Volume 12 Number 1 1984 Nucleic Acids Research

A computer program to enter DNA gel reading data into a computer

Rodger Staden

MRC Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, UK

Received 8 August 1983

ABSTRACT

This paper describes a simple program that uses a digitizing device to enter DNA sequences directly from autoradiographs into a computer.

INTRODUCTION

We describe a method to enter DNA sequence data directly from autoradiographs into a computer. Two previous papers(1,2) have described similar programs but we believe our method to be useful both because it is very simple and fast to use, and because it is an integral part of our comprehensive system for handling shotgun sequence data(3). This program (GELIN), which uses a digitizing device, is therefore a replacement for our earlier input program BATIN. The digitizing device or digitizer is generally a translucent two dimensional surface on which an autoradiograph can be mounted. If a special pen is pressed onto the surface of the autoradiograph the computer connected to the digitizer will record the x,y coordinates of its position. These coordinates can be interpreted by a program.

In order to read an autoradiograph the user need only define the four areas that contain the four sequencing lanes and the bases to which they correspond and then use the pen to point to each successive band as he moves up the gel; the program examines the coordinates of each pen position to see in which of the four areas it lies and assigns the corresponding base to be stored in the computer. Each time the pen tip is depressed to point to a position on the surface of the digitizer the program sounds the bell on the terminal to indicate to the user that a point has been recorded. As the sequence is read the program displays it on the screen both as rows of characters and by drawing a picture of the autoradiograph.

The program uses a menu to allow the user to select commands or to enter

the uncertainty codes(3) for areas of the gel that are difficult to interpret. A menu is simply a series of boxes drawn on the surface of the digitizer that each contain a command or uncertainty code. When the user puts the pen down in these special regions the program interprets the coordinates as commands and acts appropriately. As well as the uncertainty codes 1,2,3,4,B,D,H,V,R,Y,X,-,5,6,7,8 and A,C,G,T the following commands are included in the menu: DELETE removes the last character from the screen, the picture of the autoradiograph, and the sequence being accumulated in the computer; RESET is used to redefine a value; START means begin the next stage of the procedure; STOP means stop the current stage in the procedure; CONFIRM means confirm that the last command or set of coordinates are correct; FILEIN means read in from a disk file a short electrophoresis run of the sequence that is about to be read so the user can add the new data, butt-jointed, to its end (the program then asks for the name of the file containing the short run and reads it in and displays it on the screen. The user can identify the join between the two gels using the picture on the screen. If he sees that the short run has been read too far, in that it contains errors that can be resolved from the long run, he can use the DELETE command to remove the incorrect bases before adding the correct ones from the long run.); FILEOUT means that the user wants to store the current reading in a disk file. Note that the FILEIN option is employed much less often since we started to use gradient gels for DNA sequencing as sufficient data is now obtained from a single run(4).

USING THE PROGRAM

Once started, the program asks the user to type the name of a file to contain the names of all the gels he is about to read (this file of file names allows subsequent processing programs to handle the sequences as a batch); to identify the origin of the digitizer coordinates; and to affix his autoradiograph firmly to the digitizing surface. It then displays on the screen the default lane order T,C,A,G and asks the user to either confirm that it is correct or to define a different one (this is simply the order of the lanes from left to right and a new order can be assigned for each gel by using the characters in the menu). The next thing to be done is to define the areas of each of the four lanes.

As the lanes are only about 3mm wide and rarely run straight, generally it is not sufficiently precise to simply define four rectangles, and so the following method is used. The user defines the positions of the centres of

the lanes level with the bottom band to be read; the program responds with the coordinates of these and the user should confirm these and then start to read the sequence. When the program is interpreting the coordinates to decide the character to assign it simply looks to see which of the four current lane centres is closest to the pen position and assigns the corresponding base to the sequence. It then resets the centre of the lane for that base: in this way the program is constantly updating the position of the lane centres in accordance with the users pen positions. As a safeguard the following checks are made: when the original lane centres are supplied by the user the program measures their separation; if at any subsequent time the lane centres approach one another to closer than 30% of this distance the program will not respond to the pen being depressed (this is also the case if the pen is depressed at a position that is too far removed from the last point). If this occurs the user can use the RESET function to redefine the current lane centre positions (this can be done at any stage). The user will know that the program is not responding to the pen because the bell will not sound. It should be noted that it is only in extreme cases of curvature combined with a base composition such that one lane is rarely used that the RESET facility will be required: almost all of our films can be read without recourse to this function.

