Abstract

A genome mapping system has been developed that reads and assembles data from clones analysed by restriction enzyme fragmentation and polyacrylamide gel electrophoresis. Input data for the system can be most effectively obtained by the use of a scanning densitometer and image-processing package, such as that described in this article. The image-processing procedure involves preliminary location of bands, cooperative tracking of lanes by correlation of adjacent bands, a precise densitometric pass, alignment of the marker bands with the standard, optional interactive editing, and normalization of the accepted bands.

Introduction

As described in our previous publications (Coulson et al., 1986; Sulston et al., 1988), a suite of software has been developed to assist in mapping the genome of *Caenorhabditis*. Here we present the routines devoted to scanning densitometry and image processing of the autoradiographs used in the project. These programs have been in day-to-day use, in essentially their current form, for over 2 years.

In our analytical procedure, genomic clones are processed by restriction digestion and isotopic labelling to yield mixtures of radioactive DNA fragments (Coulson et al., 1986). These mixtures are separated on thin polyacrylamide gels, groups of such samples being interspersed with standard markers. An autoradiogram of a typical gel is shown in Figure 1. The task of the data acquisition routine is to detect the sample bands and to normalize their positions with respect to the marker bands.

System and methods

The code is written in VAX FORTRAN 77, and is currently run on a VAX 8600 with a VMS operating system. Films are scanned by a custom-built scanning densitometer and interpreted by an image-processing system. An AED 767 or 1024 raster graphics workstation is used for viewing and editing the processed data. The programs are available on request.

Densitometer

The densitometer (Figure 2) is designed to scan 420 × 170 mm X-ray film. By means of rollers, the film is driven lengthways past a slit at a constant speed of 2 mm/s. The slit is illuminated from below by a fluorescent tube and is imaged by a lens above onto a 1728-element linear CCD array. The image reduction is such that the CCD array receives light from a strip of film 0.1 mm wide. Stripes of data are read out at 0.05-s intervals, i.e. every 0.1 mm along the film. The CCD logic only integrates light for 0.01 s before each readout. This reduces smearing of pixels in the longitudinal direction, which occurs due to continuous film movement. The CCD output is converted to optical density by a logarithmic circuit and digitized by an 8-bit ADC.

Photodiode dark currents are individually corrected prior to the log amplifier. Sensitivity variations between individual diodes, together with illumination non-uniformity, are handled by adding an optical density correction. This is done by software in the controlling computer. The fluorescent tube is fed from a high frequency oscillator (>20 kHz), powered by a stabilized supply. This produces a ripple-free light output. Care was needed in selection of fluorescent tubes, as some makes have plasma position instabilities.

Pre-treatment of film

Care is taken to ensure that, as far as possible, the films are optimally exposed and free from extraneous marks. Unless the loading wells are all clearly visible already, a line is drawn through them to provide emphasis (see TlAR below). Labelling is confined to the region above this line.

Standard

At the beginning of a new project, a list of standard band positions is generated and placed in the file STANDARD.DAT. The list should be chosen judiciously to represent a typical film. In our own case, we necessarily transferred the positions that had been used graphically for manual input (Sulston et al., 1988); other projects have been started by using program NEWSTD to create STANDARD.DAT from part of the output of TlAR.

Implementation and algorithms

Scanning

Scanning is controlled by the program MAP, running in a local processor (PDP 11/03). Optical density values are summed in groups of four across the film to give a final pixel size of 0.1 (length) × 0.4 (width) mm, and the resulting integers are stored...
on disk in a custom built file server system to which the VAX 8600 also has access. After 4200 readings, the scan terminates and a flag number is placed in a handshaking file on the disk to activate the software in the VAX.

Image processing

The principal steps in image processing are:

(i) transposition of the picture data, to increase speed in subsequent operations;
(ii) a preliminary densitometric pass, to locate most of the band-like features;
(iii) a search for stacking of these features along the length of the film, to detect lanes;
(iv) a precise densitometric pass, following the centres of the lanes; and
(v) alignment of the marker bands with the standard.

