A general DNA analysis program for the Hewlett-Packard Model 86/87 microcomputer

Eric H. Harley

Department of Chemical Pathology, University of Cape Town Medical School, Observatory 7925, South Africa

Received 6 June 1985

ABSTRACT

A program is described to perform general DNA sequence analysis on the Hewlett-Packard Model 86/87 microcomputer operating on 128 K of RAM. The following analytical procedures can be performed: 1. display of the sequence, in whole or part, or its complement; 2. search for specified sequences e.g. restriction sites, and in the case of the latter give fragment sizes; 3. perform a comprehensive search for all known restriction enzyme sites; 4. map sites graphically; 5. perform editing functions; 6. base frequency analysis; 7. search for repeated sequences; 8. search for open reading frames or translate into the amino acid sequence and analyse for basic and acidic amino acids, hydrophobicity, and codon usage. Two sequences, or parts thereof, can be merged in various orientations to mimic recombination strategies, or can be compared for homologies. The program is written in HP BASIC and is designed principally as a tool for the laboratory investigator manipulating a defined set of vectors and recombinant DNA constructs.

INTRODUCTION

This program was written to fulfill the need for a versatile DNA sequence analysis package for the Hewlett-Packard (HP) model 86 or 87 microcomputer. These are widely used laboratory microcomputers intended principally for local laboratory scientific applications and on-line data processing. They are designed to operate predominantly in HP BASIC, the programming and graphics features of which are well suited for developing and continuously modifying a DNA analysis program. Although hardware additions can be made which enable the HP 86/87 to accept programs written in other versions of Basic, or in Pascal, these downgrade the random access memory to 64K and do not enable the built in graphics capabilities or hard-wired programming statements to be used to best advantage.

The program will be provided on either 5.25 or 3.5 inch flexible discs to anyone on request to the author, subject to a fee in advance of $50 for the disc, postage, and processing.

© IRL Press Limited, Oxford, England.
The program is designed expressly for the active researcher engaged in construction and manipulation of recombinant DNA molecules who does not necessarily have any programming knowledge but needs rapid and frequent access to sequence analysis relevant to his day to day experiments and constructions.

HARDWARE

For maximum economy the full system is designed to operate on the following minimum hardware:

1. Hewlett Packard Model 86 microcomputer with monitor (e.g. HP 82913A monitor) giving 24 rows of 80 column monochrome display. Alternatively: the Hewlett Packard 87 microcomputer, which has a built in monitor.
2. Single disc drive, e.g. HP 9130 A flexible disc drive using standard 5.25 inch floppy discs.
3. Dot matrix printer, e.g. CITOH model 1550.
4. Hewlett Packard advanced programming ROM.
5. Memory module to expand the 64 K basic memory of the model 86/87 by at least another 64 K.

FILE STRUCTURE

The main program file 'DNA' occupies 38 K of memory and is supported by the following data files:

1. 'SEQNCE', a list of the names of DNA sequences stored on the disc.
2. 'RE' A list of all the known type II restriction enzyme recognition sequences plus the name of one of the most used isoschizomers cutting at that sequence (this is used for option 5 below).
3. A data file for each sequence, stored and retrieved as a simple string, each nucleotide residue occupying 1 byte of disc storage.

OPERATION

All that the unskilled operator requires to know is the simple start up procedure (type LOAD "DNA", press the ENDLINE key, and then the RUN key). From then on he is prompted by simple instructions mostly requiring him to choose between a number of options, either by typing the number corresponding to the option required or by typing an appropriate letter, e.g. 'Y' or 'N' (Yes or No), or 'D' or 'P' (display or print).

