Volume 10 Number 1 1982 Nucleic Acids Research

A semi-automated method for the reading of nucleic acid sequencing gels

Thomas R.Gingeras, P.Rice and R.J.Roberts

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

Received 1 October 1981

ABSTRACT

A collection of computer programs is described which permit automatic entering of nucleotide sequence data directly from an autoradiograph into a computer. This collection, called DIGITPAD, makes use of a digitizing tablet for the data entry and allows the rapid and accurate transfer of the sequence into the computer.

I. Introduction

Any computerized processing and analysis of nucleic acid sequence data requires that the primary sequence data be faithfully recorded at the outset. The development of very thin polyacrylamide urea containing gels (1) has permitted the resolution of products, from a single sequencing reaction, up to a chain length of 250 to 300 nucleotides per loading. Two sources of error are associated with the manual transfer of sequence information from an autoradiograph into a computer. The first involves carelessness in reading the gel (for example, mistaking channels, or skipping nucleotides). The second occurs as a result of typographical errors while copying the sequence into the computer.

We have developed an approach to overcome such errors by automatically transferring data directly from the original autoradiograph into a computer. Our approach utilizes an electronic digitizing tablet that is controlled by a set of programs called DIGITPAD.

II. The Principles Behind the DIGITPAD Programs

A digitizing tablet (Fig. 1) operates by sending to the computer the location of any point on its surface once this point has been touched by a signal pen. This location is represented as a set of digitized X and Y coordinates. The DIGITPAD programs contain two modes, one used for initialization, the other for data collection. During initialization, areas on the digitizing tablet are defined to be associated with particular functions. During data

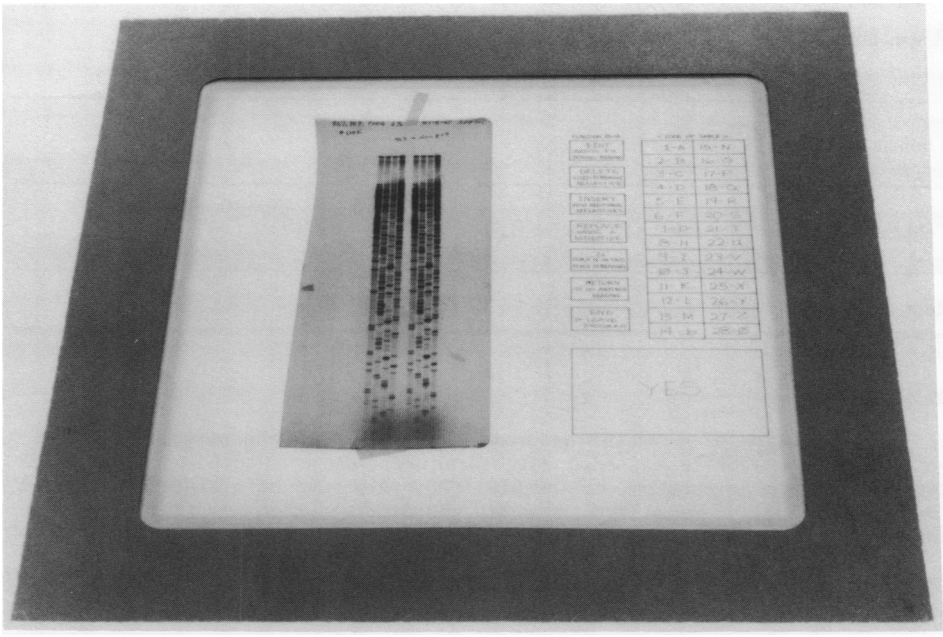

Figure 1: The digitizing tablet used to read sequencing gels. Pictured on the surface of the tablet is an autoradiograph and a menu which activates various functions encoded in the DIGITPAD programs (see text). The limit of resolution for this tablet is 0.1 mm. Reading of the autoradiograph or activation of one of the functions from the menu is done by means of a signal pen.

collection, positional information is translated into nucleotide representation and then processed into a sequence.

This is done by placing an autoradiograph on the surface of the pad and identifying each of the nucleotide channels by touching the four corners of each channel with a signal pen. The result is that each channel is defined by a quadrilateral such that any location within this quadrilateral, subsequently touched by the pen, is automatically assigned the appropriate base. Consequently, an autoradiograph is read by touching each band with a signal pen and the location is recorded as the corresponding base.

