Apple II Pascal programs for molecular biologists

Bruno Malthiery[1], Bernard Bellon[2], Dominique Giorgi[3] and Bernard Jacq[2§]

[1]Centre Inter Regional Informatique de Lorraine, Château du Montet, Avenue Général Leclerc, 54500 Vandoeuvre Les Nancy, [2]LGBC-CNRS Luminy, case 907, 13288 Marseille Cedex 09, and [3]U 31 INSERM, 46 Bd de la Gaye, 13258 Marseille Cedex 09, France

Received 25 August 1983

ABSTRACT

A collection of PASCAL programs designed for the Apple II microcomputer is presented. These DNA sequence handling and analysis programs are interactive and may be used even by people with no computer experience.
The package allows the user to enter a sequence from the keyboard, to modify it, to generate the reverse complement, to create new sequences from parts of other ones, to display or print sequences in various formats. Some analysis tasks are also performed : Translation, searches for restriction sites, for homology with subsequences, either perfect or with an adjustable match percentage. In addition, two programs are also included : The first one allows DNA data sequences generated with a BASIC program under the CP/M operating system to be used with these PASCAL programs. The second one is designed for the automatic assembly of DNA fragments sequences, obtained with the GILBERT-MAXAM or M13 techniques, into a complete sequence.

INTRODUCTION

In the last six years, our knowledge of genome organisation in prokaryotes and eukaryotes has dramatically increased after the introduction of recombinant DNA technology. The availability of powerful DNA and RNA sequencing methods has led to the determination of the complete sequence for several small genomes (plasmids, phages, mitochondria) and large regions of longer genomes, both in prokaryotes and eukaryotes. In the same time, computers became increasingly used to handle and analyse these sequences. More recently, different nucleotide sequence data banks have been settled : in Lyon (FRANCE), Heidelberg (EMBO data bank), Los Alamos (USA) and Washington (USA). In different places, very sophisticated sequence analysis softwares have been developed, which can be used in conjonction with such large databases. And in some cases, it is now possible to work on line with a remote computer where these databases and programs have been implemented, through an international communication network. A laboratory microcomputer equipped with a MODEM can be used for such a purpose and it is quite certain that this will represent the future trend for most of small molecular biology research small groups : they will have a microcomputer in the laboratory for the entry of home made sequences, data editing and a lot of simple searches (restriction sites, translations, strict homologies, ...). The use of a MODEM coupled to the microcomputer will allow them to access to large librairies and more elaborated programs.

After the pioneering work of some people (1,2,3) a number of programs have now been written which cover a large range of problems encountered in a molecular biology laboratory (see ref. 4 for a collection of such programs). The large majority of them, however, has been written to be used on mini or main frame computers. As outlined above, we believe it is essential that programs for microcomputers may be available, either for practical or financial reasons. Some programs for microcomputers have been published (5,6,7). Here we describe a sequence handling and analysis PASCAL software to be used on an APPLE II microcomputer.

MATERIAL AND METHODS

Hardware

The hardware necessary to use the programs is the following :

. Apple II plus (48 Ko) or II e (64 Ko), or III.
. Apple language card (16 Ko) with Apple II plus only
. Monitor (Philips, Sanyo,...)
. 2 Apple 5" 1/4 disk drives.
. Printer and interface
. Optional : Z 80 MICROSOFT softcard (see below the CPM/UCSD program).

Language and general software description

The programs described in this article were written in the Apple version of UCSD PASCAL. Some subroutines were written in the 6502 Assembly language to decrease execution time.

The programs are very interactive and can be used by people having no previous computer experience. After having inserted the two programs diskettes in the drives and turned the Apple on, the user has first to type "X" on the keyboard and then to answer "MENU" to the question "execute what file" asked by the operating system. From this moment, the user is them kindly taken by the hand by the program until the end of his session. He has only to answer the questions which appear on the screen or to do what is asked by the computer (e.g. : "PUT DATA DISK IN DRIVE 2", "TAPE RETURN", ...). A little knowledge of the PASCAL operating system is necessary : When new diskettes have to be used, they need to be formatted before. An option has been inserted in the program which allows sequence files to be deleted when necessary or transferred to another diskette.

We have tried to make the programs as "USER-FRIENDLY" as possible :

. an option "EXPLANATIONS" tells the user what each program is doing.

. In most cases, care has been taken that an erroneous answer to a question does not make the execution of the program to abort prematurely (e.g. : to give an incorrect name for a restriction enzyme, or for a sequence file, ...). If such an error is detected, the user is invited to give an other answer.

