Volume 12 Number 1 1984 Nucleic Acids Research

# A DNA sequence analysis package for the IBM personal computer

L.Mark Lagrimini, Steven T.Brentano and John E.Donelson

Department of Biochemistry, University of Iowa, Iowa City, IA 52242, USA

Received 8 August 1983

## ABSTRACT

We present here a collection of DNA sequence analysis programs, called "PC Sequence" (PCS), which are designed to run on the IBM Personal Computer (PC). These programs are written in IBM PC compiled BASIC and take full advantage of the IBM PC's speed, error handling, and graphics capabilities. For a modest initial expense in hardware any laboratory can use these programs to quickly perform computer analysis on DNA sequences. They are written with the novice user in mind and require very little training or previous experience with computers. Also provided are a text editing program for creating and modifying DNA sequence files and a communications program which enables the PC to communicate with and collect information from mainframe computers and DNA sequence databases.

## INTRODUCTION

Although it has only been two years since the first issue of Nucleic Acids Research devoted solely to the use of computers in DNA sequence analysis (Vol. 10, No. 1 (1982)), there have been several changes in the field. First, the number of researchers collecting DNA sequence data has increased dramatically. Previously, only those laboratories which were in a position to make a major commitment to DNA sequencing ventured into the area. Today any laboratory can purchase a relatively inexpensive kit and begin to sequence DNA fragments almost immediately. This has created an increasing abundance of DNA sequences and has led to the establishment of several databases enabling the researcher to access hundreds of thousands of base pairs of DNA sequence (1-4). Another development in the last two years has been the increasing availability of inexpensive yet powerful personal computers. This has decreased the necessity of sharing time on large computers with the attendant difficulties of downtime, accessability and cost. The personal computer has the added advantage of being portable which prevents the problems encountered when changing computer systems during relocation. Finally, along with the recent growth of microcomputers used in scientific

applications, there has been a parallel growth in laboratory software which increases the range of tasks the personal computer can perform.

Two years ago IBM, the world's largest computer manufacturer, had just introduced its version of the personal computer, the IBM PC. Although not the first microcomputer on the market, the IBM PC is rapidly becoming an industry standard for personal computers. Its speed and memory capacity make the PC an excellent substitute for the large and expensive mainframe computers for which the early DNA sequence analysis programs were written (5,6). The IBM PC's impressive graphics capabilities also obviate the need for expensive graphics terminals and hardcopy units. We have taken advantage of these features in developing a software package called "PC Sequence" (PCS) for DNA sequence analysis. Based on our experience with other DNA sequence analysis programs (5-7) during the past eight years, we tried to design the package in such a way that it combines the best features of previous programs in a format that can be easily and conveniently accessed by the novice user.

All of the programs are written in compiled IBM PC BASIC. PCS is designed to run locally on the PC, but also enables the PC to be operated as a terminal to communicate with other computers. With a modem telephone hookup these programs can be run using DNA sequences obtained from sequence databases (1-4). PCS can carry out the following operations: format given sequences into numbered double stranded output, identify the cleavage sites of all commercially available restriction enzymes within a sequence, determine the location of any user supplied DNA string, either invert or calculate the opposite strand from the entered sequence, translate the entered sequence into the three possible amino acid reading frames, translate a single reading frame into single letter amino acid code, calculate the amino acid composition, base composition, protein molecular weight and codon usage from the protein coding strand, create a linear restriction map in graphics from the entered sequence, and display a dot matrix which gives a rapid comparision of two entered sequences. PCS also includes a communications program for displaying and collecting sequences from other computers or databases. A text editor facilitates entering of DNA sequence from the keyboard and use of the special characteristics of the EPSON MX-80 printer. All of these operations are accessed from a main menu screen containing information about using each of the programs. In addition to screen viewing most of the individual analysis programs allow output to be saved to disk, or obtained as a hardcopy print.

