Volume 12 Number 1 1984 Nucleic Acids Research

Computer programs for handling nucleic acid sequences

C.Keller, M.Corcoran and R.J.Roberts

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received 10 July 1983

ABSTRACT

Programs have been developed that will accurately display restriction enzyme maps, open translational reading frames and the results of specific searches on a high resolution graphics terminal (Retrographics VT640). Many of our earlier programs have been upgraded to allow online access to GENBANK, the NIH data bank of DNA sequences and also to deal with sequences of any length. Progress has also been made in the automation of the DNA assembly programs.

INTRODUCTION

For several years now we have been developing computer software to aid both in the assembly of DNA sequences and in the subsequent analysis of those sequences. Our recent efforts have been directed towards the development of software that gives a graphic representation of the information content of a sequence using a high resolution graphics terminal. The following is a brief description of the programs that have been developed at Cold Spring Harbor Laboratory over the last few years. All the programs described are written in FORTRAN 77 and were developed on a PDP-11/44 running the RSX-11M operating system. In general they are readily transported to a VAX, running VMS. The graphics options are designed for a VT640-Retrographics CRT terminal and a Bausch and Lomb HIPLOT DMP-29 multi-pen plotter.

PROGRAMS

AASEARCH will search a DNA sequence for regions able to code for a polypeptide of defined composition and length. The program translates the nucleotide sequence into the corresponding amino acid sequence and then searches all 6 reading frames for the occurrence of peptides, of the specified length, that have the required composition. The user can limit the search to any part of the input sequence and/or its reverse complement.

CLONE is used to the reconstruct recombinant DNA sequences (1).

CODTOT is an enhanced version of the program by Staden (2), which outputs the codon tables complete with the appropriate amino acids in the standard genetic code format. It also calculates precise molecular weights and amino-acid composition.

COMBINE is a program that combines a standard sequence file and a suitable information file (one created by the program INFO), to produce a new file that provides an annotated version of the sequence file (1). An extensive new menu allows the user to select a wide variety of output formats.

DSPLAY is a program which provides a simple graphic display of the reading frames present in a sequence (1). It now accepts sequences of any length.

DRIVER is an enhanced version of the program ASSEMBLER for building a complete sequence from primary data (3). Recent modifications allow many different sequencing projects to be handled at once and provide an interface to the programs M13 and SEQ, which provide data base management facilities for projects using the M13 cloning-sequencing strategy. The user can enter raw sequence data either by typing that data into a TRIAL file or directly from the autoradiograph using the digitizer option present in the program (4). The latter facility now uses a sonic digitizer (GRAF/BAR from Science Accessories Corporation) and has been enhanced by the addition of a voice synthesizer (TYPE'N TALK available from VOTRAX), which provides immediate confirmation of correct input. The program will align the primary data and create an OUTPUT file, that shows the alignment. A new meld function then reads this output file and condenses the two overlapping segments into a single sequence, prompting the user to resolve any discrepancies between the two pieces of data. The original entries in the REGION file are then deleted and replaced by the new meld. A record of all such melds is maintained as an archive of all primary data.

FIND searches data base files created by the programs SEQ, M13, and INFO and will give a condensed report of all sequence reactions carried out on each clone (1).

FORMAT is used to format sequence files into the form required by our programs. Input files must contain a one line header followed by the sequence in free format. Multiple entries can be present in the input file.

FRAGSIZE is a program that will calculate the molecular weights of restriction fragments based upon their gel mobilities. It is an implementation of the method devised by Southern (5).

GBINDEX is a program which reorganizes the files distributed by GENBANK, the NIH sequence data bank, for use by our system. Currently there are ten categories of sequence in GENBANK (Mammalian, Eukaryotic, etc...) and for each category three separate files are created. The first is a direct access index file which contains a list of the sequence headers for that category plus pointers to the appropriate record in the second direct access file in which the sequences themselves are stored. This file also contains the annotation provided by GENBANK. The third file created is just a sequential list of the headers in each category for use in our own on-line help facility. The sequences from GENBANK can then be accessed from most of our programs by a standard interface which reacts to the sequence file name GENBANK by offering a choice of categories and then taking the user sequentially through the files in the category chosen. Once a sequence is selected the user can choose to display the relevant annotation (a good source of literature references) or can merely retrieve the sequence itself for use by the program.