. In some cases, a program needs to write a temporary file on a diskette. If this is the case, the program first tests the available place, and, if not sufficient invites the user to put another diskette in the drive.

. Finally, other letters than A, C, G, T are accepted by the programs : N is recognized as A, C, G, T or U ; Y as C, T, U ; R as A or G ; K as G, T, U ; X as A, T, U ; J as A or C and S as C or G.

RESULTS

In its present version (version 1.0, july 1983) this sequence analysis software proposes 17 options to the user, which appear on the screen as a "MENU". After a particular program has been executed and the results displayed on the screen or printed upon user's choice, the general "MENU" is proposed once again, and a new task can be performed. The available options will be briefly described below. Most of them are also proposed in similar packages for microcomputers (5,6,7) whereas some others are new or may represent (to our point of view) improvements to previously published ones.

Option 1 : CPM/UCSD

Some advantages in the use of the Apple II microcomputer over larger systems have been previously described (6). An other one is that it can be used under different operating systems and with different languages (BASIC, PASCAL, LISP, ...). The BASIC language is probably the first which should be learned and used by programming beginners. The Microsoft BASIC (used under the CP/M operating system with the Z 80 softcard) has become a "de facto" standard in the BASIC languages family. It includes many features not found in APPLESOFT BASIC, and in particular the "INSTR" string function, which is very useful in string manipulation.

For these reasons, we have written a set of programs in MICROSOFT BASIC for the Apple computer (to be published elsewhere). Having entered about 50 Kb of different sequences using these BASIC programs, we found it useful to be able to use these data with PASCAL programs. This is possible with the CPM/UCSD program :

The user is asked to put a diskette containing CP/M sequence data files in drive 1 and a blank formatted diskette in drive 2. The programs reads the directory of files and asks if all sequences must be transferred. If it is not the case only the desired sequences are transferred. The sequences stored as arrays of strings in CP/M are then transformed in files of characters and can be used with the other PASCAL programs.

Option 2 : Sequence data file transfer.

This option is used to transfer a sequence file from a diskette to another one. The CP/M allows only 48 files to be created and stored on a diskette, whereas the PASCAL operating system allows 77. The file transfer program can then be used to store 29 supplementary files on a PASCAL copy of a 48 files CP/M diskette.

Option 3 : Enter a sequence.

This option allows the entry of a new sequence, from the keyboard and stores it on a file. The user is first asked if the sequence is a protein or a nucleic acid. In the latter case, it will ask if it is circular or not. The user has then to give the name of the sequence (max. 10 Characters), and, if desired, can enter two comments, up to 255 characters each, concerning the sequence. The sequence is then entered, by strings of 40 characters (proteins are entered in the one-letter code). The sequence is stored as a packed array of characters and the upper limit for the storage is that of the diskette : approximately 135 Kb of sequence. After the last character, the user types a "*", and the length of the sequence is automatically calculated. It has to be noted that each sequence is stored as an independent file whereas all the comments concerning the sequences are stored in a single file.

Option 4 : Creation of a sequence from old ones.

This option can be used to construct a new, hybrid sequence from parts of other sequences stored on the diskette.

Option 5 : Creation of the reverse complement of a sequence.

This option makes the complementary strand of a sequence in the 5' 3' sense. It must be used, for instance, when the user wants to translate the complementary strand of a sequence file (option 9). If the sequence contains an ambiguous nucleotide such as Y, R, K, X, ... (see the general description section) the corresponding complementary ambiguous nucleotide is generated.

Option 6 : Modification of a sequence.

This program is a tiny editor and can be used either to correct a sequence (the old sequence is erased and the new one is stored under the same name) or to generate a new sequence with a new name from an old one, with only slight modifications (case of a family of related proteins like immunoglobulins for instance). The user is first required to type the name of the sequence to modify, and then the number of the first modification area : if the answer is 2010, nucleotides 2001-2040 of the desired sequence appear on the screen. At this stage 5 commands are available : "I" (insertion), "D" (deletion), "M" (replacement), " " (5' 3' deplacement), "CTRL-D" (end of modifications). The use of the " " key allows the user to put the cursor on a precise nucleotide. The selection of a particular mode (e.g. deletion) makes then new commands to be available, which in that case are : SPACEBAR (to delete one nucleotide), ESC (to quit), CTRL-D (to valid the operation). The user can select a new modification area (always in the 5' 3' direction) and proceeds. The numerotation displayed on the screen is the one before modification, and each modification step is typed out on the printer.

