Volume 10 Number 1 1982 Nucleic Acids Research

A DNA sequence handling program

A.D.Delaney

Department of Biochemistry, University of British Columbia, Vancouver, B.C., Canada V6T 1W5

Received 10 November 1981

ABSTRACT

A computer program that aids in recording, editing, and analysis of the base sequences of DNA and RNA is presented. A tape containing copies of the program and the user manual for it are available at cost.

INTRODUCTION

The development of modern methods for DNA sequencing has made the use of the computer necessary in handling and analysing the large amounts of data produced. A number of programs, some quite general[1,2,3], and others directed towards more specific ends[5,6,7,8,9,10,11], and a review[4] of what is available have been published. Two centralized sequence data banks have been made available to researchers via computer to computer telephone communication[12,13].

This paper presents a single program that contains most of the features desirable in a general purpose sequence handling system. It does not require that a user know the details of the computer system beyond startup procedures and the naming of data files. Once installed its use should be self evident to the computationally naive. It responds through the terminal to the user's instructions with answers, queries for more instructions or error messages. Thus the system is interactive and though batch use is possible, the program was not designed specifically with this in mind.

The program is designed to be implemented on most systems with minimum alteration. This is particularly important when groups with various computing equipment wish to use the same program without a great investment in time by a programmer. The programming language used, PASCAL, is not as widely accepted as FORTRAN, but its standard features include capability to handle files and strings of characters. It is a modern structured language, which makes its programs easier to write and understand and therefore to

debug and alter. PASCAL is also designed for compactness, so that its compiler is widely available on minicomputers.

Among the functions provided by the program are storage for established sequences, displaying of sequences, comparisons for homology within or between sequences,and searches for features such as specific sequences, repeated sequences and hairpin loops. Sequences to be searched for may include residues specified only as purine (R), pyrimidine (Y), or any (N), as well as specific nucleotides. These sequences presently are limited in length to 10,000 nucleotides, but it is very simple to change this, the actual limit depending on the computer system resources. Execution time and the amount of output from the program are also limiting factors, and these can be adjusted by parameters controlling the homology routines. The homology algorithms and their control are discussed in the user manual. The program provides facilities for the storage and handling of data accumulating during the active determination of a sequence including the construction of complementary strands to given sequences, comparison between sequences and the melding of overlapping sequences. Details can be found in the user manual.

THE DATA STRUCTURE

A number of devices and storage areas are used by the program during its operation. All instructions are sent to the program via the keyboard on the terminal. All error messages and prompts are written by the program on the terminal screen.

The program does all its computational work in three work areas. These can hold two sequences and a list of oligonucleotides. Most computations are done on the sequence in the primary work area, and those computations which require two sequences also use the backup work area. The oligonucleotide list area is the active area where short sequences to be used by the SEARCH command are kept.

A permanent storage file is required by the program where sequences are stored for future access and which is searched for sequences requested by the user. This file can be changed during the interactive session, thus the user can move data from one file to another.

Sequences can also be copied to the primary work area from the keyboard. Oligonucleotides can be copied from the sequence file to the oligonucleotide list area. Several other instructions can be used to move sequences about between the three work areas.

Information in the work areas is lost when the current session is terminated, whereas information in the permanent file is available for future use. Thus it is a good idea to transfer any sequences typed in or obtained by processing to the appropriate permanent file by OUTPUT instructions.

PROGRAM OPERATION

Operation of the program is controlled by instructions from the user and by a number of parameters specified by the user. Whenever the program requires an instruction it prompts by showing "NEXT COMMAND PLEASE" on the terminal screen. Most of the instructions are a single word, though some require a few words or numbers on the same line. The instructions are intended to describe their actions in abbreviated English and are outlined below. The words in upper case are the actual commands; these cannot be changed by the user. The words in lower case are information to the program from the user, for example sequence names and delimiting numbers.

GLOSSARY OF INSTRUCTIONS

GET name	copies the sequence of that name from the permanent sequence file to the primary work area
GET FROM KEYBOARD	copies a sequence from the keyboard to the primary work area
GET OLIGO name	copies the oligonucleotide of that name from the permanent sequence file to the oligonucleotide list area
GET OLIGO ALL	copies all sequences which are in the sequence file and which are less than 25 nucleotides long to the oligonucleotide list area
CLEARO	clears the oligonucleotide list
DELETE name	removes the sequence of that name from the permanent sequence file
OUTPUT	copies the sequence in the primary work area to the sequence file
PRINT	prints out the sequence
WORK	displays the names and lengths of the sequences in the three work areas
PAGE	advances the printer to the next page and

	prints the time and date at the top
LIST	lists the names of all the sequences in the permanent sequence file
NAME name	gives the sequence in the primary work area the name typed in on the line
COPY	copies the sequence in the primary work area to the backup area
SWITCH	exchanges the locations of the two sequences in the two sequence work areas
REVERSE	reverses the sequence in the primary work area
COMPLEMENT	generates the strand complementary to the sequence in the primary work area
EXTRACT lower upper (limits)	extracts the portion of the sequence between the indicated residues, keeping only the extracted portion in the primary work area
BREAK breakpoint	splits the sequence into two, putting the latter portion in the backup area
JOIN	makes one sequence by joining the backup sequence to the right end of the sequence in the primary work area
MELD	overlaps and joins the backup sequence to the primary sequence if the right end of the primary sequence is homologous to the left end of the backup sequence
CHANGE	makes corrections to the sequence
SEARCH	searches the sequence for the members of the oligonucleotide list, for example, a list of restriction sites.
SEARCH FOR seq, seq, seq,...	searches the sequence for the sequences on the instruction line
BASES	produces a table of base compositions
CODONS	produces a table of codon frequencies
TRANSLATE	prints a translation of the sequence into an amino acid sequence
TRAN2	prints the two sequences and their translations and lines them up