III. Operation of DIGITPAD

A. INITIAL SETUP

The program SETBOX is used to define the boundaries for each of the boxes represented in the menu (Fig. 1). This program is only run during the establishment of the menu area of the digitizing tablet. The boundaries (x, y coordinates for each of the four corners of each box) are kept permanently (in

a file called BOX.DAT) for reference by the program READ. Figure 2 provides an example of the dialogue between the user and the SETBOX program. Answers to the SETBOX questions are given from a terminal keyboard. Once SETBOX has been run, use of the keyboard is minimal during the running of READ and most input is directly from the digitizing tablet.

B. DATA COLLECTION

The program READ is concerned with the input and manipulation of sequence data from an autoradiograph into the computer. READ has four functional parts: 1) Introduction, 2) Initialization, 3) Readings and 4) Editing.

1. Introduction: The introductory part of READ provides instruction for a novice user and gives an explanation of the function boxes listed in the menu. If the user is already familiar with the program, this section can be bypassed. Figure 3 illustrates the interaction between the user and the

```
First pick a reference point on the tablet.  After this a hit to the left of
this point will end the program.

Put in YES              Function Box

Each box is input in following order:

1.   Upper left hand corner
2.   Upper right hand corner
3.   Lower right hand corner
4.   Lower left hand corner


(343.8, 171.0)                                  (502.8, 171.0)

          ................................................
          .                                              .
          .                      YES                     .
          .                                              .
          .                                              .
          ................................................

(343.8, 69.0)                                   (502.8, 69.0)


Is this all right?  (Y-N)
.
.
.
Put in EDIT             Function Box
```

Figure 2: This is an example of the dialogue that a user encounters during the activation of the SETBOX program. This program establishes the boundaries for each of the boxes represented in the menu (see Figure 1). It is only run, after the initial magnetization (or re-magnetization) of the digitizing tablet.

```
WELCOME TO THE WONDERFUL WORLD OF THE DATA PAD

DIGITPAD is a program to input data from sequencing autoradiographs.  Most input
comes from the digitizer except the headings and the selection of the file.
This input comes from the keyboard.

     When using either a Function Box or a box in the LOOKUP table, put pen as
close to center as possible.

     Now follows a brief explanation of the Function Boxes and LOOKUP table.

     Put in the indicated number (from keyboard) and you will see the explana-
tion until you hit return:

1)   YES Function Box
2)   EDIT Function Box
3)   DELETE Function Box
4)   INSERT Function Box
5)   REPLACE Function Box
6)   "N" Function Box
7)   RETURN Function Box
8)   END Function Box
9)   LOOKUP Table
10)  All Of The Above

     If you hit return the program will continue with no further explanation.

10



YES BOX         Largest box on pad - appears in lower right hand corner and is
                used for answering (Y-N) questions--A hit in the box sends YES.
                A hit out of the box sends NO.
EDIT BOX        First Function Box
                used to obtain EDIT function during reading.  This is the only
                way, during reading, to get to the DELETE, INSERT, or REPLACE
                functions.
DELETE BOX      Second Function Box
                used to remove a nucleotide.  Available only during MATCH and
                EDIT subroutines.
INSERT BOX      Third Function Box down
                used to insert additional nucleotides during MATCH.  May INSERT
                only two during  EDIT.  May INSERT up to three nucleotides during
                MATCH.
REPLACE BOX     Fourth Function Box down
                used to change a nucleotide during MATCH or EDIT.
"N" BOX         Fifth Function Box down
                used to record N as a nucleotide during reading and MATCH.
RETURN BOX      Sixth Function Box down
                used to indicate end of a single reading.
END BOX         Seventh Function Box down
                used to terminate the program.

LOOKUP TABLE    Table to right of Function Boxes
                used to input number (between 1 and 28), or letters. The program
                will determine whether a number or letter is required. Try to hit
                middle of box to cut down on error. Hit return when finished.
```

Figure 3: An example of the Introductory part of DIGITPAD. At the start of the DIGITPAD program, a brief set of explanations are optionally available, which describe the Function Boxes and LOOKUP Table listed on the menu.

computer during this introduction and provides a brief description of each of the function boxes shown in the menu (Fig. 1).

2. Initialization: This portion of READ is used whenever the sequence from a new autoradiograph is ready to be entered into the computer. The program requests a letter (from A-Z) from the user (from keyboard) which is used to name an output file (standardized as TRIALA, TRIALB, etc.) and a summary file (standardized as GARBAGE.A, GARBAGE.B, etc.) (Fig. 4). All the data from reading any set of autoradiographs will be stored in these two files. Figure 5a; b shows an example of the contents of each of these files. The information passed to the TRIAL file (Fig. 5a) is formatted so that it can be recognized by the set of programs called ASSEMBLER which we described earlier (2).