Option 7 : printing a sequence.

The program allows the user either to display a sequence (or a part of it) on the screen, or to print it on paper. 3 different editing formats are proposed. Each of them includes the possibility of double strand output (see fig. 1).

Option 8 : deletion of a sequence file.

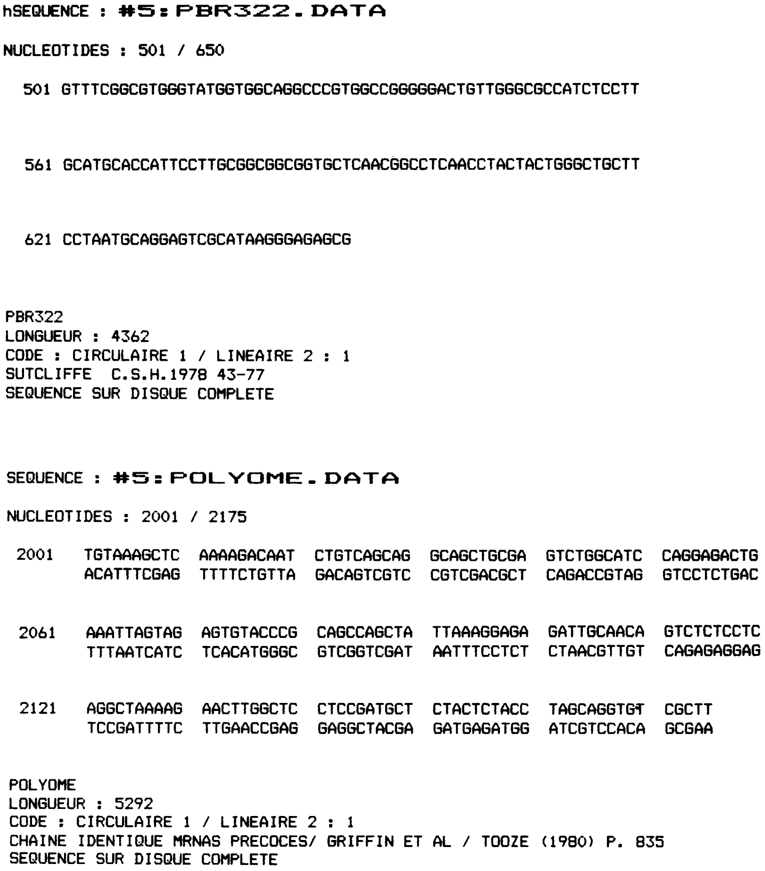
This option permits the deletion of one or more sequence files from the diskette. The comments being stored on a separate file (comment-data), the program removes the corresponding comments from this file.

Option 9 : Translation of a DNA sequence.

The user can select the region to be translated, the output (screen or printer), the one or three letter amino-acid code, the number of reading frames and the output format (number of codons per line). When a termination codon is encountered, a "*" is printed in the peptide sequence. If the DNA sequence contains ambiguous nucleotides such as N, Y, R, ... (see the general description section) and if unambiguous amino-acids result from their translation, these amino-acids are printed. If more than one amino-acid could result from the translation of the corresponding ambiguous codon, a blank is left in the peptide sequence. Finally, the number of stop codons is printed for each reading frame, as well as the number of codons for which the translation led to an ambiguous amino-acid. A typical output is presented in fig. 2.

Option 10, 11, 12 :

These options are used in a search for restriction sites. Option 10 is used to add new restriction enzymes, their corresponding recognition sites