The following is a list of the hardware that is required to run PCS:

IBM PC + 64K bytes random access memory
2 x 320K byte double sided disk drives
B/W or Color monitor with 80 x 24 character resolution
Color graphics adaptor
IBM/MX-80 dot matrix printer (with GRAFTRAX)

Optional hardware:

Asynchronous communications adaptor
Modem

Optional software:

Word processor, e.g. Wordstar
IBM BASIC compiler
Terminal emulation program

## OPERATION

PCS routines are stored on one double sided disk which is operated from disk drive A. All data files containing DNA sequence are assumed to be in drive B, and new files created (such as from the editor, sequence collected from other computers, or other PCS disk output) will be saved to drive B. We have utilized the IBM DOS batch file "AUTOEXEC.BAT" on the program disk to automatically run the sequence analysis programs when the PC is turned on or after the system is reset. If the PC is already on, the programs can be initiated by typing in the command "PCS". When the program is started, a menu of individual programs can be run, or information about PCS and its operation can be obtained. For example, when PCS is loaded as described above, it begins operating automatically with the menu:

```
                IBM PC DNA SEQUENCE ANALYSIS PACKAGE

1.  PCFORMAT: Formats an entered sequence into numbered double stranded
              output
2.  PCRES:    Searches for all commercial restriction enzyme sites or
              user supplied string
3.  PCTRANS:  Determines three letter amino acid code for all three
              reading frames, or single letter code for 1 reading frame
4.  PCOPSTR:  Produces opposite strand from entered sequence
5.  PCINVERT: Inverts entered sequence
6.  PCCALC:   Calculates; amino acid composition, base composition,
              protein MW, and codon usage
7.  PCMAP:    Creates a linear restriction map of entered sequence in
              graphics
8.  PCMATRIX: Compares two sequences in dot matrix format ( from either
              DNA or single letter amino acid code)
9.  PCCOM:    Communications program for file transfer and work
              station use
```

```
10. PCEDIT:   General purpose text editor and MX-80 printer program

          Press the number for the program you wish to run
          Press "I" for information
          Press <esc> to Quit

  >_
```

The user now selects any of the options, receives information about the programs, or exits from PCS. For example, if PCTRANS is selected by pressing the "3" key at the ">" prompt, PCTRANS is then loaded and run, asking the user for the necessary information as needed:

```
This is a program to translate DNA sequence into its amino acid coding
sequence for all three reading frames.

Enter the name of the file which contains the sequence to be translated.
>B:PUC8.SEQ

Would you like this output to disk?  (Y or N)  >Y

Enter the name of the drive and file which will contain the translated
sequence.
>B:PUC8.TRN

Would you like this output to the printer?  (Y or N)  >N

Translation start position.  >

Translation stop position.  >60

The sequence loaded is 2720 bases long.

Asp   Val   Thr   His   Ser   Cys   Thr   Gln   Leu   Ile
  Met   ***   Pro   Thr   Arg   Ala   Pro   Asn   ***   Ser
    Cys   Asn   Pro   Leu   Val   His   Pro   Thr   Asp   Leu
G A T G T A A C C C A C T C G T G C A C C C A A C T G A T C
                  10                  20                  30

Phe   Ser   Ile   Phe   Tyr   Phe   His   Gln   Arg   Phe
  Ser   Ala   Ser   Phe   Thr   Phe   Thr   Ser   Val   Ser
    Gln   His   Leu   Leu   Leu   Ser   Pro   Ala   Phe   Leu
T T C A G C A T C T T T T A C T T T C A C C A G C G T T T C
                  40                  50                  60
```

The user has entered "B:PUC8.SEQ" which is the file containing the DNA sequence to be translated in drive B. A copy of the output will also be stored on disk B using the name "PUC8.TRN", but no printed output is requested at this session. Pressing return at the start position prompt causes the program to start at nucleotide position 1, while number 60 has been specified as the stop site. The sequence is then loaded, the length reported, and the three reading frames printed simultaneously to the screen and the disk along with the corresponding numbered DNA sequence. After PCTRANS is finished, it returns to the beginning where the user then may run the program again, go back to the menu screen to choose another option, or exit from PCS.