The headings for each set of readings are then requested by the computer (Fig. 4). The headings are input from the keyboard, and each new reading will be preceded by this request if a new heading is desired. A NO is a hit on the surface of the tablet anywhere except within the confines of the YES box.

The next questions from this initialization section concern the autoradiograph. The program asks how many channels have been used in the sequencing reaction (4 or 8 channel format) (Fig. 4). The program looks for a numeric response derived from the tablet. Any number other than 4 or 8 produces an error message. If a 4 channel format is chosen, the program will ask for the position of the 5 top-most corners, followed by the 5 bottom-most corners of the channels. This is used to define a quadrilateral for each of the four reaction channels.

To determine the order of the channels, the program asks for the nucleotide in the left-most channel. Table 1 shows the default order of the reactions based upon the nucleotide represented in the left-most channel. These defaults allow for reading autoradiographs resulting from either the Sanger dideoxy-chain termination method (3), or from the Maxam-Gilbert chemical modification method (4).

The computer responds to the user choice by providing a graphic representation of the reaction channel order (Fig. 4). The user is now asked to accept or reject this order by using the YES function box in the menu. If accepted, the initialization step of the program is completed.

3. Reading: When the computer is ready to accept data from the autoradiograph, the user is asked to: KEY IN DATA POINT OR FUNCTION. Data is entered by simply touching the band on the autoradiograph that corresponds to a sequential reading of the gel from bottom to top. If a reading is to be

```
.
.
.
.
Input through keyboard a letter A-Z
Into what file are you entering data?

A


Input through keyboard (Y-N)
A 'Y' will choose voicebox option
A 'N' will choose NO voicebox option

N


What heading would you like to use?
Input from keyboard

Test-1


Storing into TrialA

Unless new heading, all further input will come from digitizer

Do you want a new heading (Y-N)?
How many channels do you want? (4-8)
<Use LOOKUP TABLE>

4

Put in top 5 points <from left to right>

Put in bottom 5 points <from left to right>
What is the left-most channel?
     <use LOOKUP TABLE>
                              CHANNEL DISPLAY

(187.0, 388.0)               (202.0, 388.0)          (216.0, 389.0)
**************************************************************************
*              (195.0, 388.0)          *          (209.0, 389.0)      *
*                   *                  *               *              *
*         T         *       G          *      C        *       A      *
*                   *                  *               *              *
*              (199.0, 145.0)          *          (213.0, 146.0)      *
**************************************************************************
(191.0, 145.0)               (206.0, 146.0)          (221.0, 146.0)

Do you find this acceptable?
Key in coordinates of data point or function
.
.
.
.
```

Figure 4: In preparation for reading an autoradiograph each film is initialized (or identified). This is done by assigning the data to be read from the gel to a specific data file (e.g. TRIALA). Also, in the case of multiple reactions per autoradiograph, each reaction is assigned an individual heading by the user (e.g. Test-1). This initialization of a film also requires the user defining the areas of the tablet which demark each of the four nucleotide channels (A graphic representation is provided complete with X, Y coordinates for the four points that define each channel.) Once this step is complete, the program (READ) is ready to accept data.

```
TRIAL FILE OUTPUT:

Test-1                                                    42
CCCCGGATCATGCCNATTCGGNCGAACTNTTNGCCACANGGN
Test-2                                                    30
CCATTCGCGGCGACACTTTGCCACACGGGC
```

```
GARBAGE FILE OUTPUT:

Test-1 number 1                    41
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
CTTTG C C A C A C GGG
Test-1 number 2                    38
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T
TTT G C C A C A G GG
```

Figure 5a (top): This is an example of how data is formatted in the main output file (TRIALA) generated by DIGITPAD. Data in this format is acceptable by other computer programs designed to assemble nucleotide sequences (see ref. 2).

5b (bottom) A record of each reading is stored in a file called GARBAGE__. This allows one to have a record of how a consensus nucleotide was reached for each position in the sequence.

changed immediately after entering it, the signal pen is placed on the tablet surface away from the function boxes or designated channels and will erase the user's most recently entered nucleotide. After an initial reading of the autoradiograph, the user may choose to a) edit the input sequence [activate EDIT], b) check the input sequence against a second reading of the same gel

TABLE 1

Designation of Nucleotide Order for 4 and 8 Channel Reactions.

User Designated Left-Most Channel	Default Channel ORDER
4 Channel Rx	
A	ACGT
G	GATC
T	TGCA
C	not used
8 Channel Rx	
None needed	CTACGTAG

[activate RETURN], or c) end the reading session and exit the program [activate END].

