Interactive computer programs for the graphic analysis of nucleotide sequence data

Verne A. Luckow^{1,3,*}, Roland K. Littlewood² and Robert H. Rownd³

¹Laboratory of Molecular Biology and ²Biophysics Laboratory, University of Wisconsin, Madison, WI 53706, and ³Department of Molecular Biology, The Medical and Dental Schools, Northwestern University, Chicago, IL 60611, USA

Received 30 August 1983

ABSTRACT

A group of interactive computer programs have been developed which aid in the collection and graphical analysis of nucleotide and protein sequence data. The programs perform the following basic functions: a) enter, edit, list, and rearrange sequence data; b) permit automatic entry of nucleotide sequence data directly from an autoradiograph into the computer; c) search for restriction sites or other specified patterns and plot a linear or circular restriction map, or print their locations; d) plot base composition; e) analyze homology between sequences by plotting a two-dimensional graphic matrix; and f) aid in plotting predicted secondary structures of RNA molecules.

INTRODUCTION

The proliferation of nucleotide and protein sequence data in recent years has provoked widespread interest in developing computer programs to aid in the search for patterns, palindromes and other regularities in sequences of potential biological significance. Several overlapping collections of programs have appeared which can be divided into two general categories: small sets of highly interactive programs often used on microcomputers which are concerned primarily with local data entry and simple analysis, and large integrated sets of programs on mainframe computers which often serve to maintain genetic databases.

We have written a set of highly interactive programs for a high performance desktop computer that facilitate local sequence data entry, editing, and rearrangement. These programs emphasize graphic display of sequence information which greatly enhances their analytical power by revealing various properties or patterns encoded within nucleotide or protein sequence data.

PROGRAM DESCRIPTIONS

Entry, Editing, Listing, and Rearrangement (SEQ)

This program provides a variety of general functions for entering, editing, listing, or generating nucleotide or protein sequence files. Sequences

are checked after each line is entered from the keyboard to make sure they contain only legal characters. Legal character sets include DNA, RNA, extended DNA, extended RNA, amino acid, and ASCII (any character is legal). Extended character sets may contain characters which represent ambiguous nucleotides. Random sequences can be generated using any character set. Sequences can be edited using the functions substitute, delete, insert, search, and convert. Unlike other sequence editors with which we are familiar, the numbering system does not change until all of the insertions or deletions are completed. Thus, deletions or insertions can be specified while progressing through the file in any order. Searching for exact matches to short patterns is permitted during editing. The convert function determines the characters in a sequence and allows all occurrences of one character to be changed to another. Sequences can be listed in a variety of formats on either the printer, plotter, or display. An important feature is the convenient coordinate or relative numbering system achieved by setting a parameter (offset), to specify the starting position of the first character in a sequence. Sequences can be listed, therefore, with coordinates that start with any positive or negative number, not necessarily one (1). Sequences are stored and retrieved from files which may have one of two possible formats. Files stored in an ASCII format can be read by terminal emulator programs which allow transfer of files between remote computer systems. The format of ASCII files is compatible with sequence analysis programs that run on our mainframe computer. An encoded binary file format which is more efficient than ASCII format files, both in terms of disc storage space and disc access speed, is used by all of the sequence analysis programs described in this paper. Additional information stored with each sequence includes name, topology (linear or circular), type (character set), offset (for the coordinate numbering system), species, reference or comments, length, date, and checksum. The checksum, calculated according the algorithm we presented in an accompanying paper (1), provides a convenient method for verifying sequence integrity when data are transferred from one computer to another or when published data are manually entered into a computer. Sequences can be rearranged by flipping (reversing and complementing), by circular permutation (rotation), or by appending one sequence to another.

Digitized Entry of Sequences (DIG)

This program uses a digitizer to enter sequence data directly from autoradiograms and to store it in computer files which can be later used to compile an overall sequence. Among the functions provided by the program are digitizer setup, read, compare, list, edit, store, and load. Several programs with sim-

Figure 1. Representative output from the SEARCH program. Portion of a linear restriction map of the plasmid pBR322.

ilar functions have been previously described (2,3).

General Pattern Searching and Restriction Enzyme Mapping (SEARCH)

This program searches for patterns such as restriction enzyme recognition sites within nucleotide sequences. Patterns may contain wild card characters which match two or more characters within the search string. The program can be used to list the locations of the sites including the positions of the cuts in both the 5' and 3' strands, the phase, the sequence surrounding and including the match, and the distance from the previous match. This is followed by a table showing the length of the predicted fragments sorted in order of decreasing size. The program can also be used to generate linear or circular restriction maps of sequences which are drawn on a graphics plotter (Figure 1). Both options list the sites which do not have matches in the sequence. Patterns available for searching include the complete set of prototype restriction endonuclease sites, subsets of the above selected by enzyme name, the set of commercially available restriction enzymes, and new sites entered by the user.

Restriction Enzyme Database (ENZYMES)

This program creates the restriction enzyme pattern files and other data files required by the SEARCH program.

Base Composition Plotting (BPLOT)

This program plots the base composition of a DNA sequence on a graphics plotter (Figure 2). Plots of A + T content closely match the denaturation maps of DNA molecules (4). The composition can refer to any base or combination of bases and is plotted as a fraction or as a percentage. The DNA sequence is analyzed on a sliding average basis. The A + T content, for example, for each base-pair position is defined as the average percent of adenine and thymine for a given number of base-pairs (average segment length) around that position. Two other parameters (cut-off window and minimum acceptance region) can be set to accept or reject the base-pairs' sliding averages, depending on their respective A + T content and position. Options include plotting the results as a one-step bar graph and treating the sequence as a circle.