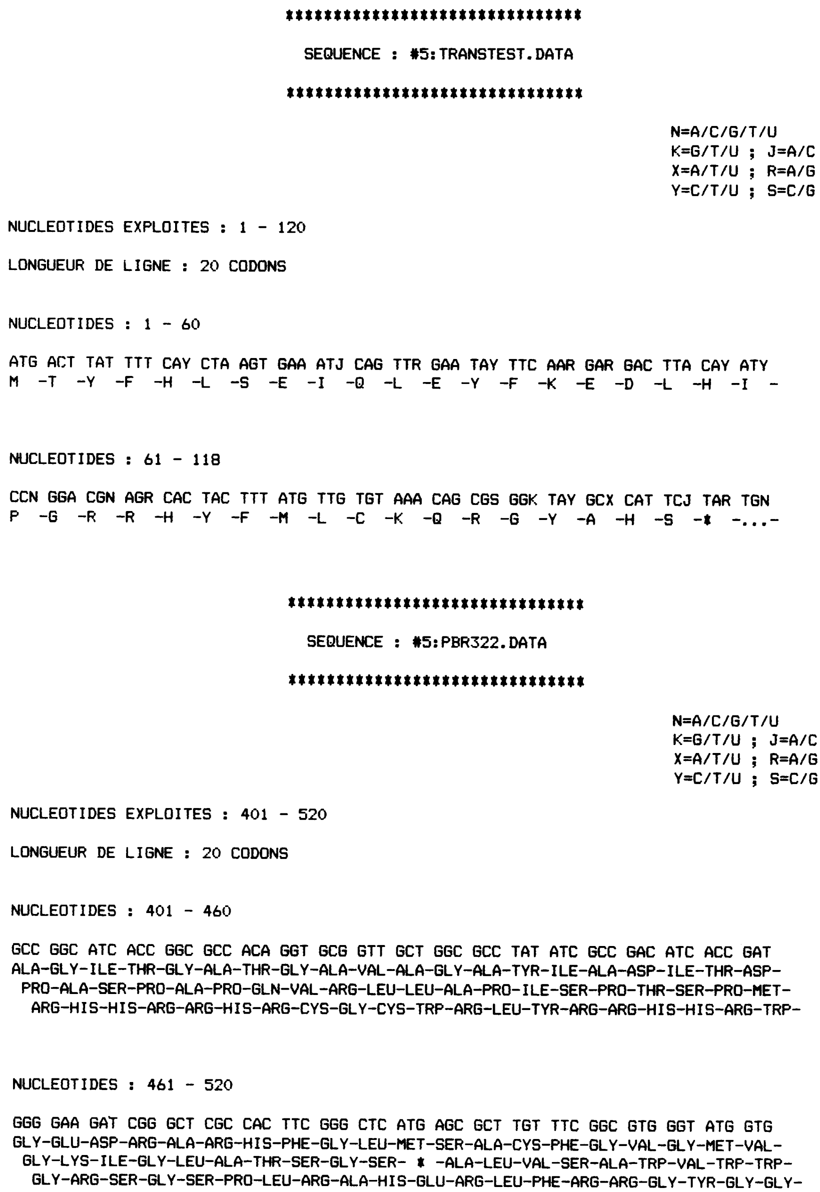
```
hSEQUENCE : #5:PBR322.DATA

NUCLEOTIDES : 501 / 650

  501 GTTTCGGCGTGGGTATGGTGGCAGGCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTT


  561 GCATGCACCATTCCTTGCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTT


  621 CCTAATGCAGGAGTCGCATAAGGGAGAGCG


PBR322
LONGUEUR : 4362
CODE : CIRCULAIRE 1 / LINEAIRE 2 : 1
SUTCLIFFE  C.S.H.1978 43-77
SEQUENCE SUR DISQUE COMPLETE


SEQUENCE : #5:POLYOME.DATA

NUCLEOTIDES : 2001 / 2175

 2001    TGTAAAGCTC  AAAAGACAAT  CTGTCAGCAG  GCAGCTGCGA  GTCTGGCATC  CAGGAGACTG
         ACATTTCGAG  TTTTCTGTTA  GACAGTCGTC  CGTCGACGCT  CAGACCGTAG  GTCCTCTGAC

 2061    AAATTAGTAG  AGTGTACCCG  CAGCCAGCTA  TTAAAGGAGA  GATTGCAACA  GTCTCTCCTC
         TTTAATCATC  TCACATGGGC  GTCGGTCGAT  AATTTCCTCT  CTAACGTTGT  CAGAGAGGAG

 2121    AGGCTAAAAG  AACTTGGCTC  CTCCGATGCT  CTACTCTACC  TAGCAGGTGT  CGCTT
         TCCGATTTTC  TTGAACCGAG  GAGGCTACGA  GATGAGATGG  ATCGTCCACA  GCGAA

POLYOME
LONGUEUR : 5292
CODE : CIRCULAIRE 1 / LINEAIRE 2 : 1
CHAINE IDENTIQUE MRNAS PRECOCES/ GRIFFIN ET AL / TOOZE (1980) P. 835
SEQUENCE SUR DISQUE COMPLETE
```

Figure 1 : Two examples of sequence output using two possible formats.

and cut positions into a restriction enzyme library. Option 11 lists the enzyme library (screen or printer). Option 12 is the restriction site search. It is possible to search either for all sites in library or for some of them. The search is made on a complete DNA sequence or on a specified part of it only (see fig. 3 for an output example).