4. Editing: This mode is activated by touching the area designated as EDIT. The sequence which has already been read into the computer will be displayed in rows 25 nucleotides long (Fig. 6). Any nucleotide can be edited by supplying its row and column number. These numbers are supplied to the program by using the alphanumeric table listed in the menu (see Fig. 1). After you have made an acceptable selection the user is given the opportunity to DELETE, INSERT, REPLACE, or insert an N into any of the positions of the original nucleotide sequence. This is done by touching the appropriate box in the menu followed, where necessary, with the replacement nucleotide. The corrected sequence is immediately displayed after editing (Fig. 6).

```
REQUESTED EDIT FUNCTION

      / 1/ 2/ 3/ 4/ 5/ 6/ 7/ 8/ 9/10/11/12/13/14/15/16/17/18/19/20/21/22/23/24/25/
- - -/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/--/
  1 / C/ C/ A/ T/ T/ C/ G /C/ G/ G/ C/ G/ A/ C/ A/ C/ T/ T/ T/ G/ C/ C/ A/ C/ A/
  2/ C/ G/ G/ G/ C/

This chart displays the given array.  The numbers at the left indicate the row.
The numbers at the top indicate the column.

Using pad input the number of the row.
2
Now using pad input the number of the column.  The number will indicate where you
have choosen
4
C G G 1 C

To correct discrepancy at left most 1, choose one of the following functions:

     DELETE
     INSERT
     REPLACE
       "N"

Requested REPLACE Function
Put in the desired nucleotide <use LOOKUP Table>

CCATTCGCGGCGACACTTTGCCACACGGCC

Key in coordinates of data point or function

Requested END Function

CCATTCGCGGCGACACTTTGCCACACGGCC

The END  there are 2 different gels
read to TRIALA
```

Figure 6: This is an example of how editing may be performed directly on an input sequence using the EDIT function.

a) Matching Mode: This mode is activated when two or more readings of the same gel have been performed after selecting the RETURN box in the menu. This is done to check the accuracy of the newly read nucleotide sequence, while the autoradiograph is still in place.

Figure 7 illustrates the results when two readings of an identical autoradiograph show several mismatches. Similarly to the EDIT function once the positions of each discrepancy have been located (given in Fig. 7 as a numerical value) the user is free to use the available functions listed in the menu in order to replace, substitute or leave ambiguous each of the positions of mismatch. In addition, each of the multiple readings from a single reaction is recorded in the GARBAGE. file so that a record of each reading is available (Fig. 4b).

b) END: By touching the function box labeled, END, the user can terminate the program at any time. The total of the number of reactions entered is then displayed. The sequence data for each set of readings is stored as a nonredundant version (i.e. no multiple readings, only the consensus of multiple readings) in a TRIAL. file. It is from this output file that data may be drawn for the further analysis.

IV. Interesting Features

The use of DIGITPAD for the automatic entry of nucleotide sequence data has several distinct advantages.

One interesting consequence of being able to read autoradiographs by means of a digitizing tablet is that a digitized map of the bands on the gel are recorded. This map is an accumulation of X and Y coordinates for each band on the autoradiograph. Consequently, any irregularity in the spacing of the bands due to localized compression in the gel, are a part of this record. Since such compressions are frequently an indication of localized secondary structure formation, a catalogue of regions of the DNA that tend to form such structures can be retrieved from the DIGITPAD raw data files (i.e. the listing of the untranslated x and y coordinates). This mapping data is potentially quite interesting since little is known about the rules that govern the formation of secondary structures in DNA molecules.