Base and Codon Content (CONTENT)

This program calculates and prints codon content and base frequency.

Graphic Matrix Homology Searching (DOTS)

This program is used to search for homologies and regularities within or between nucleotide or protein sequences by displaying a graphic matrix (Figure 3). This method, first introduced by Maizel (5) and Hunkapillar (6), compares consecutively each character of one sequence with all characters in a second sequence, and places a dot in the matrix for every position where the two char-

Figure 2. Plot of the A + T composition of pBR322. Thick line - averaging segment length is 200 bases. Thin line - averaging segment length is 50 bases.

Figure 3. Comparison of the RNA sequence corresponding to the incompatibility RNA of pBR322 (10) to its reverse complement.

acters are identical. The program contains a noise reduction or filtration procedure which is controlled by a parameter (matchlength) that determines the minimum number of characters required before a stretch of dots are placed in the matrix. A parameter may also be set which controls the minimum number of bases required for each match (minmatch). Another powerful procedure allows groups of characters in each sequence to be converted to other characters. This procedure can be used to visualize distant protein homology by converting individual amino acid characters into four broad classes based on chemical reactivity of amino acid side groups: non-polar, polar, basic, and acidic. Additional options include generating a sequence's reverse complement, allowing GU pairing to contribute to reverse complement matching, printing the homology regions, and directing the dot matrix output to either the graphics display and

Folding 322INC
Nucleotides 1 to 107
Energy = -47.5

Figure 4. Secondary structure of the incompatibility RNA of pBR322 (10) predicted by the program of Zuker and Stiegler (9) and drawn by the ZPLOTB program.

printer, or the plotter.

Search for Hairpins and Palindromes (HAIRPIN)

This program searches for potential hairpins and palindromes.

Search for Homologous Repeats (MATCH)

This program compares sequences for regions of perfect homology. Both HAIRPIN and MATCH are derivatives of earlier programs by Staden (7,8).

Display of RNA Secondary Structure (ZPLOTB)

ZPLOTB is an interactive plotting program used in conjunction with the programs of Zuker and Stiegler (9) to draw the predicted secondary structures of single-stranded RNA molecules. These programs calculate optimal computer folding of large RNA sequences using thermodynamic information as well as auxiliary information from chemical or enzymatic studies which reveal base pairing interactions and single-stranded regions.

ZPLOTB is a derivative of an earlier Fortran program (G. Osterburg, unpublished) which has been translated into HP BASIC and enhanced to provide interactive control over the drawing of the predicted structures and allow pivoting of subsections of the sequence. Options control the labeling and numbering of the bases as well as the drawing of tic marks at each position on the molecule. Tic marks have the advantage of distinguishing individual bases in congested drawings. Subsections of the sequence may be pivoted by specifying two positions on the sequence and a pivoting angle. Pivoting is limited to the struc-

ture between these two positions and occurs around the XY coordinates of the first pivot point. The second position specified need not be numerically higher than the first. Multiple pivots are allowed for each drawing.

The drawing of the molecule can be directed to a graphics screen or to a hard copy graphics plotter (Figure 4). We routinely direct plots to the graphics screen and adjust the pivots until a satisfactory structure is obtained before plotting the molecules on paper.

Inter-Computer Communications (TERM)

This is a terminal emulator program which allows exchange of sequence data files between our computer and remote computer systems. This permits collection of sequence information from genetic databases, generally maintained on large computer systems, as well as the transfer of sequence data in the reverse direction, for analysis by programs which offer capabilities not found in our local collection of programs.

IMPLEMENTATION

These programs are written in HPL (high performance language), available for Hewlett-Packard series 200 desktop computers. ZPLOTB is written in HP enhanced BASIC 1.0. These programs were developed on a HP9826A computer interfaced to a HP9872C plotter, a HP2671G thermal graphics printer, and a Numonics digitizer.

CORRESPONDENCE

Correspondence concerning these programs should be directed to V. Luckow.

ACKNOWLEDGMENTS

V.A.L. is a predoctoral trainee supported by Public Health Service training grant GM-07215 from the National Institutes of Health. This work was supported in part by U.S. Public Health Services research grants GM-14398 and GM-30731 to R.H.R.

*Present address: Department of Molecular Biology, The Medical and Dental Schools, Northwestern University, Chicago, IL 60611, USA

REFERENCES

1. Littlewood, R.K., Luckow, V.A., Rownd, R.H. Submitted to Nucleic Acids Res.
2. Gingeras, T.R., Roberts, R.J. (1982) Nucleic Acids Res. 10, 103-114.

3. Lautenberger, J.A. (1982) *Nucleic Acids Res.* 10, 27-30.
4. Funnell, B.E., Inman, R.B. (1979) *J. Mol. Biol.* 131, 331-340.
5. Maizel, J.V., Lenk, R.P. (1981) *Proc. Nat. Acad. Sci. USA* 78, 7665-7669.
6. Steinnitz, M., Frelinger, J.G., Fisher, D., Hunkapillar, T., Pereira, D., Weissman, S.M., Uehara, H., Nathenson, S., Hood, L. (1981) *Cell* 24, 125-134.
7. Staden, R. (1977) *Nucleic Acids Res.* 4, 4037-4051.
8. Staden, R. (1978) *Nucleic Acids Res.* 5, 1013-1015.
9. Zuker, M., Stiegler, P. (1981) *Nucleic Acids Res.* 9, 133-148.
10. Selzer, G., Som, T., Itoh, T., Tomizawa, J. (1983) *Cell* 32, 119-129.