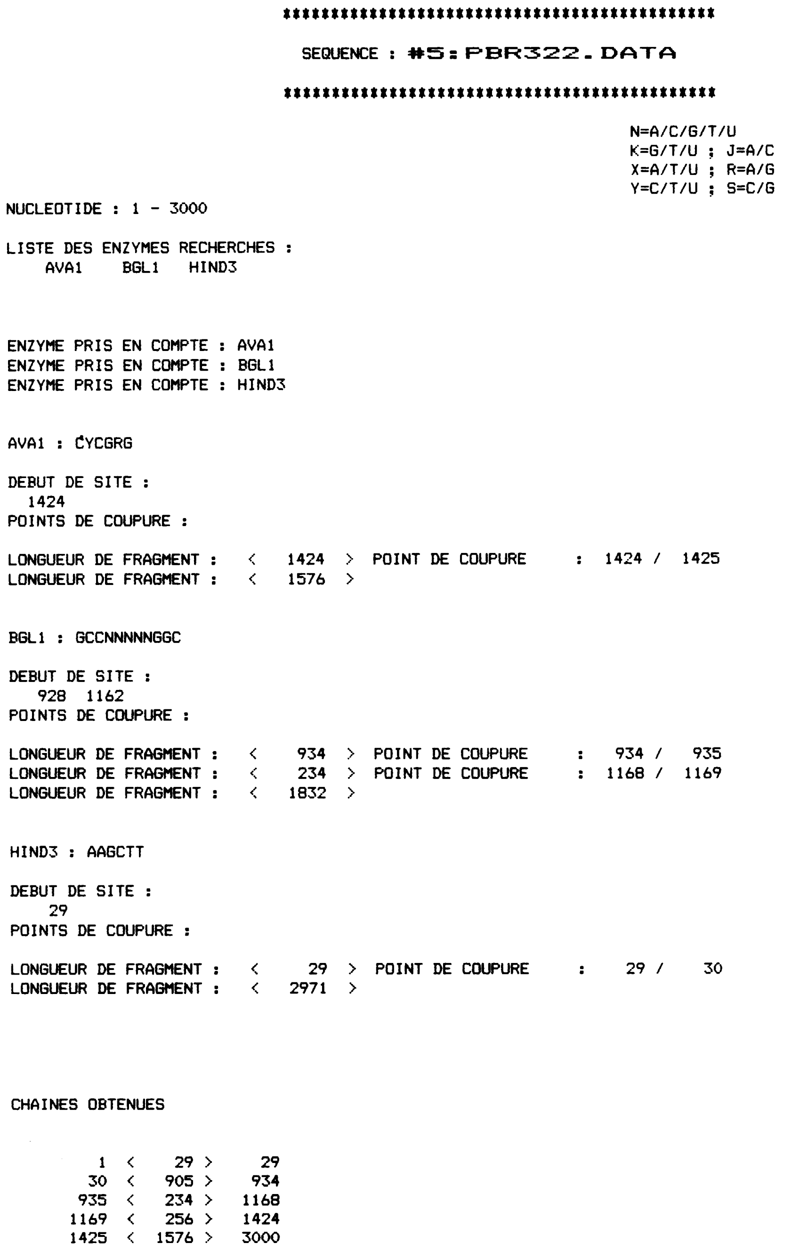
Option 13, 14 :

These two options are used to find homologies between small sequences and a sequence (nucleic acid or protein) present in the data file. Option 14 searches for the occurrences of a maximum of 15 different sequences up to 100 nucleotides each in a choosen sequence (or in a part of it). Only strict homologies are taken into account, nevertheless the use of the ambiguous



