Volume 14 Number 1 1986 Nucleic Acids Research

Automatic reading of DNA sequencing gel autoradiographs using a large format digital scanner

J.K.Elder, D.K.Green and E.M.Southern[1]

MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, and [1]MRC Mammalian Genome Unit, Department of Zoology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Received 17 July 1985

ABSTRACT

We have developed a large format digital scanner for several applications in nucleic acid analysis. Here we describe the scanning of autoradiographs of DNA sequencing gels and a set of programs for reading the base sequence. The programs correct distortions in the gel, recognize bands by their characteristic shape and assign bases to bands by weighting band position and intensity. The sequence read in this way is as accurate as that read by an expert.

INTRODUCTION

As techniques for sequencing DNA have improved, the reading of autoradiographs has increasingly become a limiting step in the process. The main improvement in reading has been the use of a digitizing tablet to enter band position directly into the computer [1, 2]. This eliminates transcription errors but retains the human skills of recognizing and interpreting the complex patterns formed by the bands. But even this method is not free of errors, and becomes tedious after several hundred bands have been entered. There is a need for full automation of the sequencing process. Scanners using charge-coupled device (CCD) array cameras have been used to digitize electrophoresis gels [3, 4]. We have designed and built such a scanner and describe here its application in scanning autoradiographs of sequencing gels, and programs for reading a base sequence from the digitized data. The performance of the system is comparable with that of an experienced person.

MATERIALS AND METHODS

Large Format Scanner

The stage (Apollo Optical Sciences Ltd) consists of a 32 x 45 cm glass plate mounted in a linear bearing carriage and driven in 5 μm steps by a step-

All software described here is written in the C programming language and is freely available on direct request for non-commercial use and on receipt of a magnetic tape.

ping motor at speeds of up to 2000 steps/s. It is illuminated by a DC-powered white fluorescent lamp across the full width of the plate. The illuminated strip is viewed through a mirror set at 45 degrees, by a 2048 element Fairchild 1500R CCD linear array camera fitted with an Olympus 50 mm f3.5 macro lens. A horizontal sliding camera mount allows the pixel size on the plate to be varied from 25 to 160 μm, with corresponding maximum field widths of 5 to 32 cm. The pixel integration time can be varied from 2 to 128 ms; we normally use an integration time of 10 ms.

Each pixel is converted to an 8-bit optical density (OD) value by a log amplifier and an analogue-to-digital converter. The full OD range is 0-2, but this can be reduced to give finer OD resolution over a smaller OD range. A lineskip facility selects one in every n scan lines, where n can be varied from 1 to 256. The lineskip controls the rate of data capture so that it can be matched to the rate at which data is written to disc. The use of lineskip does not imply that areas of the field between selected scan lines will not be represented in the final scan data; the stage speed is set independently, and can be chosen so that pixel edges of successive selected scan lines touch, or even overlap. A 2048-bit pixel mask passes selected pixels in each scan line. This facility is useful for restricting the scan data to regions of interest, for example the tracks in one-dimensional gels.

There are two scanning modes: in pixel mode, each pixel is passed as an 8-bit value; in integration mode, for each contiguous set of pixel mask bits, the pixels are summed and passed as a 16-bit value. Pixel mode is normally used for two-dimensional data; integration mode is useful for one-dimensional gels in which the tracks and bands are straight, for example, agarose gels of restriction fragments. As described below, pixel mode is necessary for scanning sequencing gels.

The scanner is controlled by a Mizar VME 7100 microcomputer with a Motorola 68000 processor, 512 kbyte memory, 20 Mbyte hard disc, a floppy disc drive, a 512 x 512 x 8 bit VME 512 display card (Computer Recognition Systems) and a monochrome Digivision monitor. All software is written in the C programming language and runs under the OS-9/68000 operating system (Microware Systems Corporation).

Scanning and Analysis of DNA Sequencing Gel Autoradiographs

Each set of four tracks is scanned simultaneously. The scan dimensions are typically 500 x 6500 pixels, corresponding to an image size of 2.5 x 32.5 cm using a pixel size of 50 μm. The data within each track of an autoradiograph is in theory one-dimensional, but because of band and track

curvature, it is necessary in practice to perform a full two-dimensional scan. Scans are therefore done in pixel mode, but because high spatial resolution is only required along the length of the gel, the mean of each set of four pixels in a scan line is computed and stored, giving an effective pixel size of 200 x 50 μm. The total time for scanning four tracks is less than 5 min.