When the lane centres have been defined the user starts to read the gel by placing the pen on each successive band, moving up the gel as he would in the normal way. It is the user who interprets and defines the sequence by the order in which he hits the four areas: the program does not decide the sequence by calculating the order of the y coordinates and assigning bases on the corresponding x coordinate. The user is interpretting the film and can therefore use his experience to take acount of the known artefacts, compressions and pile-ups(5). At any time the user may assign an uncertainty code by using the menu or he can stop reading the current sequence by hitting the STOP command.

The program responds to the STOP command by asking the user to hit FILEOUT OR STOP. If the user wants to store the last gel reading in a disk file he should hit FILEOUT in the menu and the program will prompt for a filename from the keyboard. The program will then ask if he wishes to enter another gel reading and if he does the program goes back to the step of affixing the autoradiograph and defining of lane order; if he does not the program stops.

DISCUSSION

We have described a simple program that forms part of a comprehensive system for handling shotgun sequencing projects. This system of programs can take a batch of gel readings from this program; screen-out unwanted vector sequences; screen-out fragments that contain restriction enzyme recognition sites that should not be present (for example if sonication(6) is used to prepare fragments for cloning the DNA will first be circularised at a restriction enzyme site, and any fragments containing this site will be those that cross the join. These should be cut using an editor before entry into the sequence database); and finally compare and align all the remaining fragments with a consensus for all the previously collected data. All of this is done automatically with no user intervention but other, interactive programs allow the user to display all of the sequences aligned one above the other with their consensus below and also to edit the sequences. To date, in this laboratory alone, these programs have been used to handle sequences totalling more than one million bases.

We have found that the program improves accuracy by removing transcription errors, by reducing user fatigue, and because the program never forgets which lane the pen is in: a common source of error with pencil and paper reading is that although the user correctly interprets the order of the bands he will confuse the lanes and write down the wrong bases. The ability of the shotgun data handling programs, outlined above, to display all the gel readings that contribute to each section of sequence aligned one above the other, makes it unnecessary for the gel reading program to contain the facility for multiple reading of sequences. Any errors are made very clear by the alignment, and infact, to save time, can be left until it becomes obvious which gel reading contains the error. It will become obvious because, in the alignment, one sequence will be different from all the others and this means that the user probably need only check the one offending gel, and need only check it once. He does not have to go through the time-consuming process of checking every missmatch every time new data is compared to his consensus and the display is such that he knows exactly which film to look at.

HARDWARE

Although we described the digitizer as a two dimensional translucent surface we actually use a digitizer that does not require a special surface for its operation(7). This device uses two microphones, positioned at the

top corners of the active area, to pick up a small sound made at the pen tip when it is pressed down: the coordinates are then calculated by triangulation. We find the device very convenient to use when mounted on a conventional light box as there is nothing to obstruct or obscure the films we are examining.

The program is written in FORTRAN 77 for a VAX 11/780 computer and uses a simple graphics terminal(8) to handle the drawing of the gel patterns and display of characters. It has also been implemented on a TORCH microcomputer which has a CPM compatible operating system(9).

ACKNOWLEDGEMENTS

I would like to thank K. A. Nasmyth and T. J. Gibson for helpful suggestions and F. Sanger, K. McKenney and T. H. Rabbitts for critical reading of this manuscript.

REFERENCES

(1) Gingeras, T. R., Rice, P. and Roberts, R. J. (1982), Nucl. Acid Res. 10, 103-114.
(2) Lautenberger, J. A. (1982) Nucl. Acid Res. 10, 27-30.
(3) Staden, R. (1982) Nucl. Acid Res. 10, 4731-4751.
(4) Biggin, M. D., Gibson, T. J. and Hong, G. F. (1983) Proc. Natl. Acad. Sci 80, 3963-3968.
(5) Bankier, A. T. and Barrell, B. G. (1983), Techniques In Life Sciences, in press.
(6) Deininger, P. L. (1983) Anal. Biochem. 129, 216-223.
(7) The digitizer is a Graphbar model GP-7, made by Science Accessories Corp., 970 Kings Highway West, Southport, Connecticut 06490, USA.
(8) The terminal we use is a VT640 which is an upgrade of the VT100 terminal. The VT100 is a standard DEC Vax terminal and the upgrade is made by Digital Engineering Inc., 630 Bercut Drive, Sacramento, Calif. 95814, USA.
(9) The CPM compatible system is a TORCH Z80 DISC PACK using a BBC Model B as a peripheral processor. The TORCH is made by Torch Computers, Abberley House, Great Shelford, Cambridge CB2 5LQ, U. K.