At start up a list is displayed of which the following is an example
Enter the number corresponding to the option required

1. Enter/delete/merge/compare
2. BK-WT
3. BK-MM
4. pBR322
5. Alu repeat
6. Test
7. SV40
8. c-myc
9. Random
10. ApoE Human
11. HPRT
12. His H5
13. HBV
14. BTG
15. BK(WT)Enhancer
16. BK(WJ)Enhancer
17. BK(MM)Enhancer
18. BK-Pst
19. TRH
20. M13mp7
21. ASS
22. pAlOcat2

Entering any of the numbers except 1 loads the corresponding sequence from file for analysis. Entering 1 calls up a subroutine displaying the following five options:

Choose from the following options
1. Enter a new sequence from the beginning
2. Continue entering an incomplete sequence
3. Delete a sequence from the file
4. Merge two of the sequences on file
5. Compare two sequences for homologies

Option 1 and 2 enable the operator to type in a new sequence or continue entering an incomplete sequence from the keyboard, entering a maximum of 160 characters (two lines of the display) at a time, following which there is the choice of continuing entry or storing the total entered in that section onto the disc. Editing of each 160 character string uses simple cursor manipulation and does not delete characters entered after the mistake was made.

Option 3 deletes the required sequence file and reorders the data file 'SEQNCE' to eliminate the gap.

Option 4 enables two sequences entered under separate file names to be merged. Both sequences are loaded from disc into memory and with the assumption that both sequences represent circular molecules either all or part of each sequence can be merged into one continuous sequence by entering the appropriate cutting sites. A sequence can be linked up in either of the two possible orientations relative to the other sequence, with the complementary form of the reversed sequence being determined if appropriate. The merged sequence can be stored on disc if required either under a new file name or in place of an existing file, and is immediately available for sequence analysis: This feature of the program is particularly useful for prediction of
restriction enzyme recognition sequences in recombinant contructions
and for determination of the resulting fragment sizes given by con-
structs in differing orientations.

Option 5 performs a search of two sequences entered under sepa-
rate file names for homologies. It can also be used as a quick method
of searching a single sequence for repeats. The algorithm employed is
simple and uses the rapid 'POS' function, which returns the position of
one string within another in HP BASIC, to search one sequence for iden-
tities with sequential segments of the other sequence. The segment,
or window, length is specified by the user. The wider the window the
more rapid is the search and the lower the background noise, but the
ability to pick up low levels of homology becomes progressively compro-
mised. A comparison of two 5000 base pair sequences take 50 minutes
with a window of 5 and 8 minutes with a window of 10. Homology
searches of corresponding regions around the origin of replication of
two strains of BK virus DNA (1,2) are shown in Fig. 1, and illustrate
the extensive variations in arrangement of the enhancer elements
present in this region.

When a sequence has been entered from the disc, or a merged
sequence has been constructed, a list of analytical options (the main
menu) is displayed:

Choose from the following options
----------------------------------
1. Display or print DNA sequence
2. Display or print complementary sequence
3. Search for a nucleotide sequence
4. Search for all known restriction sites
5. Map sites
6. Convert to amino acid sequence
7. Alter nucleotide sequence
8. Base frequency analysis
9. Search for repeated sequences
10. Reinitialize data

Enter the appropriate number

Option one : as in most of the subsequent options, the operator
is asked whether he wants the sequence displayed or printed. He is
then asked if he wishes to display/print the whole sequence, or a speci-
fied part of it, and the sequence is then displayed or printed in the
following format, illustrated for a short test sequence in Fig. 2.

Option two : as for option one but the complementary sequence
Fig. 1. Search for homologies between DNA segments enclosing enhancer elements from two strains of BK virus, using window sizes of 2 or 9.

is computed and displayed/printed from the specified start and stop position in the 5' to 3' direction.

Option three: This enables the operator to search for restriction enzymes or other sequences of any length. Use of the 'POS' function produces almost instantaneous retrieval of the position of the search sequence within the main sequence. Searches for sequences containing arbitrary bases, e.g. for the restriction enzyme Bgl I which would be typed in as GCCNNNNNGGC, use a somewhat more complex algorithm and take somewhat longer, for example, a search for Bgl I in pBR322 takes about 10 seconds. After the positions have been found and displayed or printed the sizes of the fragments produced (if the search sequence was for a restriction enzyme) may be determined and displayed or printed in order from the start of the sequence and sorted with respect to size. Alternatively, the sizes of end-labelled partial digest fragments obtained by the procedure of Smith and Birnstiel (3) may be obtained. If two or more enzyme sites have been searched for, the request for fragment sizes will give the results for the composite enzyme digest, as in the example shown in Fig. 3.

