Volume 12 Number 1 1984 Nucleic Acids Research

DNA sequence analysis on the IBM-PC

William F.Schwindinger and Jonathan R.Warner

Departments of Biochemistry and Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Received 17 August 1983

ABSTRACT

We have developed, for the IBM-PC microcomputer, a menu driven, interactive set of programs which provide the functions routinely used for DNA sequence data analyses.

INTRODUCTION

We report here a menu driven, interactive set of programs for storing, manipulating, examining, and comparing DNA sequences on the IBM-PC. The programs are written entirely in BASIC (1), which makes them immediately available to every user of the IBM-PC, and readily adaptable to each individual user's needs.

The programs are designed with economy and simplicity in mind. The available programs perform those functions which are most essential and the most frequently used for recording, assembling and interpreting DNA sequence data. Their design is modular, which allows for individual programs to be upgraded or replaced, and provides for maximum economy in memory usage.

Although these programs provide capabilities which are similar to available sequence data analysis packages (2,3), we hope that in developing them specifically for the IBM-PC we have taken advantage of some of its unique features, and that by choosing BASIC as a programming language we will encourage users to adapt the programs to suit their individual needs.

METHODS

The programs were developed on an IBM-PC with a monochrome display, 128K of memory, two double sided, double density 5 1/4" floppy disk drives, a 128K RAM+ disk (Seattle Computer), and an 80 column dot matrix printer (C.ITOH). As a minimum configuration the programs require 64K of memory, one disk drive, and any IBM compatible printer.

RESULTS AND DISCUSSION

The programs are menu driven and to a certain extent self-documenting. The two page main menu is shown in Figure 1. The individual programs are described below, along with the amount of memory each occupies when manipulating 3,000 nucleotides, and their approximate execution times.

Enter is used to type sequence data into the computer. It is designed to facilitate typing with one hand while following a printed sequence or the bands on a gel with the other. To this end it provides the option of using the typewriter keyboard or the numeric keypad to enter the data. In addition to displaying each nucleotide as it is entered, it produces a different tone for each, so that the user's attention need never be diverted from the data. Finally it provides some primitive editing capabilities for correcting errors caught in the process of entering the data.

The sequence data is stored in a diskette file, in a format which is upwardly compatible with programs such as SEQ (4). Each file may contain up to ten comment lines, which must begin with a semicolon. These must be followed by a title line. The

A.

KEY	FUNCTION	DESCRIPTION
F1	MENU	Prints page two of this list.
F2	ENTER	Accepts your sequence from the keyboard.
F3	EDIT	Allows you to make changes in a file.
F4	READ	Reads a file containing your sequence.
F5	PRINT	Prints your sequence.
F6	REST	Finds restriction enzyme sites in your sequence.
F7	SEARCH	Finds all positions of a specified sequence.
F8	TRANSLATE	Translates your sequence in all possible frames.
F9	COMPII	Compares two files.
F10	CONTROL	Changes the parameters which control program flow.

Depress the appropriate function key to continue.

B.

KEY	FUNCTION	DESCRIPTION
F1	MENU	Prints page one of this list.
F2	FANCY	Prints and translates your sequence in one frame.
F3	USAGE	Determines the codon usage in your sequence.
F4		Not implemented.
F5		Not implemented.
F6		Not implemented.
F7		Not implemented.
F8	TYPE	Prints a specified file.
F9	PRINTER	Changes the typeface used by the printer.
F10	CONTROL	Changes the parameters which control program flow.

Depress the appropriate function key to continue.

Figure 1. Page one (A.) and page two (B.) of the main menu.

sequence data follows, consisting of upper case letters: A, C, G and T denoting the four nucleotides and N denoting an undetermined base. The file is terminated with a 1. This program puts the time and date the file was created on the first comment line.

Edit is a specialized, full screen text editor, which allows for nucleotides to be inserted, deleted or corrected anywhere in the sequence. It also allows for revisions of the title or comment lines. Finally it puts the time and date the file was last edited on the first comment line.