MAPR (parameter input and control). Requests a file name, parameters to define marker lane arrangement, and the relative intensity threshold for band acceptance. For a given operator these parameters can be set up as defaults. MAPR then writes a command file (GSCAN.COM) that controls the image processing, and submits it to the VAX batch queue.

MAPROT (transposition and compression of picture data). Waits until the flag number appears in the handshaking file, then reads in the picture data, transposes the matrix so that the fast reading direction is along the length of the film, and writes back to the file chosen in MAPR. At the same time the data is compressed to bytes. The transposition is achieved via a small array, chosen to lie within the RAM working space available to the program, so that all disk accesses (including those involved in virtual memory) are performed efficiently. In this way, the c.p.u. time required is kept down to about 15 s.

T12DNS (preliminary densitometric pass). Detects band-like features by two nested loops. The outer loop detects peaks of density in each stripe of pixels along the length of the film. This routine smooths the data by adding pixels in overlapping groups of three or five, finds all maxima and minima, draws a base line from top to bottom passing through all minima (with the proviso that it may not rise more than a pre-set amount between minima), calculates integrated density for each peak, discards peaks below a pre-set absolute threshold, and calculates a new threshold as a pre-set fraction of the mean of the remaining peaks. The algorithm is very sensitive to small peaks.

Fig. 1. Autoradiogram of a mapping gel. The five lanes with closely spaced bands are markers. The band present in all the sample lanes is derived from the vector. Several overlaps can be seen, because the clones were not picked at random but were pre-selected by hybridization to a mixture of probes. Reprinted from Sulston et al. (1988).
and is dependent upon appropriate thresholds and smoothing to eliminate noise.

The inner loop detects features, by looking for peaks in adjacent stripes that are aligned across the width of the film. It accepts the surviving peaks from each stripe in turn and searches for overlaps with peaks in the previous stripe according to a pre-set tolerance. When an overlap is found, appropriate elements in a feature array are updated or initiated; termination of features is indicated by absence of an overlap.

The output routine calculates mean feature dimensions, and passes features that do not deviate by more than a pre-set amount from the mean. The overall effect of successively discarding deviants from the mean is to provide a reasonable representation of the most common feature type. The output includes approximate values for the inclination of features to the width axis of the film; these inclinations can in principle be integrated to give contour values and used to correct for 'smiling'. However, the density of features in the sample lanes of our mapping films is too low to give reliable contour information; we therefore prefer to rely upon temperature uniformity achieved by a metal plate clamped to the gel, which allows a linear interpolation between the marker lanes.

**D9AR** (*lane detection*). Because its decision making is based on small numbers of pixels, the feature data produced by T12DNS is noisy: closely spaced features are not well resolved and spurious features are generated. D9AR uses the centres of the detected features in order to track the lanes. First, it creates a histogram of feature density in a pre-set zone across the width of the film near the top; then it finds the edges of the central block of features, corresponding to the outer marker lanes, and divides the distance between these markers into \((n-1)\) equal parts, where \(n\) is the number of lanes. Provided the edges of the film are clear, this procedure reliably gives a starting point within each lane because the lanes are evenly spaced near the top of the film. The lane-following routine now takes over, moving first up and then down from the starting points; it is able to delineate curves and distortions provided that they are not too severe (Figure 3). The principle is the generation of a width histogram (as before) which is multiplied by a triangular wave function having maxima at the current lane positions. The multiplication is repeated with a series of small distortions in
**Fig. 4.** Marker alignment. (A) Graphical representation of alignment between standard (upper track) and output of TIAR for marker lane (lower track). (B) Illustration of cost calculation for a portion of the alignment indicated in (A). Standard at left side, data from TIAR above. For each sensible possible match, the lowest cost for an alignment up to and including the match is shown. This is found by considering possible previous matches and picking the one whose cost plus increment is smallest. A line is drawn back from the new match to the previous match that minimizes this sum, which is the desired optimal cost so far. The procedure progresses from top left to bottom right; then the lowest total alignment cost is chosen (49 in this example), and the corresponding optimal alignment is found by chasing back along the backward connections (heavy line). Since distortions are penalized as well as unmatched bands, the optimal alignment will tend to be close to a straight line in the diagram. In practice, the procedure starts off by forcing the topmost (well) bands to be matched, and the cost calculation is then performed for the entire lane before chasing back.
the spacing of the maxima in the triangular function, the lane positions are reset to the highest scoring pattern, and the process is repeated at the next level along the film. When the whole film has been scanned in this way, the resulting pattern is smoothed slightly to remove any abrupt changes in lane position or spacing.