Option four: a complete search for the position of all known type II restriction enzymes can be displayed or printed, accessing the data file 'RE', and for a sequence the size of pBR322, takes about 12.5 minutes. The data file is currently up to date with the restric-

Test sequence from 1 to 70

ATGGCCCTGC CTCTCCCCAG ATGCTACGAT GTAGTACGC TCTTTTATT CCTCTACTAG TACA......

Fig. 2. Format of DNA sequence printout or display
Fig. 3. Printout of a search of pBR322 for the recognition sites of the restriction enzymes Pst I and Taq I, and calculations of the fragment sizes given by a digestion with both enzymes.

Fig. 4. Plot of Pst I and Taq I sites mapped onto pBR322.
Termination codon map for pBR322

Frame 1

Frame 2

Frame 3

(Upper rows read 5' to 3', lower read 3' to 5')

Open reading frames (longer than 250 bases) found:

| Frame | From Base | To Base | Length | Pos. of AUG | Length AUG-ter | AUG in frame? |
|-------|-----------|---------|--------|-------------|----------------|---------------|
| 1 (5'-3') | 1071 | 1515 | 444 | 1179 | 336 | Yes |
| 2 (5'-3') | 94 | 1273 | 1179 | 259 | 1014 | Yes |
| 3 (5'-3') | 65 | 423 | 358 | 86 | 537 | Yes |
| 3 (5'-3') | 1808 | 2105 | 297 | 1817 | 288 | Yes |
| 1 (3'-5') | 870 | 489 | 381 | 786 | 297 | Yes |
| 1 (3'-5') | 4167 | 3294 | 873 | 4152 | 858 | Yes |
| 2 (3'-5') | 1120 | 622 | 498 | 1078 | 456 | Yes |
| 3 (3'-5') | 1973 | 1514 | 459 | 1910 | 396 | Yes |

Fig. 5. Search for open reading frames in pBR322. The graphical representation illustrates the positions of termination codons and the printout details the positions and sizes of the longest possible open reading frames.

Option six: This section offers a series of choices for identification and analysis of translation products of the DNA. Firstly, a search for open reading forms can be performed, the operator choosing the minimum acceptable size for the frame and whether he wants all in frame and out of frame AUGs displayed or only the first in frame AUG. An example of the graphics display and print out of the first in frame AUG for each open reading frame greater than 250 is shown for pBR322 in Fig. 5. Secondly, a specified part of the sequence can be translated in any of the three reading frames, in either direction, into the corresponding amino acid sequence, with the option of including the DNA sequence underneath, as illustrated in Fig 6 for 100 nucleotides of a randomly generated sequence. The standard prokaryotic and eukary-

Translation of Random sequence

Reading frame read from 1 to 88 in frame 1

Asp.Phe.Lys.Arg.Asn.Thr.Arg.Val.Cys.Ser.Glu.Gln.Ala.Ile.Pro.Pro.Phe.Ser.Val.Ile.
GAC TTC AAG AGA AAT AGC AGA GTA GTG GCC GAG GCT ATC CCG CCT TCC TTC ATA

Fig. 6. Printout of translation of 100 base pairs of a random sequence.
Yotic termination codons are depicted as *a* (amber), *u* (umber), and *o* (ochre).

Translation in the reverse direction is automatically performed if the specified stop position has a lesser value than the start position. When translation is completed the following list of choices for analysis of the polypeptide sequence is displayed.

Choose from the following options

1. Plot frequency of basic and acidic amino-acids
2. Hydrophobicity plot
3. Codon usage
4. Copy plot on printer
5. Return to main program

Examples of the plots for frequency of basic and acidic amino acids, hydrophobicity and codon usage are illustrated in Fig. 7 for the coding region of the human hypoxanthine phosphoribosyl transferase gene (5). The hydrophobicity plot uses the normalised consensus values of Eisenberg (6) and averages over any specified number of bases. The plots are displayed on the monitor and choice four enables the contents of the graphics screen to be output to the dot-matrix printer.