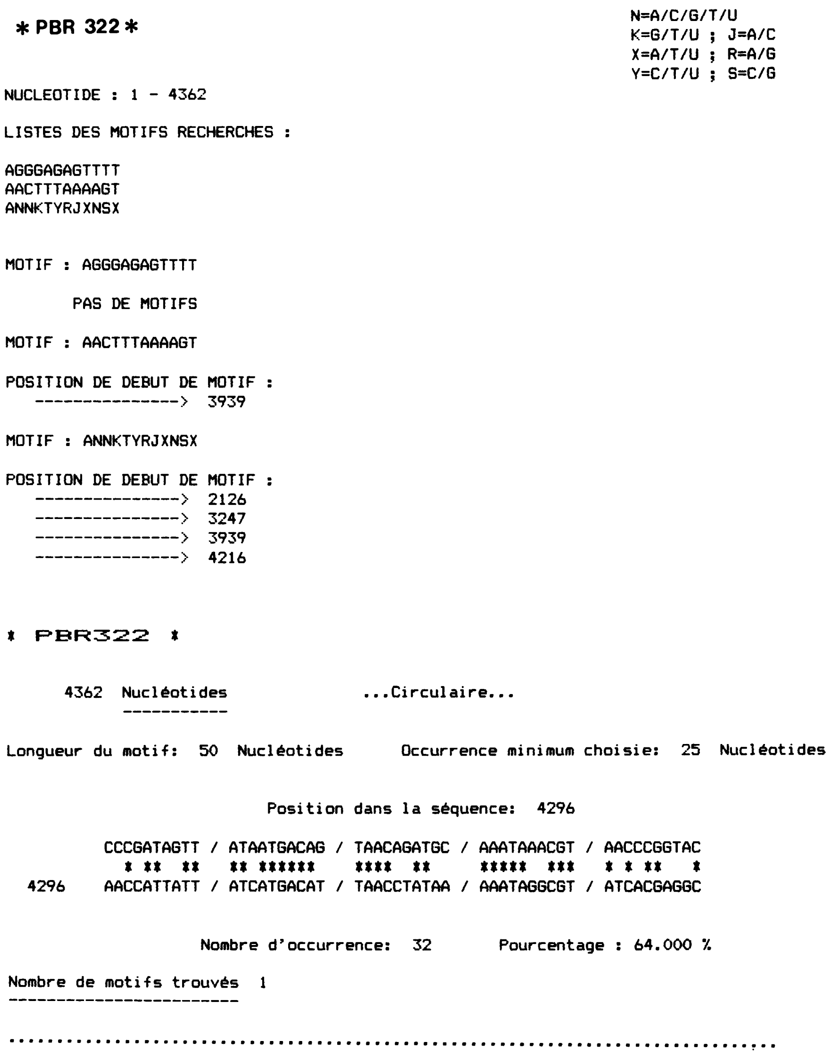
```
                                  *********************************

                                     SEQUENCE : #5:TRANSTEST.DATA

                                  *********************************

                                                                                     N=A/C/G/T/U
                                                                                     K=G/T/U ; J=A/C
                                                                                     X=A/T/U ; R=A/G
                                                                                     Y=C/T/U ; S=C/G

NUCLEOTIDES EXPLOITES : 1 - 120

LONGUEUR DE LIGNE : 20 CODONS


NUCLEOTIDES : 1 - 60

ATG ACT TAT TTT CAY CTA AGT GAA ATJ CAG TTR GAA TAY TTC AAR GAR GAC TTA CAY ATY
M   -T  -Y  -F  -H  -L  -S  -E  -I  -Q  -L  -E  -Y  -F  -K  -E  -D  -L  -H  -I  -


NUCLEOTIDES : 61 - 118

CCN GGA CGN AGR CAC TAC TTT ATG TTG TGT AAA CAG CGS GGK TAY GCX CAT TCJ TAR TGN
P   -G  -R  -R  -H  -Y  -F  -M  -L  -C  -K  -Q  -R  -G  -Y  -A  -H  -S  -*  -...-


                                  *********************************

                                     SEQUENCE : #5:PBR322.DATA

                                  *********************************

                                                                                     N=A/C/G/T/U
                                                                                     K=G/T/U ; J=A/C
                                                                                     X=A/T/U ; R=A/G
                                                                                     Y=C/T/U ; S=C/G

NUCLEOTIDES EXPLOITES : 401 - 520

LONGUEUR DE LIGNE : 20 CODONS


NUCLEOTIDES : 401 - 460

GCC GGC ATC ACC GGC GCC ACA GGT GCG GTT GCT GGC GCC TAT ATC GCC GAC ATC ACC GAT
ALA-GLY-ILE-THR-GLY-ALA-THR-GLY-ALA-VAL-ALA-GLY-ALA-TYR-ILE-ALA-ASP-ILE-THR-ASP-
 PRO-ALA-SER-PRO-ALA-PRO-GLN-VAL-ARG-LEU-LEU-ALA-PRO-ILE-SER-PRO-THR-SER-PRO-MET-
  ARG-HIS-HIS-ARG-ARG-HIS-ARG-CYS-GLY-CYS-TRP-ARG-LEU-TYR-ARG-ARG-HIS-HIS-ARG-TRP-


NUCLEOTIDES : 461 - 520

GGG GAA GAT CGG GCT CGC CAC TTC GGG CTC ATG AGC GCT TGT TTC GGC GTG GGT ATG GTG
GLY-GLU-ASP-ARG-ALA-ARG-HIS-PHE-GLY-LEU-MET-SER-ALA-CYS-PHE-GLY-VAL-GLY-MET-VAL-
 GLY-LYS-ILE-GLY-LEU-ALA-THR-SER-GLY-SER- * -ALA-LEU-VAL-SER-ALA-TRP-VAL-TRP-TRP-
  GLY-ARG-SER-GLY-SER-PRO-LEU-ARG-ALA-HIS-GLU-ARG-LEU-PHE-ARG-ARG-GLY-TYR-GLY-GLY-
```