GSEARCH is a generalized search routine program. It will search a sequence file for all occurrences of a requested sequence which may include a variety of degeneracies at each position. The program can handle a sequence of any length and will also circularize the sequence if required.

HEADLIST is a program to list the headers in a file containing multiple sequence entries. It will also print the lengths of each sequence, the total number of bases in the file and the number of entries in the file.

INFO is used to create the information files used by the program COMBINE (1).

M13 is used to manage the data base of M13 clones obtained during a sequencing project (1).

MAPPER is used to find restriction enzyme sites in a sequence (1). New options within the program allow searches for all known restriction enzyme sites or for a particular subset of sites (eg. sites for all commercially available enzymes or a user-created list). Three major output options are available. The first is a linear listing of the sites as they occur in the sequence. The second is a double listing of sites, ordered by size or by position. This can optionally include the products of either single or multiple digests and, in the latter case, end-labelled fragments can be specifically marked. A new third option allows graphic output of restriction enzyme maps, either linear or circular, with a variety of choices for customizing maps. A hard-copy, suitable for publication, of any desired map can be made (Fig 1).

PEPTIDE translates an input DNA sequence, of any length, into the amino acid sequence. Options within the program aid in the analysis of peptides derived by cleavage with trypsin or other proteases. The amino acids at which the protease cleaves may be entered and their positions will be highlighted in the final output. In addition if specific amino acids, such as methionine, are labelled then the program will single out those peptides that contain this amino acid for display in printed or graphic form (Fig 2). Also all open translational reading frames can be displayed graphically, both at a CRT or as hard copy, as a high resolution version of the DSPLAY output or as a bar diagram (Fig3).

PRIMER finds restriction fragments which serve as potential primers for dideoxy sequencing (1).

PRIMERCHK compares an input oligonucleotide sequence with an input sequence file to find the positions of maximum homology. Only direct comparisons are made (ie. no allowances are made for looped out structures). The program is useful for checking potential homology between a synthetic oligonucleotide and the vector sequence to ensure that it will only hybridize to the intended sequence.