At present, scan data is transferred to a VAX-11/750 computer for analysis. In the future, all data analysis will be performed on the Mizar computer.

Track Straightening The stored image of four tracks of an autoradiograph consists of 125 x 6500 8-bit OD values. The first stage of the analysis is to define track boundaries. The gel is divided into sectors 25 mm in length. For each sector the OD values in each of the 125 pixel columns are summed and integrated OD values searched for an interleaved sequence of five minima and four maxima constrained by equispacing: the minima define the boundaries and the maxima the centres of the tracks. This gives an estimate of the mean track width. For each sector the search is repeated, with the equispacing constraint relaxed about the mean track width. This relaxation is necessary to take account of local variation in track width. Local cubic interpolation is applied to the final template positions to obtain smooth track boundaries. Each track is then reduced to the global minimum track width by discarding pixels near boundaries. The four tracks are straightened simultaneously by transferring the data into an imaginary scan line as it is moved down the tracks in a direction perpendicular to the mean of the track boundaries.

Construction of Band Templates The shape of bands is not constant from track to track nor indeed within a track. However, band shape provides a valuable means of discriminating between images which are bands and those which are not. To make use of this information, we analyze each track in the same sectors as those used for track straightening and within each sector define a template characteristic of the sector. The template consists of a set of contiguous pixels which, when placed on a band, will run along its centre. Band templates are constructed by choosing within each sector a left-to-right path of varied slope in a manner to approximately maximize the inter-column pixel-value correlation. For a sector of scan lines m_0 to m_1 and pixel columns n_0 to n_1, we start from the leftmost column of pixels and examine columns to the right in three directions, using the following procedure. For each column j, define order k band directions d_j^k by

$$d_j^0 = 0$$

$$d_j^k = d_j^{k-1} + t \qquad (k = 1, \ldots, n),$$

where, for each k, t is chosen to minimize

$$\sum_{i=m_0}^{m_1} |z[i][j] - z[i + d_j^{k-1} + t][j + k]| \qquad (t = -1, 0, 1),$$

where z[i][j] is the pixel value in line i and column j. That is, we start by finding the offset which minimizes overall adjacent column differences, then the offset which, when combined with the first, minimizes overall differences between columns two units apart, and so on, as far as four-separated columns, thus allowing for finer gradients than if only adjacent columns were considered. We now consider the sector as partitioned into four sets of four-separated columns and, fixing the position of the first set, align the others by least squares to form the band template.

Track Profiles Since the template corresponds in shape to the principal features in the sector, the result of summing OD values across the template as it moves through the sector is a profile of those features in the sector which most closely correspond in shape to the strongest bands, rather than a simple one-dimensional OD profile. The mean gradients of the templates in each pair of adjacent tracks are used to register the track profiles with respect to each other by local linear displacement. In applying the template, only the central two-thirds of the track is used, in order to avoid the merging of multiple bands which occurs near track boundaries.

Band Location In regions of the tracks where bands are well separated, there is no difficulty in locating their position, but where bands overlap, it is necessary to use slope and curvature information, to detect and locate them. To do this, the first and second derivatives of each profile are estimated using quadratic convoluting functions [5].

Sequencer Inter-band distances increase with distance from the top of the gel. These distances are important in identifying bases from band positions. The characteristic inter-band distances are calculated from the spacing of the major bands in each track, and a quadratic function is fitted by least squares to give inter-band distance as a function of band position. Then, starting from the lower end of the gel, successive bases are identified by examining a window ahead of the current position between 0.3 and 1.75 of the unit band separation. All bands within the window are assigned weights

according to their intensity and their closeness to the unit distance. The weighting function used is

$$w(h, x) = (0.5 + 0.67x)h \qquad (0.3 \leqslant x \leqslant 0.75)$$

$$w(h, x) = (1.75 - x)h \qquad (0.75 \leqslant x \leqslant 1.75)$$

where x is the band distance from the current position expressed as a proportion of the unit band separation and h is the band height. The band with maximum weight is chosen as the next base in the sequence. Other bands with high weights are recorded as possible alternatives. Empty windows signal the possible absence of a base.