Figure 2 : Typical output from the translation option.
The upper part of the figure is the translation of an hypothetical sequence containing ambiguous nucleotides. The user has choosen to analyse one reading frame only and to use the one-letter amino-acid code.
The lower part is a translation of a pBR 322 segment in the 3 reading frames. The "*" means a stop codon and the "..." means that the translation of the codon leads to an ambiguous amino-acid.



```
                          ************************************************

                             SEQUENCE : #5:PBR322.DATA

                          ************************************************

                                                                N=A/C/G/T/U
                                                                K=G/T/U ; J=A/C
                                                                X=A/T/U ; R=A/G
                                                                Y=C/T/U ; S=C/G
NUCLEOTIDE : 1 - 3000

LISTE DES ENZYMES RECHERCHES :
    AVA1    BGL1   HIND3


ENZYME PRIS EN COMPTE : AVA1
ENZYME PRIS EN COMPTE : BGL1
ENZYME PRIS EN COMPTE : HIND3


AVA1 : ĊYCGRG

DEBUT DE SITE :
  1424
POINTS DE COUPURE :

LONGUEUR DE FRAGMENT :   <   1424  >  POINT DE COUPURE     :  1424 /  1425
LONGUEUR DE FRAGMENT :   <   1576  >


BGL1 : GCCNNNNNGGC

DEBUT DE SITE :
   928  1162
POINTS DE COUPURE :

LONGUEUR DE FRAGMENT :   <    934  >  POINT DE COUPURE     :   934 /   935
LONGUEUR DE FRAGMENT :   <    234  >  POINT DE COUPURE     :  1168 /  1169
LONGUEUR DE FRAGMENT :   <   1832  >


HIND3 : AAGCTT

DEBUT DE SITE :
    29
POINTS DE COUPURE :

LONGUEUR DE FRAGMENT :   <     29  >  POINT DE COUPURE     :    29 /    30
LONGUEUR DE FRAGMENT :   <   2971  >




CHAINES OBTENUES


         1  <    29 >     29
        30  <   905 >    934
       935  <   234 >   1168
      1169  <   256 >   1424
      1425  <  1576 >   3000
```

Figure 3 : Representative output from the restriction site search program. The list of restriction sites to be found is printed. The length of the restriction fragments obtained and the positions of cleavage are reported.



```
* PBR 322 *                                    N=A/C/G/T/U
                                               K=G/T/U ; J=A/C
                                               X=A/T/U ; R=A/G
                                               Y=C/T/U ; S=C/G
NUCLEOTIDE : 1 - 4362

LISTES DES MOTIFS RECHERCHES :

AGGGAGAGTTTT
AACTTTAAAAGT
ANNKTYRJXNSX


MOTIF : AGGGAGAGTTTT

       PAS DE MOTIFS

MOTIF : AACTTTAAAAGT

POSITION DE DEBUT DE MOTIF :
   ---------------> 3939

MOTIF : ANNKTYRJXNSX

POSITION DE DEBUT DE MOTIF :
   ---------------> 2126
   ---------------> 3247
   ---------------> 3939
   ---------------> 4216
```

```
* PBR322 *

      4362 Nucléotides              ...Circulaire...
           -----------

Longueur du motif:  50  Nucléotides      Occurrence minimum choisie:  25  Nucléotides

                              Position dans la séquence:  4296

           CCCGATAGTT / ATAATGACAG / TAACAGATGC / AAATAAACGT / AACCCGGTAC
             * **  **   ** ******    ****  **     *****  ***   * * **   *
4296       AACCATTATT / ATCATGACAT / TAACCTATAA / AAATAGGCGT / ATCACGAGGC


                       Nombre d'occurrence:  32      Pourcentage : 64.000 %

Nombre de motifs trouvés  1
------------------------

.............................................................................
```