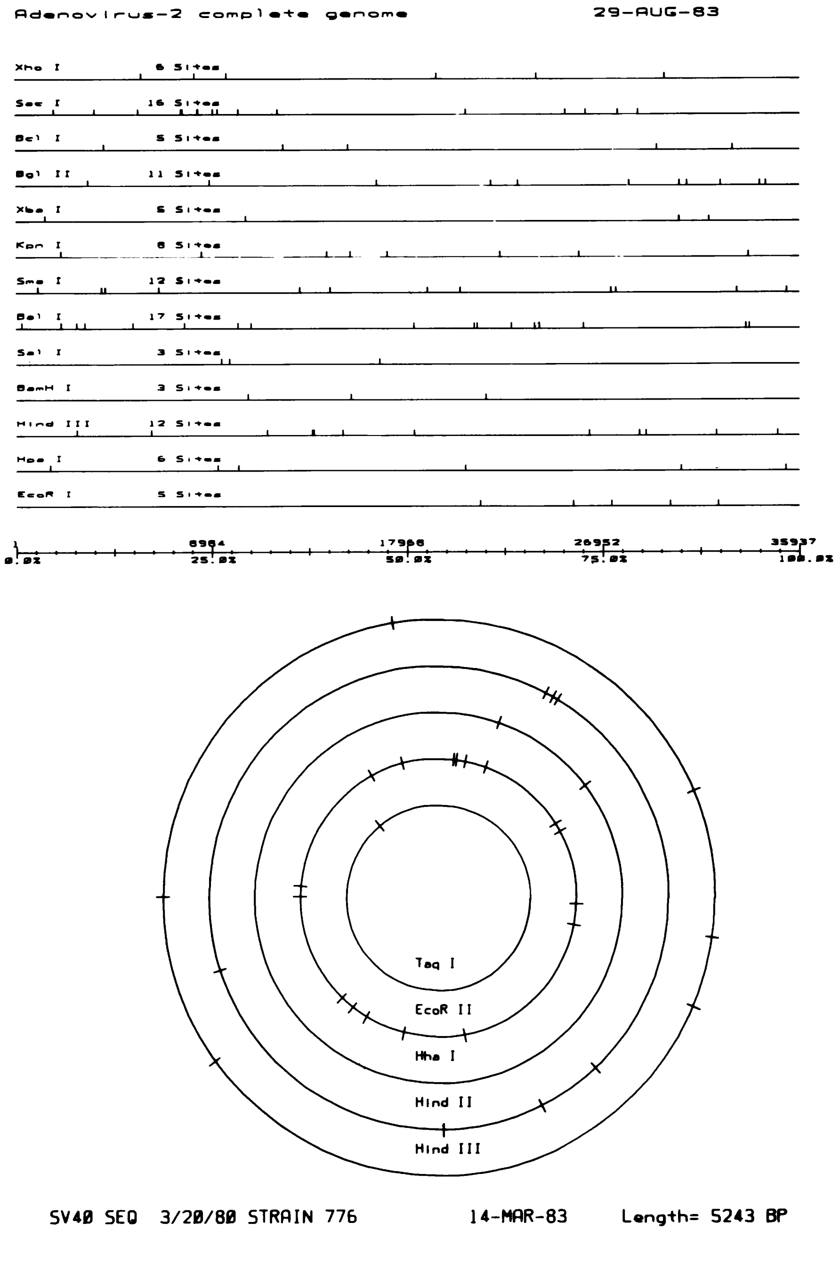
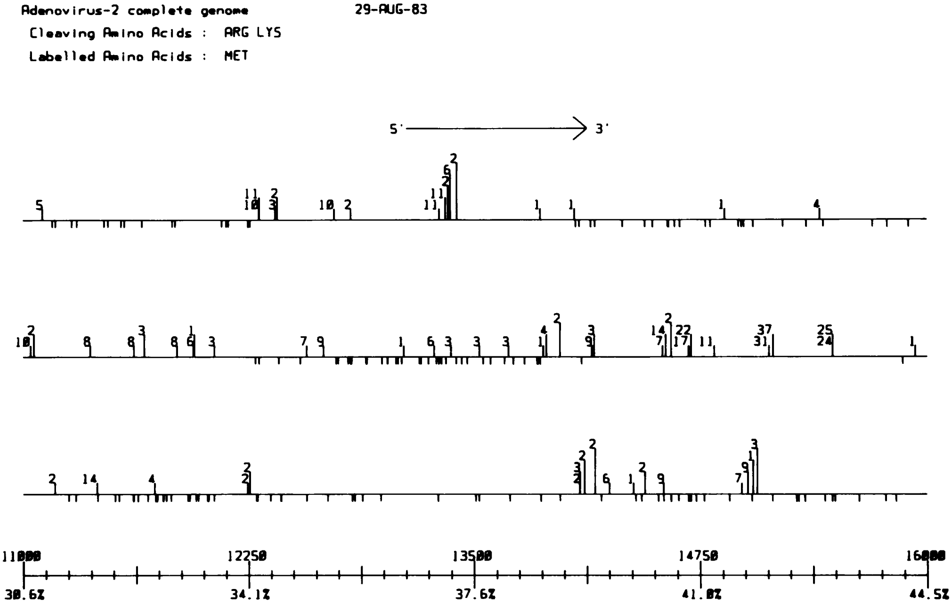


Figure 1. Examples of hard-copy output from the program MAPPER. Upper panel : a map of Adenovirus-2 DNA. Lower panel : a map of SV40 DNA. In this figure as well as in figures 2 and 3 the hard-copy is an exact copy of the graphic output displayed by the VT640 CRT terminal.



```
Adenovirus-2 complete genome          29-AUG-83
 Cleaving Amino Acids :  ARG LYS
 Labelled Amino Acids :  MET
```

Figure 2. Example of hard-copy output from the program PEPTIDE, showing the positions of methionine residues within tryptic peptides. All three reading frames are shown, with the vertical lines above each frame indicating a methionine residue at the indicated position within a tryptic peptide. The vertical lines below each reading frame show the positions of terminator codons within that reading frame. The scale shows absolute base coordinates above the line and percentage coordinates, based on 100% for the whole genome, below the line.