SHORT DESCRIPTION OF THE INDIVIDUAL PROGRAMS

1. PCFORMAT calculates the opposite strand from single stranded input

and prints the sequence as double stranded output with numbering at 10 base pair intervals. The following is an example of the output:

```
         10         20         30         40         50         60
GCTAGCTAGT GGTATGGCAG TAATTTGGGA TAGCGATTTT AAGCGCCTGC TAGGGCCTAC
CGATCGATCA CCATACCGTC ATTAAACCCT ATCGCTAAAA TTCGCGGACG ATCCCGGATG
```

As with all of the programs the output can be directed to the disk for permanent storage and/or to the printer for an immediate hardcopy.

2. PCRES will search a single stranded DNA sequence for the locations of 94 commercially available restriction enzyme sites. A data file called "resenz.dat" contains a list of the restriction enzyme recognition sites. This file can be easily edited by the user to include new restriction sites as more enzymes become available. The output of this program displays the name of the enzyme, the recognition site, the location of the site with respect to the origin and the sizes of the DNA fragments that will be generated upon cleavage. A second function provides the option of searching for any DNA string that is entered by the user. For example, if the user wishes to search a region for the string 'TATA', this sequence is entered at the start of the program. If the user wishes to search a sequence for a number of user supplied strings, a data file can be created using the editor which contains these strings. The following is an example of the output from PC RES.

```
pSVP12.seq is 9663 base pairs
 .
 .
 .
PVU I / CGATCG                  LOCATION                SIZE

                                 5860                   9663

PVU II /CAGCTG                  LOCATION                SIZE

                                 188                    2011
                                 4192                   4004
                                 7840                   3648
 .
 .
 .
NO MATCH WAS FOUND FOR THE FOLLOWING ENZYMES

ACC Ib/GTCTAC      APA I/GGGCCC      BGL I/GCC-----GGC
BST E II/GGT-ACC   SST I/GAGCTC      XBA I/TCTAGA
 .
 etc.
 .
```

3. PCTRANS determines the predicted amino acid sequence in all three reading frames from the DNA sequence. The user can supply the start and

stop positions for the translations. Termination codons are represented by "***". This program will also create a single letter amino acid sequence for one reading frame, which terminates at a stop codon or at the end of the sequence. The output of this program is demonstrated in the sample session above.

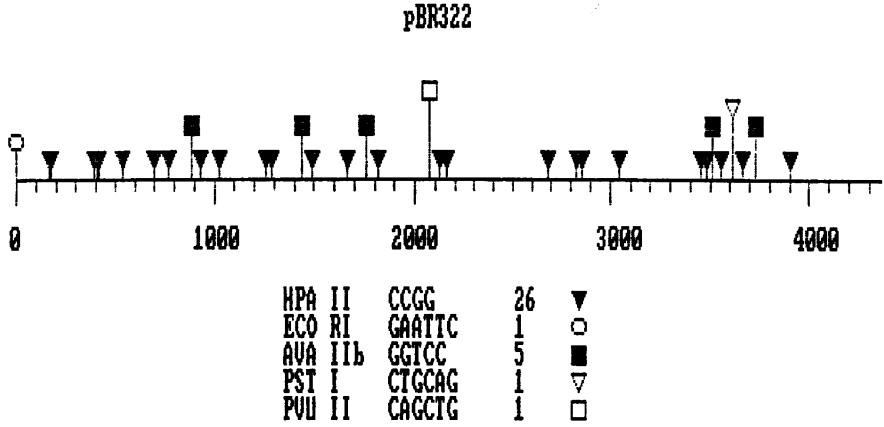
4. PCOPSTR creates a disk file of the opposite strand from a single strand input sequence. This allows for quick alignment of sequences when they are entered in the opposite orientation.