The main reason for the development of this method of data entry reading was to minimize the various sources of error associated with the manual recording of sequence data. However, one potential source of error associated with this new method was soon apparent. During the process of entering data from the tablet, it is very easy to lose one's place on the autoradiograph after viewing the terminal screen to check the accuracy of the newly entered data.

```
Requested RETURN function

CCCCGGATCATGCATTCGGGCGACACTTTGCCACAGGG

In the display trial 1 will appear on the top line and trial 2 will appear on
the bottom line.  The middle line will be the composite array with numbers
symbolizing discrepancies.
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GC1 A T T CGG2CG3 4 5 A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 1, choose one of the following functions:
     DELETE, INSERT, REPLACE or "N"
Requested INSERT function.  You may INSERT up to 3 nucleotides.
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCC A T T CGG2CG3 4 5 A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 2, choose one of the following functions:
     DELETE, INSERT, REPLACE or "N"
Requested "N" function.  N is placed in this position:
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCC A T T CGGNCG3 4 5 A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 3, choose one of the following functions:
     DELETE, INSERT, REPLACE or "N"
Requested DELETE function.  This will erase that nucleotide.
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCC A T T CGGNCGG 4 5 A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 4, choose one of the following functions:
     DELETE, INSERT, REPLACE or "N"
Requested REPLACE function.  N is placed in this position:
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCC A T T CGGNCGG A 5 A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 5, choose one of the following functions:
     DELETE, INSERT, REPLACE or "N"
Requested DELETE function.  This will erase that nucleotide:
C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCC A T T CGGNCGG A   A C 6
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T

To correct discrepancy at left most 6, choose one of the following functions:
     DELETE, INSERT, REPLACE, or "N"
Requested REPLACE function.  N is placed in this position.  The changes have
been made on this group of 25.

They are as follows:

Trial 1 appears on the top line; Trial 2 appears on the bottom line; The middle
line contains the corrected reading.

C C C C G G A T C A T GCC A T T CGG CGG C G A C A
C C C C G G A T C A T GCCNA T T CGGNCG  A   A C T
C C C C G G A T C A T GC  A T T CGGGCG  A C A C T
```

Figure 7: Multiple readings of a single reaction from an autoradiograph increases the reliability of the newly read data. Discrepancies between two reading can be detected by activating the MATCH mode of DIGITPAD. This is an illustration of the matching of 2 readings from the same gel. There are six positions of discrepancy noted. The correction of each position is illustrated. All of these corrections are performed using the Function Boxes and LOOKUP Tables listed in the menu. The composite sequence (listed between the first and third lines of sequence) is stored in the output (TRIAL) file. Individual readings (top and bottom strings of nucleotides) are stored in the GARBAGE file.

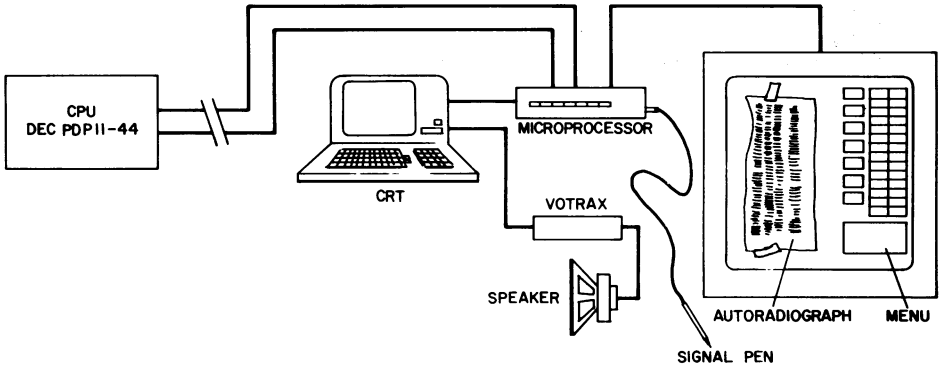

Figure 8: This is a wiring diagram indicating the relationship of the various parts of the gel reading station. The presence of the voice synthesizer (VOTRAX®) is optional.

A solution to this problem has recently been found by using an auditory feedback system to monitor the identity of each nucleotide that was being recorded. This was achieved by attaching a Votrax® "Type and Talk" speech synthesizer to the digitizing tablet setup. The program that drives the voice synthesizer can be optionally activated by the user and so is not essential to the setup. It is, however, an extremely useful addition.

V. Hardware and Software Specifications

Figure 8 includes a picture and a wiring-diagram showing the setup of the equipment which has been described in this paper. The digitizing tablet and companion microprocessor is an ID-TT-20 Translucent digitizing tablet (Summagraphics, Fairfield, CT). It utilizes an RS232 serial communications interface (Summagraphics) which connects the tablet to a PDP 11/44 mini-computer (Digital Equipment Corp). The CRT terminal is linked in series with the digitizing tablet, and the voice synthesizer (Votrax) is connected through the printer part of the CRT terminal.

DIGITPAD and its companion documentation are available upon request.

ACKNOWLEDGMENTS

A special thanks is due to J. Milazzo for his help in the initial stages of development of this program. Also, we wish to thank M. Ockler and N. D'Anna for their help in the preparation of this manuscript. This work was supported by a grant from the National Cancer Institute (CA 27275).

REFERENCES
1. Sanger, F. and Coulson, A.R. (1978) FEBS Letters 87: 107.
2. Gingeras, T.R. and Roberts, R.J. (1980) Science 209: 1322-1328.
3. Sanger, F., Nicklen, S., Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74: 5463.
4. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74: 560.