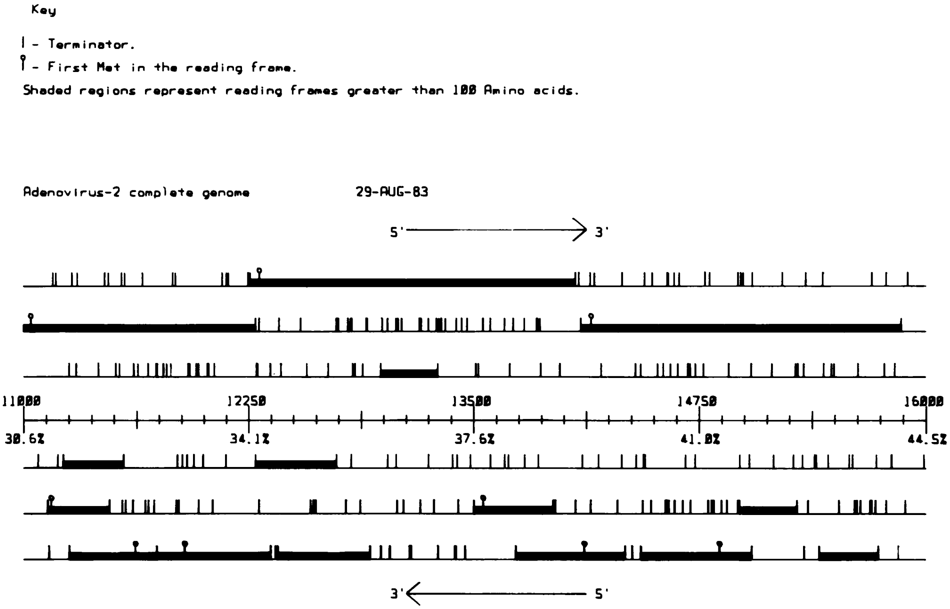
REVCUT provides a reverse translation of a protein sequence and determines which restriction enzymes will be unable to cleave the gene coding for that protein (1).

REVSEQ is used to provide a printed version of a sequence in a variety of different formats (1).

RFRAME finds all occurrences of initiators and terminators in a sequence and displays them in all six reading frames (1).

SCROLL is used to provide a detailed picture of how a complete sequence was derived by aligning the primary data against the final sequence. The output alignment is similar to that described elsewhere (6).

SEQ is used to manage the M13 sequence reactions carried out during a sequencing project (1).



Key

| - Terminator.

ǃ - First Met in the reading frame.

Shaded regions represent reading frames greater than 100 Amino acids.

Adenovirus-2 complete genome 29-AUG-83

Figure 3. Example of hard-copy output from the program MAPPER, showing the location of terminators and open reading frames within a 5kb region of the Adenovirus-2 genome.

SPREAD will display all occurrences of a defined sequence, plus their flanking sequences, present in an input sequence file (1). High resolution graphic representation of the output is available. The program will also now accept multiple sequence files and batch input, which is useful for long and/or multiple searches.

TABLES is a an enhanced version of our earlier program to aid in the determination of restriction enzyme recognition sequences (1). The program can create and study tables for a large variety of sequence types, including all sequence patterns known to be recognized by restriction enzymes. It now has the capability for on-line addition of new sites and a greatly improved algorithm for generating the master tables of sites.

UPDATE is used to write the pointers in the master data files used by M13 and SEQ (1).

ZFIND finds runs of alternating purines and pyrimidines in a sequence and hence indicates likely candidate regions for Z-DNA formation.

ACKNOWLEDGEMENTS

The authors thank R.M. Blumenthal, R.E. Gelinas, T.R. Gingeras, J.P. Milazzo, P.J. Rice and D. Sciaky, without whose help and encouragement the present versions of these programs would not exist. This work was supported by a grant from the National Cancer Institute (CA 27275).

REFERENCES

1. Blumenthal, R.M., Rice, P.J. and Roberts, R.J. (1982) Nucleic Acids Res. 10: 91-101.
2. Staden, R. (1977) Nucleic Acids Res. 4: 4037-4051.
3. Gingeras, T.R., Milazzo, J.P., Sciaky, D. and Roberts, R.J. (1979) Nucleic Acids Res. 7: 529-545.
4. Gingeras, T.R., Rice, P. and Roberts, R.J. (1982) Nucleic Acids Res. 10: 103-114.
5. Southern, E.M. (1979) Anal. Biochem. 100: 319-323.
6. Staden, R. (1980) Nucleic Acids Res. 8: 3673-3694.
7. Gingeras, T.R, Milazzo, J.P and Roberts, R.J. (1978) Nucleic Acids Res. 5: 4105-4127.