typing ALL. Only those files which the operator typed Y after are compared to sequence A.

The output of this program depends on the length of the region compared and whether the region is on the right or left end of sequence A. If the region has fewer than 80 bases and is from the right end of sequence A the printout will look like fig. 6. If the region of comparison is greater than 80 bases the printout will look like fig. 7. Stars show matching bases.

Option 5. Print a Sequence

When this program starts the operator is asked whether the sequence should go on the screen or printer. There is a choice between a single or double stranded output. Then the operator types the sequence file name. Finally the sequence is printed out with location numbers every ten bases. The limit is 10000 bases.

Option 6. List files in DRIVE 2

This program will list the files on the disk in DRIVE 2 on the screen or printer.

```
 A COMPARISON OF Z4ALU4 AND Z4ALU5
 SEARCHING FOR DIRECT HOMOLOGIES

THE UPPER SEQUENCE IS  Z4ALU5
THE LOWER SEQUENCE IS Z4ALU4

       551        561        571        581        591        601        611        621
CTACCATTCAGCCAGCTAGCTGGTGTGAGCCCTGCTACCTTCTTGACACAACCACAGTTGTACCGCCGTTCCACTCGGG
*    * *        ** *   *  * *    **             *    *   *   *** ****** ** *****
CATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTACATTAATTGCGTTGCGCTACTGCCGTTTCAGTCGGG
       112        122        132        142        152        162        172        182

       729        739        749        759        769        779        789        799
TCTACCAACCCATCATTGGTGGTGCCCTCTTTTTTTAGATTTCTTATGAGTTATAGTTACCATGGCCCTTTCACTCGGC
  **        *    *** *   *        *     *  *.   * * *        * *  *** ***** ****
CATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTACATTAATTGCGTTGCGCTACTGCCGTTTCAGTCGGG
       112        122        132        142        152        162        172        182
```

Figure 6

Option 7. Create a file

This program will enter a new sequence and store it on a file. The limit is 10000 bases. The operator is asked to type in the name of the new file and the disk it should go on. Next the sequence is typed in with a star after the last base. The star will tell the computer that input has ended. The sequence is written on the disk with the new file name. After the sequence has been written on the file the operator has the choice of seeing the new sequence on the screen or printer with line numbers every ten bases. At this point if the operator does not want to look at the new sequence, pushing control (CTRL) RESET will put control back at the option level.

Option 8. Translate to Amino Acid Code

This program starts by asking if the operator would like to choose the

```
 A COMPARISON OF Z4RSA78 AND SEQA30
 SEARCHING FOR COMPLEMENTARY HOMOLOGIES

THE UPPER SEQUENCE IS  SEQA30
THE LOWER SEQUENCE IS Z4RSA78

       355        365        375        385        395        405        415        425
CAGTTTCTTCCATTCAACCAACTAGCTGCATTGAACTCTGCTTCTTATTTGCAACAACAACAACTACCATTCAGCCAGCT
 *  * *   **  ***  *              **     * *    ********** ******************
AAACTGCCATCAACCAAATACTAGCTGCATGAACTCTGCTGCTTATTGCAGCAACAACAACTACTACCATTCAGCCAGCT
        10         20         30         40         50         60         70         80

       435        445        455        465        475        485        495        505
ACCTGCTGCCTACCCCCAGCAATTTCTTCCATTCAACCAACTGGCAGCATTGAACTCTCCTGCTTATT
* ***************  ****************************************** *******
AGCTGCTGCCTACCCCCGGCAATTTCTTCCATTCAACCAACTGGCAGCATTGAACTCTCATGCTTATG
        90         100        110        120        130        140        150
```

Figure 7

starting point of translation. If the operator types Yes the program will translate only the region the operator picks. If No is typed the program will locate a protein within the sequence, convert the sequence within the protein to three letter amino acid code and give the molecular weight of the protein. The limit is 10000 bases. After typing Yes or No the operator is asked for the name of the sequence file. Then the file is read and the number of bases is printed.