LINEUP	prints the two sequences lined up for easy comparison
HAIRPINS	searches for hairpins within the sequence
REPEATS	searches for repeats within the sequence
INVERTED REPEATS	searches for inverted repeats within the sequence
INVERTED DYADS	searches for inverted dyad symmetries within the sequence
COMPARE	searches the two sequences for regions of direct homology
COMPARE FOR COMPLEMENTS	searches the two sequences for regions of complementary homology
COMPARE FOR INVERTED REPEATS	searches the two sequences for regions of inverted homology
COMPARE FOR INVERTED COMPLEMENTS	searches the two sequences for regions of inverted complementary homology
STOP	ends execution of the program

THE SET INSTRUCTIONS

The parameters controlling program operation can be changed using the "SET" instruction. The sequence file and the printfile names can be changed at any time. The latter can be any file name or special names indicating the terminal screen or the printer.

There are a number of numerical parameters controlling the analytical routines. The maximum and minimum distance between homology searches can be set, i.e., these would control loopsize limits in a hairpin search. The quality of an acceptable homology can be controlled by setting the minimum number of matches, the minimum proportion of matches, the minimum length, and the maximum length of loopout in a homology.

Oligonucleotide searches for sequences with mismatches can be performed, and the search can be restricted to any of the reading frames if desired. Codon analyses and translations can be performed in any one or all three phases.

All numerical parameters have default values, so that they do not have to be set before each session. The default values can be found in the instruction manual.

PROGRAM IMPLEMENTATION

The program uses approximately 125 kilobytes of memory when executed on an Amdahl V6-II computer running under an MTS operating system. The Amdahl computer architecture is essentially identical to that of the IBM-370. The program requires 65 kilobytes, the PASCAL library routines 25, and the stack another 35. The space requirement for the stack can be reduced by changing the maximum sequence length which can be analysed. Further space savings may be possible if only the routines required for a given session are loaded. This size of program would require a medium sized computer or a small computer with a virtual memory system. One of the versions on the tape is completely in standard Pascal and should be implementable on any computer with sufficient memory and a standard Pascal compiler. The program has also been implemented by Dr. V. Ling, Department of Medical Biophysics, University of Toronto, on a Digital Corporation VAX-11 machine.

CONCLUSION

This program is reasonably complete, but improvements are being made constantly. High on the priority list for a future version are
(1) an internal program data structure invisible to the user,
(2) more sophisticated and efficient homology routines, and
(3) a MELD routine which can work on an unlimited number of sequences and provide error feedback to the user.

This publication makes the current version of this program available to interested researchers. The content of future versions will depend to a great extent on feedback from users. A detailed instruction manual, as well as the source code for the program is available for the cost of a magnetic tape and postage. Also included on the tape is a file containing the sequences of MS2, ØX174, pBR322, SV40, and the known restriction enzyme sites to May 1981.

ACKNOWLEDGEMENTS

Many thanks to Ian Gillam, Shizu Hayashi, Michael Smith, William Addison and Gordon Tener for suggestions concerning the program and for help with the manuscript. This study was supported by a Medical Research Council of Canada Grant (MT-1279), to G. M. Tener.

REFERENCES

1. Korn, L.J., Queen, C.L., and Wegman, M.N. (1977) Proc. Nat. Acad. Sci. U.S. 74, 4401-4405
2. Queen, C.L., and Korn, L.J. (1980) Methods in Enzymology 65, 595-609

3. Staden, R. (1977) Nucleic Acids Res. 4, 4037-4051
4. Gingeras, T.R. and Roberts, R.J. (1980) Science 209, 1322-1328
5. Staden, R. (1979) Nucleic Acids Res. 6, 2601-2610
6. Gingeras, T.R., Milazzo, J.P., Sciaky, D., and Roberts, R.J. (1979) Nucleic Acids Res. 7, 529-545
7. McCallum, P. and Smith, M. (1977) J. Mol. Biol. 116, 29-30
8. Staden, R. (1978) Nucleic Acids Res. 5, 1013-1015
9. Gingeras, T.R., Milazzo, J.P., Roberts, R.J. (1978) Nucleic Acids Res. 5, 4105-4127
10. Fuchs, C., Rosenvold, E.C., Honigman, A., and Szybalski, W. (1978) Gene 4, 1
11. Staden, R. (1980) Nucleic Acids Res. 8, 817-825
12. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C. and Orcutt, B.C. (1980) Science 209, 1182
13. Jordan, E. (1980) Science 210, 1074