Option seven: this enables four types of alteration to be made to any part of a sequence in system memory: alterations of single bases, insertion of any number of bases up to 80 at any one time, deletion of any number of bases, and renumbering of the whole sequence from a new starting position. After the changes to the sequence have been made, the altered sequence can be stored if desired under the original file name or under a new one, and the program returns to the main options list for analysis of the altered sequence. The deletion option is particularly useful if a complete restriction search is required on only a part of a sequence on file.

Option eight: This prints the absolute numbers and proportions of each base correct to one decimal place in any specified part of the sequence. If required, an incremental graphical representation can be obtained, as shown in Fig 8.

Option nine: this searches through the sequence in system memory for perfect repeats of any specified length greater than five bases, and returns the positions of the repeats, together with the preceding
Fig. 7. Graphical representations of analyses of the product of translation of the coding region of an HPRT cDNA sequence. From above down: plot to emphasize regions rich in basic or acidic amino acids; hydrophobicity plot averaged over 10 bases; codon usage tabulation.

and following six bases of each repeat.

Option ten: The main use of this option is to rapidly restore the initial conditions after searching for several restriction enzyme sequences with option three, otherwise fragment numbers will accumulate with each sequence searched for and found, and maps constructed with option five become overly complex.
Fig. 8. Incremental plot of relative base frequencies in the 654 base pair coding region of the human HPRT gene.

DISCUSSION

The minimal hardware required by this program provides for maximum economy and convenience without significant loss of performance or quality of output. The graphics dump routine provides an adequate hard copy of the graphical representation of results on the dot-matrix printer without having to resort to the expense and extra bulk on the workbench or office desk of a plotter and its associated ROM and software. One of the principle aims of this system is to provide a practical and reasonably portable system which is always ready to hand for the researcher at a moments notice. A minor disadvantage of the system is that although video display and printer are both 80 columns wide the graphics dump routine foreshortens the plot significantly. For most of the plots this is of little consequence, but for the circular mapping routine it is necessary to display a request as to whether the plot is intended for copying onto the printer. If the answer is yes then the plot appears on the screen as an oval in order to appear on the hard copy as a circle.

The random access memory of the system as described will enable a sequence as large as the 16,500 base pair mammalian mitochondrial DNA sequence to be analysed as a single unit. The memory capacity of the HP model 86/87 can be expanded readily to three times the 128 K used here by plugging in further memory modules and would only require very minor software modifications to utilise this added capacity. However, the 128 K of the described system was found to be quite adequate for all day to day requirements on molecules up to size of mammalian mitochondrial DNA.
In addition to the main program are 3 short additional programs: 'SEQEX', for transferring individual DNA sequences between discs; 'RE3 TRANSF', which transfers the list of restriction enzyme between discs, and 'SEARCH', a program for updating or editing the restriction enzyme file and with a procedure for searching a short sequence of less than 150 bases, typed in from the keyboard, for all known restriction enzyme recognition sequences. Data files included contain as standard with the program an up to date list of restriction enzymes and their recognition sequences for use by option one of the main program and by 'SEARCH', and two DNA sequences, a short test sequence for practising program operation, and a 5000 base randomly generated sequence which forms a useful basis for comparison with naturally occurring DNA sequences. This leaves space in mass storage for about 200,000 bases of DNA sequence for the operator's use.

ACKNOWLEDGEMENTS

Financial assistance from the MRC and the University of Cape Town is gratefully acknowledged.

REFERENCES

1. Yang, R.C.A. and Wu, R. (1979) Nucleic Acids Res. 7, 651-668.
2. Seif, I., Khoury, G. and Dhar, H. (1979) Cell 18, 963-977.
3. Smith, H.O. and Birnstiel, M.L. (1976) Nucleic Acids Res. 3, 2387-2388.
4. Roberts, R.J. (1985) Nucleic Acids Res. 13 Suppl, r165-r200.
5. Jolly, D.J., Okayama, H., Berg, P., Esty, A.C., Filpula, D., Bohlen, P., Johnson, G.G., Shively, J.E., Hunkapiller, T., and Friedmann, T. (1983) Proc. Natl. Sci. USA 80, 477-481.
6. Eisenberg, D. (1984) Ann. Rev. Biochem. 53, 595-623.