If No was typed after the first question the operator chooses a start position and an end position and the program searches for the first occurrence of the start codon (ATG). After locating the start codon a stop codon (TAA, TGA and TAG) is searched for using the same frame as the start codon. If either a start c.or stop c. is not found the operator is asked for a new start position. If a start c. and stop c. are found then the program prints the sequence with index numbers starting at the start position and ending at the end position. The amino acid three letter code is printed below the sequence from the start c. to the stop c. At the end of the sequence the molecular weight of the protein is printed. After this the operator is given the choice of having the display that was just on the screen printed on the printer. Then the operator can have the program search for another stop c. or start c. The start c. can be left as is and the sequence can be searched for the next stop c. or if the operator chooses not to search for another start c. or stop c. the program will ask for the start position again. A negative start position will stop the program.

If Yes was typed after the first question the operator is asked if the translated sequence should be printed on the screen or printer. Then a start position and end position are typed and the sequence is translated. A negative start position will stop the program. This is a modified form of a program written by Allen Delaney (4).

Option 9. Edit a sequence

There is no limit to the length of the sequence that may be edited. The operator is asked to type in the name of sequence file to edit. A position number is asked for and after it has been typed the operator is given the choice of either Inserting or Deleting. If the choice is to Delete, the program asks how many bases should be Deleted. If the choice is Insert, the operator is instructed to type them in. The character to signal the end of Insert is a star. Then the program asks for another position number and the cycle is repeated until a negative position number is typed. All edits must be entered in increasing order. To replace a base or a number of bases

```
NAME OF SEQUENCE FILE YOU WANT TO EDIT ?
#5:Z4ALU4
 WHEN YOU WANT TO STOP EDITING TYPE IN A NEGATIVE POSITION NUMBER.

EDIT AT POSITION NO. 10

I(nsert, D(eleteD

Delete how many bases ? 2

EDIT AT POSITION NO. 10

I(nsert, D(eleteI

MAXIMUM number of bases that may be inserted per edit is 5000.
Type the bases to be Inserted . When you are done type a * .
AG*
EDIT AT POSITION NO. -1


Type the name you want the new version to have . It has to be
 different from the previous name. The file being edited must
 remain in DRIVE 2.
Z4ALU6
 TYPE THE VOLUME NUMBER OF THE DRIVE YOU WANT THE NEW FILE PUT ON.
 DRIVE 1 THE VOLUME NUMBER IS 4, DRIVE 2 IS 5
5
WORKING
        10        20        30        40        50        60        70        8
0
GCACGGATGAGGAATTTCACGATGGAATCGAATGGTCATAGCTGTTTCCTGTAGTGGAAATTGTTATCCGCTCACACTT
C

        90       100       110       120       130       140       150       16
0
CACACAACATACTAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTACATTAATTGCGTTG
C

       170       180       190       200       210       220       230       24
0
GCTACTGCCGTTTCAGTCGGG


 NEXT OPTION PLEASE
```

Figure 8

(fig. 8), type the position number and Delete. Then type that same position number and Insert. It also works to reverse the process, Insert first and then Delete. After a negative position number has been typed the operator is asked for a new name for the edited sequence. The old sequence file is not destroyed. The edited sequence is written on the screen and on the new file in DRIVE 2.

These programs were all written with many small subroutines so modifications could be made without too much difficulty. Complete ready to use copies

are available upon request. The program text will also be copied on a disk if changes need to be made. For a copy of both the text and the ready to use code send a stamped self addressed envelope and two blank, single density, soft sectored, 5 1/4" disks to Roger Larson.

Acknowledgement

We would like to thank Bonnie Allen for her help in preparing this manuscript. This work is supported by grants from the Minnesota Agriculture Station MN15-030 and from the Department of Energy, DE-AC02-81 ER 10901.

References

1. Messing, J., Crea, R. and Seeborg, P. H. (1981) Nucleic Acids Research 9, 309-321.
2. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
3. Staden, R. (1977) Nucleic Acids Research 4, 4037-4051.
4. Delaney, A., U. of British Columbia, Vancouver (personal communication).