Figure 4 : Typical output from option 13 and 14.
At the top of the first output, the 3 different sequences to be identified in pBR 322 are printed. The third one contains ambiguous nucleotides. The position of the first nucleotide of the occurence is reported.
The second output is an example of the rapid search program.
The user has entered a 50 nucleotide sequence from the keyboard and searched for all possible occurences having a minimum of 50 % homology, in the entire pBR 322 sequence.
1 region of homology was found and the search took 11 seconds.
The upper sequence is the one entered from the keyboard and the lower one is the part of the pBR 322 sequence in which the homology was found.

code (see general description section) in the sequence to be searched for allows matches with incompletely defined sequences to be reported. Option 14 is a very rapid program to search for occurrences with an adjustable homology percentage. Sequences up to 255 nucleotides long can be searched for. A sub-option proposes the choice to search either for a sequence entered from the keyboard or for a part of a stored sequence. A search for all homologies above a defined minimum percentage between a 100 nucleotides subsequence and PBR 322 (4362 nucleotides) takes 20 seconds only, once the PBR sequence is loaded in the memory (fig. 4).

Option 15 : Assembly of DNA sequence fragments

This program is a first step towards an automatic computer-aided assembly of the DNA fragments obtained in the course of a sequencing project. In an ideal way, a program designed for such a purpose would be able :

. To store and handle a large number of fragments, 100-300 nucleotides long.

. To determine if overlaps exist between them in any of the 4 possible combinations of strands (same strand or complement) and orientations (5' 3' or 3' 5').

. To merge the overlapping fragments in a continuous sequence.

The third problem is not a trivial one, especially when discrepancies involving insertions/deletions exist between overlaps. Programs have been published which partially fulfil the above requirements (9,10) but were designed for mini or main frame computers.

The program proposed in option 15 is able to handle 70 fragments up to 500 nucleotides each. The sequence to be reconstructed must not exceed a final length of 6000 nucleotides. We have tested the determination of overlaps and the merging process with a set a subsequences covering 2000 nucleotides of PBR 322. The program succeeded in the reconstruction of the sequence when no discrepancy was introduced in overlapping stretches. If differences exist, which are neither insertions nor deletions, the sequence is still re-assembled and N is put at each position occupied by conflicting nucleotides. We are presently modifying the program in such a way that insertions and deletions could be handled.

Options 16 and 17 :

Option 16 displays a short explanation text for each option of the program. Option 17 is used to stop execution of the programs.

CONCLUDING REMARKS

A large variety of computers and programming languages have been used to design, write and run programs necessary to analyse the growing mass of sequence data. The writing of a complete package being a time-consuming activity, it would be advantageous to be able to use already available programs. However the software compatibility problems still represent a limiting step in the use of foreign programs (see ref. 7 for a discussion of such problems). We describe herein a PASCAL DNA handling and analysis package for the APPLE II microcomputer. A program called CPM/UCSD is included, which converts CP/M data files in PASCAL compatible ones. It is thus possible to APPLE owners to use either MICROSOFT BASIC or PASCAL programs once the format of the files has been modified. Although such data conversion programs are not a definitive solution for compatibility problems, we hope they may favor program exchange between people using a given type of computer with different languages.

ACKNOWLEDGEMENTS

This work was supported by an ATP from the Centre National de la Recherche Scientifique. Thanks are due to Brigitte Magnan for the typing of the manuscript. Special thanks are offered to F. Cuzin (Centre de Biochimie, Nice) to make his programs as well as his collection of sequences available to us. A more complete english version of the programs will be available in the middle of 1984. Program code and/or text are available on request. For both text and code copies, please send 4 blank 5" 1/4 diskettes to Bernard Jacq.

§ To whom reprints request should be addressed.

REFERENCES

1 Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051
2. Korn, L. J., Queen, C. L., Wegman, M. N. (1977) Proc. Natl. Acad. Sci. (USA) 74, 4401-4405
3 Gingeras, T. K., Roberts, R. J. (1980) Science 209, 1322-1328
4. Nucl. Acids Res. (1981) : Special issue devoted to the application of computers to research on nucleic acids, IRL Press, D. Söll and R. Roberts eds, 10, 1-456
5. Conrad, B. and Mount, D. W. (1981) Nucl. Acids Res. 10, 31-38
6. Larson, R. and Messing, J. (1981) Nucl. Acids Res. (1981) 10, 39-50
7. Pustell, J. and Kafatos, F. C. (1981) Nucl. Acids Res. 10, 51-60