5. PCINVERT reverses the sequence so that the 5' end becomes the 3' end and vice versa. The following is an example of PCOPSTR and PCINVERT:

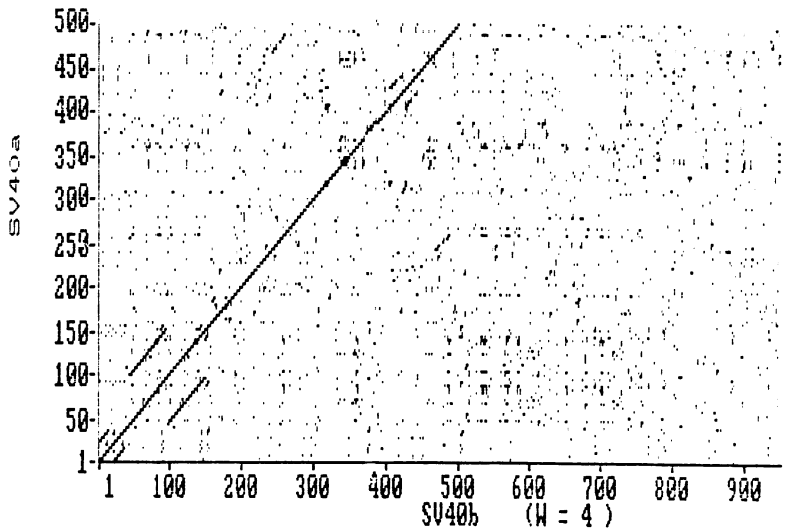
| | PCOPSTR | PCINVERT |
|---|---|---|
| original | opposite | reverse |
| ATCGGATCGC | TAGCCTAGCG | CGCTAGGCTA |

6. PCCALC determines the codon usage, the amino acid composition by weight and number, the base composition, and the calculated protein molecular weight from the entered DNA coding sequence. The user supplies the start point for translation, and the translation will continue to either a user specified stop position, the first stop codon encountered, or the end of the sequence.

7. PCRESMAP utilizes the graphics capabilities of the IBM PC to generate a pictoral restriction map from the entered sequence. This program will display the locations and number of cleavage sites for up to 10 different restriction enzymes at a time. It is also possible to quickly generate a hardcopy output on IBM or EPSON dot matrix printers with GRAFTRAX plus. An example of a map generated on pBR322 is shown below:



8. PCMATRIX is a dot matrix program (7), which compares two DNA sequences in a two dimensional plot. The two sequences compared can total 16,000 nucleotides in length. The user may select the start and stop positions of the comparison and choose a window for scoring matches. For example, at a window of 5 only a perfect match of 5 nucleotides between each sequence will score a match. Identical sequences are denoted by a diagonal line with a positive slope, direct repeats are denoted by two diagonal lines with the same slope as the main line. Deletions and insertions can be detected by a jagged line, and inverted repeats are shown (using a window of one nucleotide) by diagonal lines with negative slopes. The user can also use the PC function keys to select an area from the plot to either replot at higher resolution or display the two DNA sequences from a region on the plot to examine the location and extent of homology. A hardcopy printout of the screen can also be quickly obtained. This program also contains a dot matrix sub-program, MATRIXAA, for comparing two proteins entered in single letter amino acid codes as from PCTRANS. The following dot matrix shows part of the SV40 late gene region compared with itself using a window of 4. The 21 and 72 bp repeats are shown in the lower left corner.




9. PCCOM is a communications program which will connect the IBM PC with other computers through a serial adapter (asynchronous communications adapter), provided that the other computer uses standard communication

protocols, character codes and control characters. The user will be able to log onto other computers and use the PC as a terminal. This will permit transfer of files from the other computer to the PC for storage on disk which prevents retyping data already present on another system. For example this program will permit communication with DNA sequence databases such as the Dayhoff or Los Alamos databases (1,2). Sequences can then be collected by the PC and stored on disk if permitted. DNA sequences obtained from databases or other computers can be used with all of the analysis programs.