Two types of interactive options are available in D9ARD (an interactive version of D9AR) if required (see below). The first allows manual marking of the edge start points; the automatic lane following then takes over. This option is useful for films with dirty edges or for selection of particular regions (for example, when the marker lanes are unevenly distributed). The second allows local adjustment of the final pattern, sometimes needed to deal with very faint or severely distorted lanes.

**TIAR (final densitometric pass).** Scans down the centre of each lane, summing a pre-set number of pixels widthways at each point, and summing in overlapping groups down the film like T12DNS. Detects and filters peaks in basically the same manner as T12DNS, but with two enhancements. The first is that a running mean of peak width at half height is calculated, and adjacent peaks are fused if they are very close together and excessively narrow. The second is that around each peak a search is made for subsidiary peaks, or shoulders, by considering the derivative of the intensity. The reason why this approach was not used from the start to find all the peaks is that it is too sensitive to noise where there is no density: the TIAR strategy restricts attention to places where there is at least one band. An important feature of the output is that it provides the position of the maximum rather than the mean of each band: this is required because bands in thin gels often, but variably, have a slight smear, so that only the position of the most dense region is significant. The output positions from the marker lanes are used by T6R, and those from the sample lanes by TAR.

**T6R (marker alignment).** For each marker lane in turn, T6R calculates an optimal alignment between the detected bands and the standard. The alignment is found by a dynamic programming method related to the Viterbi decoding algorithm (Viterbi, 1967). This minimizes a cost function containing penalty terms for local distortions and unmatched bands in either the detected or the standard pattern; band intensities are not explicitly matched, but preference is given to the stronger bands in the detected pattern by increasing the penalty for leaving them unmatched. The method works by starting at the top of the film and building up a set of possible alignments, considering each possible pair of matching bands in turn (see Figure 4). For each pair the method tests all the matches that could come just above it, by adding the cost of the new increment to that of the alignment above the tested match. The best one is picked and the cost of the resulting alignment up to the current match is saved, so that the procedure can be iterated. When this process reaches the bottom of the film, the optimal alignment is found by chasing back up the list of optimal previous matches. Having aligned the marker lanes, T6R calculates a normalization grid for the sample lanes by linear interpolation between the nearest marker lanes and bands. We attempted to improve accuracy by averaging over several bands, but found that in practice the use of more than two causes substantial distortion at the top and bottom of the film.

**T7A (computation of distorted standards).** Using the normalization grid in reverse, calculates standard band positions for every lane. Originally generated simply as a debugging aid, the display of these positions, superimposed upon the picture data, now provides a convenient check on lane following and marker alignment (see TARD below).

This completes the routines run by GSCAN.COM. A film such as that shown in Figure 1 requires some 2 min of c.p.u. time. Usually, each film is scanned and image processed while the previous one is being edited. The operator now uses the AED display to check the lane alignment.

**Editing**

**TAPS (display of picture data).** In order to match the 4096 pixels along the length of the film to the 1024 width of the AED memory the pixels are added in groups of four. This operation also has the effect of restoring the aspect ratio of the image (already compressed 1:4 widthways by MAP, above), but means that the perceived resolution is less than that available to the computer.