Read moves sequence data from diskette files into the active memory of the computer, where it can be manipulated by the other programs. It may be used to access all or any portion of the sequence in a diskette file. It also provides the option of generating the complementary strand. Sequences of up to 15,000 nucleotides can be manipulated by these programs, but in practice sequences longer than 3,000 nucleotides are processed quite slowly. Read occupies less than 16K of memory, and requires 0.2 seconds for every 100 nucleotides to be accessed.

Print lists a sequence and numbers it from the first nucleotide. This establishes the numbering of the sequence for the other programs. This program occupies less than 16K of memory, and requires approximately 3.5 seconds for every 100 nucleotides of sequence to be printed.

Rest searches the sequence for 62 common, palindromic restriction enzyme recognition sites and prints the positions of the nucleotides before which each enzyme would cut. This program occupies less than 24K, and requires 3.5 seconds for every 100 nucleotides of sequence to be searched.

Search provides the capability of searching for an oligonucleotide sequence within a larger sequence. It asks for the oligonucleotide sequence and for the degree of homology required. It prints each occurence of the sequence and the degree of homology observed. It requires 4.5 seconds for every 100 comparisons to be made, and thus will find all regions at least 70% homologous to a given decanucleotide in a 1000 bp sequence in about 3 minutes.

Translate prints the putative amino acid sequence for each of the three reading frames. It occupies less than 16K of memory, and requires 17 seconds for each 100 nucleotides of sequence to be translated.

CompII is used to compare two files of sequence data. It prints the two sequences, one above the other, placing a dash in the second sequence where it is identical to the first. In this way it can readily establish the accuracy of data entered into the computer in duplicate. It also allows for nucleotides to be inserted, deleted or changed anywhere in either sequence, and for either of the two sequences to be saved.

Another element of CompII is its ability to move the second sequence relative to the first and to merge the two sequences. This capability is useful for assembling a

sequence from the data derived from a number of sequencing gels.

Control is used to redirect the input and output of the other programs. It defines the drive containing the sequence data, and routes the output to the screen, to the printer, or to a diskette file.

Fancy is designed to print the sequence in a format which should be suitable for publication, after minimal editing. It prints the nucleic acid sequence separated into groups of 10 nucleotides in non-coding regions and groups of 3 in coding regions, prints the derived amino acid sequence, and numbers the sequence assigning +1 to the first nucleotide in the first exon. It is capable of handling a gene with up to ten exons. The program occupies less than 40K, and requires 10 seconds for every 100 nucleotides to be processed.

Usage determines the pattern of codon usage for a given reading frame, and the amino acid composition, molecular weight and charge of the putative protein product. It occupies less than 32K of memory, and requires approximately 6 seconds for every 100 nucleotides in the frame.

Type provides the capability of printing any diskette file. It is most useful for printing a file created by fancy, after any necessary editing has been done.

Printer is used to set the type of print produced by the printer. It sets the number of characters per inch, the number of lines per inch, and the position of the left margin. Although it is necessarily printer-dependent, it should be adaptable to most printers.

These programs are available upon request. Documentation and hints for revising the programs will be included. To obtain a copy send a blank 5 1/4" floppy diskette to W.F. Schwindinger.

ACKNOWLEDGEMENTS

This work was supported by NIH grants GM 25532 and CA 13330. W.F.S. is a Medical Scientist Trainee under an NIH grant GM T32 7288.

REFRENCES

1. IBM Personal Computer BASIC, Version 1.10 (1981).
2. Pustell, J. and Kafatos, F. C. (1982) Nucleic Acids Res. 10, 51-61.
3. Fristensky, B., Lis, J. and Wu, R. (1982) Nucleic Acids Res. 10, 6451-63.
4. Brutlag, D. L., Clayton, J., Friedland, P. and Kedes, L.H. (1982) Nucleic Acids Res. 10, 279-294.