10. PCEDIT is a text editor that is easy to learn and permits the user to create new text files or make modifications to existing files. Usually this program will be used to enter and modify DNA sequences for use in the analysis programs, but is also suitable for limited word processing applications. When an MX-80 printer equipped with GRAFTRAX is used, this program can easily access its special functions such as underlining, italics, sub- and superscripts, compressed text, etc. These allow enhanced printouts from either original text or the output from any of the analysis programs.

All of the sequence analysis programs are designed such that only the capital letters A,G,C, or T are accepted, along with "-" dash for ambiguous bases. Because of this feature, input files may include comments and descriptions as long as they do not use the above characters.

The use of BASIC's error handling subroutines enables us to provide helpful messages to avoid premature program termination if a mistake is encountered while executing PCS. For example, if a nonexistent file name is typed, the program will respond with "There is no file by that name. Please re-enter file name, and include correct drive letter if necessary". The user is then allowed to re-enter the file name. There is also error trapping for no disk in drive, printer off, out of paper, disk full, sequence too long for the amount of machine memory, etc.

## DISCUSSION

We have developed PCS as a comprehensive DNA sequence analysis software package from the IBM PC. The user will be able to carry out all routine DNA sequence analyses using these programs. We have also included an editor and communications program to assist in entering sequences by hand or directly from other computers. These programs are written in compiled BASIC and run with the times shown in Figure 1.

| | Run Time (Min.) | | |
|---|---|---|---|
| Program: | Print to screen | + Disk | + Disk + Printer |
| PCFORMAT | 2:08 | 2:30 | 7:25 |
| PCRES | 2:12 | 2:46 | 11:00 |
| PCTRANS | 3:16 | 5:10 | 17:20 |
| PCOPSTR | :33 | 1:03 | 2:35 |
| PCINVERT | 1:00 | 1:37 | 2:48 |
| PCCALC[i] | 1:01 | 1:20 | 1:40 |
| PCRESMAP[ii] | :53 | --- | 4:10 |
| PCMATRIX[ii] | 9:00 | --- | 12:10 |

Fig. 1. Execution times of PCS programs on a standard IBM PC using pBR322 as the input sequence (length = 4362 bp) with the following exceptions and notes: (i) PCCALC was used on the ampicillinase coding region of pBR322 (length = 789 bp), (ii) PCMATRIX compared a 500 bp sequence with a 1000 bp sequence using a window of 5.

The IBM PC's 16 bit microprocessor enables programs to run considerably faster than on other popular 8 bit microcomputers such as the Apple II. PCS employs advanced error trapping to decrease the probability of premature program termination due to user error. Since many scientists sequencing DNA do not have an extensive background in computer programing, we have attempted to make these programs user friendly. PCS begins by displaying a menu of routines with all of the functions available. The menu program which controls the execution of all the analysis programs also contains three screen pages of additional information about the programs, and helpful directions for using the programs. In addition, thorough documentation is provided which contains details and examples of all of the programs and a troubleshooting guide for help in time of difficulty.

The output is concise and easily interpretable. Changes can be made in the output by the editor for user customizing. The programs with graphics output can quickly produce hardcopies of the restriction maps and dot matrices on the inexpensive IBM/MX-80 dot matrix printer.

The PCS package is available from the authors for a small cost to cover materials and postage.

ACKNOWLEDGEMENTS

This work was supported by USPHS research grant AI 18954.

REFERENCES

1. Orcutt, B.C., George, D.G., Fredrickson, J.A. and Dayhoff, M.O. (1982) Nucl. Acids Res. 10:157-174.
2. Kanehisa, M.I. (1982) Nucl. Acids, Res. 10:183-196.
3. Walgate, R. (1982) Nature 296:696.
4. Lewin, R. (1982) Science 217:817-818.
5. Staden, R. (1977) Nucl. Acids Res. 4:4037-4051.
6. Queen, C.L. and Korn, L.J. (1980) Methods in Enzymology 65:595-609.
7. Novotny, J. (1982) Nucl. Acids, Res. 10, 127-131.