RESULTS AND DISCUSSION

Sequencing gels differ from the ideal in four respects: the tracks are not straight, the bands within the tracks are not straight, the inter-band spacing is not constant and band intensity is not uniform. In reading a sequencing gel by eye, these problems are overcome by the pattern recognition abilities of the brain, and by applying rules which are learnt from experience. Most of the pattern information necessary for this analysis is lost if the tracks are simply scanned with a one-dimensional densitometer. The large format scanner used in this work enables us to produce full two-dimensional scans to which image analysis procedures can then be applied. We have solved the problem of reading sequences automatically from the gel by first converting the image into an "ideal" gel with straight tracks, and then using band templates to locate the positions of bands. The sequences read from these band positions using simple rules, are as accurate as those read by an experienced person.

The image of the gel is represented in the computer as a 125 x 6500 array of OD values. Figure 2a shows the digitized image of a section of the autoradiograph in Figure 1. The first step in the analysis of the image is to define the track boundaries as described in Materials and Methods, and to straighten them without altering the relative positions of bands in different tracks (compare Figures 2a and 2b). Distortions in the gel are not uniform (see Figure 1) and so correction procedures such as the straightening of track boundaries and identification of bands are applied over limited regions of the gel.

There are two difficulties in identifying bands: some are faint and some

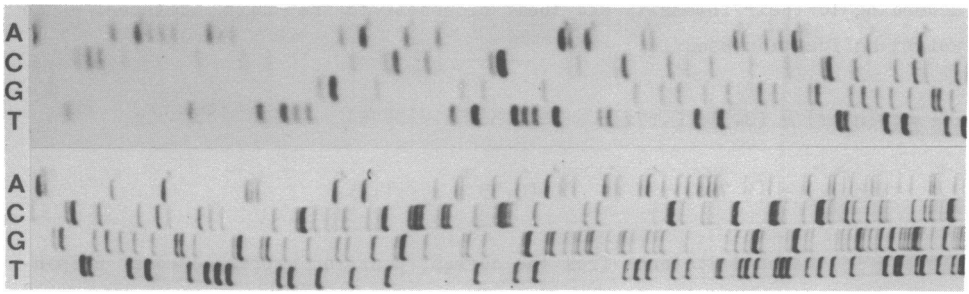

Figure 1. Four tracks from a representative sequencing gel, with an overlap between the two halves in the illustration. The gel, which was 40 x 35 x 0.03 cm, was prepared using 1 x TBE in the light solution and 2.5 x TBE in the heavy solution as described [6]. dATP[α-^{35}S] was used in labelling reactions.

overlap. Also, because bands are curved and are not perpendicular to track boundaries, their relative positions are sometimes difficult to define. To identify bands and to correct for distortion, we find the shape of major bands within a local region of a track and use this as a template to produce a representative OD profile of the track in that region. The slopes of band templates are used to register the tracks with respect to each other so that bands in different tracks are in the correct order when the tracks are superimposed.

The importance of these corrections can be seen by comparing the bands in Figure 2c with those in Figures 2a and 2b.

We have found that the use of the band template permits the detection of bands which are barely visible (results not shown) and, together with OD slope and curvature information, the detection of individual bands in overlapping groups (for example the run of Cs in Figure 1).

The final reading of the sequence requires a reading of the order of the bands in the four tracks. Starting from the lower end of the gel, successive bases are identified by choosing bands, according to simple rules, from those present in a window ahead of the current position.

Errors

The sequence shown in the autoradiograph in Figure 1 has been fully determined. The first 240 bases are:

```
GTCCCACAAA ATCAAACTCT TTGGACAGCC ACCATGTGCC TTTGTAACAT     50
TCCGAAGCGC TGCTGAACGA GACAAGGCCT TGCGAGTGCT GCACGGTGCT    100
CTCTGGAAAG GCTGTCCGCT CAGCGTACGC CTGGCCCGAC CCAAGGCTGA    150
CCCCATGGCT AGGAAGAGGC GGCAAGAAGG TGATAGTGAG CCATCAGTAA    200
CACAAATTGC CGATGTGGTG ACCCCTCTGT GGACAGTGCC
```