**TARD (alignment check).** Cancels the least significant bit plane of the AED display. Uses this bit plane to display T7A's output in a contrasting colour. If the alignments are satisfactory, the operator proceeds to TAR. If necessary, the lane following can be corrected by means of the interactive options in D9ARD (see above). Marker alignment does not fail unless the bands are very faint or smeared; in case of desperation, the operator can retouch offending regions by hand and then rescan the film.

**TAR (interactive editing).** Requests gel number (for subsequent indexing of film library). Displays T7A’s output over marker lanes only, together with the set of bands detected for one sample lane at a time. By means of a cursor that moves from band to band the operator can selectively delete artefactual bands before moving on to the next sample lane. As each lane is completed, the accepted bands are normalized and written to the database.

Correctly exposed autoradiograms from clean gels need little editing, provided that the relative intensity threshold is set appropriately. TAR has an option for automatic removal of the common bands (derived from the cloning vector) from the sample lanes. Thus fully automatic data entry can in principle be achieved. Overall, however, project time is saved by use of the editor rather than by insistence upon perfect gels.
Discussion

In choosing electrophoresis in thin polyacrylamide gels as the separation technique for the project, we were influenced by the superior resolution provided by polyacrylamide as compared with agarose. For image analysis, however, the former medium presents two particular problems—at least in the manually poured gels that we use.

The first is that the lanes are not straight, so that a simple geometrical algorithm for delineating them (which is satisfactory for agarose gels, e.g. Gray et al., 1984) is not practicable. Initially we explored analysis of density at low resolution as a means of finding lanes (cf. Elder et al., 1986), but found two limitations with this approach. One was that only the marker lanes have a consistent enough density to contribute to lane following by this procedure. More important is that, occasionally, background density (presumably resulting from the rapid movement of unincorporated radioactive monomer down the gel) forms a track different from that containing the bands. We therefore decided that the routine should first search for band-like features and then assemble the features into lanes.

The second problem is that the potentially high resolution of polyacrylamide can only be realized if each gel is accurately calibrated. For this purpose, the marker should be a DNA mixture providing plenty of bands, because the rate of molecular migration changes slightly but erratically from point to point down the gel. A Sau3AI digest of lambda DNA, originally chosen because its well-textured appearance facilitated manual input (Figure 1), has proved to be convenient and effective for automatic input as well. The image analysis system had to be capable of recognizing the marker pattern without constraints upon the type of distortion to be expected. The Viterbi algorithm has proved ideal for this task. With further programming effort, the algorithm would undoubtedly be capable of a still more robust performance; in particular, by providing a secondary standard that adapts towards the observed pattern after each correct alignment, it should be possible to avoid the need to tailor the standard to the operator.

A third potential problem in these gels is that of 'smiling', caused by the outer, cooler lanes running more slowly than the inner, hotter ones. The effect is greatly reduced by clamping an aluminium plate over the front of the gel assembly; with six sample lanes between each pair of markers, linear interpolation then provides a sufficiently good approximation. Although smile information can be extracted from T12DNS, the sample lanes are too sparsely populated to provide accurate contours throughout the gel.

Before long, the system that we are using may be superseded by methods employing direct readout machines (Ansorge et al., 1986; Smith et al., 1986) or image plates (Miyahara et al., 1986; Whiting et al., 1988). However, it should be remembered that the autoradiographic approach still presents certain advantages, in its low capital cost and its provision of a high resolution image for direct comparison of fingerprints, which must be weighed against the advantages of more fully automatic methods. Of course, the algorithms that we have presented here could equally well be used for image analysis based on data collected by other means.

Acknowledgements

We are greatly indebted to M. Fordham and M. Thompson for constructing the scanner, and to Denis Rutovitz and other members of the Clinical and Population Cytogenetics Unit in Edinburgh for showing us their scanner in prototype form. We have benefited from discussion with numerous colleagues, but particularly with Toby Gibson and John White.

References


Received on August 31, 1988; accepted on December 6, 1988

Circle No. 3 on Reader Enquiry Card