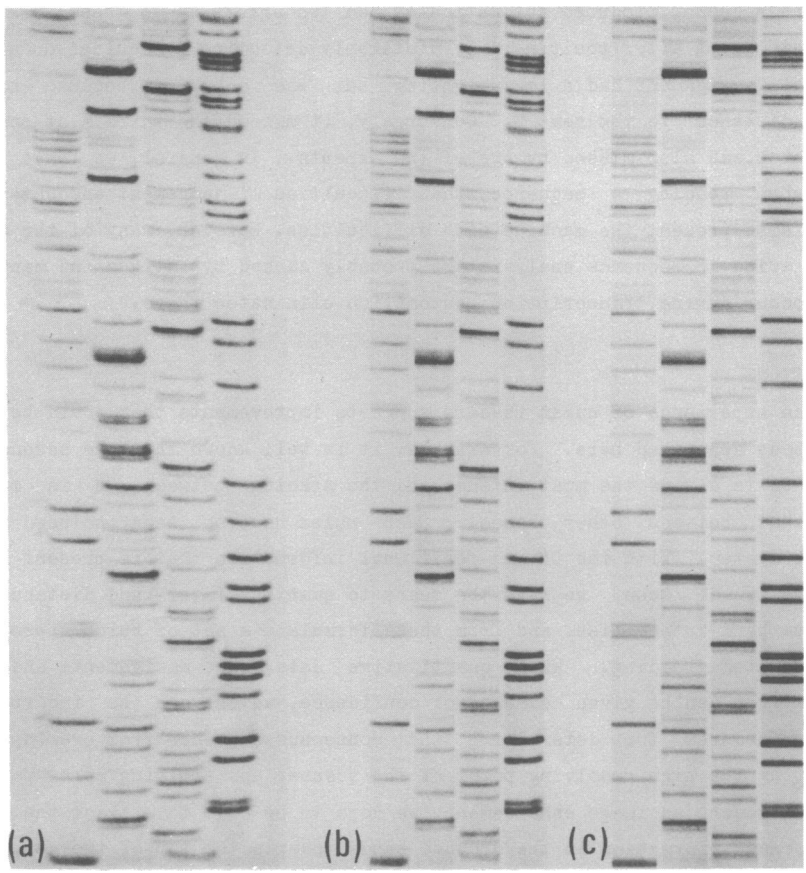

Figure 2. Correction of gel distortions. (a) A section of the digitized image of the autoradiograph in Figure 1. (b) The image in (a) after track straightening. (c) An "ideal" gel reconstructed from the four OD profiles produced by application of the band templates to the tracks in (b).

Three individuals with experience of reading sequences read the sequence from the four tracks shown in Figure 1. All three made at least one error:

(1) one skipped a track and read the C at position 112 as A;

(2) two included the clear band in the C track at position 123, which was shown to be an artefact from the sequence of the opposite strand;

(3) one missed the A at position 206 which ends a run of three As.

Apart from these errors, the three individuals agreed with each other, and with the above sequence as far as base 217. Two of them read a further 30-40 bases, but with increasing error rate and disagreement. Five novices also read the gel; their error rates were much higher.

The sequence produced automatically had two errors up to base 222. It inserted an A after position 175, mistakenly detecting a shoulder on a doublet. This assignment had a low weighting, but was accepted because of the large distance to the next G. Conversely, it missed the third A at position 206, which was also missed by one of the experts. In general, we have found that when reading a sequence, the difficulties of judgement encountered by humans also present the machine with difficulties. However, many of the errors which arise in sequence analysis are probably caused by fatigue and many mistakes occur during transcription; automation eliminates these.

CONCLUSIONS

The experience of human readers suggests improvements that could be made to methods described here. For example, it is well known that the second in a run of Cs is always the most intense and the spacing between Cs is smaller than that between other bases. Such rules have not been included in the present system. With the OD and positional information that is present in the two-dimensional scan, we have the means to quantify inter-band distances and relative band intensities, and from these formulate a set of rules appropriate to automated reading. With quantitative data, base assignments and their alternatives can be given measures of confidence, which can be incorporated into procedures for determining the consensus sequence from overlaps. By exploiting the high resolving power of the scanner and applying more powerful analyses such as image enhancement, we hope to be able to extract the wealth of sequence information in the crowded bands towards the top of the gel, which the eye is incapable of resolving. With these improvements, the automated system should be capable of a performance exceeding that of humans.

ACKNOWLEDGEMENTS

We thank Robin McDermott for contributing the major part of the scanner design, Donald McLeod for providing the autoradiographs and Denis Rutovitz for critical reading of the manuscript.

REFERENCES

1. Staden, R. (1984) Nucleic Acids Res. 12, 499-503.
2. Komaromy, M. and Govan, H. (1984) Nucleic Acids Res. 12, 675-678.
3. Gray, A.J., Beecher, D.E. and Olson, M.V. (1984) Nucleic Acids Res. 12, 473-491.
4. Toda, T., Fujita, T. and Ohashi, M. (1984) Electrophoresis 5, 42-47.
5. Savitzky, A. and Golay, M.J.E. (1964) Anal. Chem. 40, 1627-1639.
6. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.