# PREPROCESSING ALGORITHMS FOR AUTOMATIC DNA SEQUENCE READING

Baozhen Fan and Donald G. Bailey
Massey University
Palmerston North, New Zealand

## ABSTRACT

Image processing algorithms for DNA sequence reading from DNA sequence gel autoradiographs are presented in this paper. Background normalization, contrast enhancement, boundary extraction and DNA sequence detection algorithms are respectively described in individual sections. The background normalization algorithm removes the variation of the background. The contrast enhancement algorithm extends the readable range of the captured DNA sequence image and increases the reliability of the DNA sequence reading system. The boundary extraction algorithm detects the boundaries of each lane set which are then used to correct geometric distortions. DNA sequence detection algorithm selects only the bands, determines the order of the bands in the sequence and joins different parts of the seque ce.

Version 4 of VIPS is used as the image processing algorithm development environment for this project.

## INTRODUCTION

The genetic information of a living organism is encoded by the DNA contained within every cell of that organism. DNA itself consists of a chain of nucleotide residues derived from the bases adenine, cytosine, guanine and thymine. Thus the genetic code can be determined by reading the sequence of bases within the DNA[1]. Although the sequence may be read manually, manual reading and transcription of the sequence into computer are tedious and therefore prone to errors. Several commercial DNA sequence reading systems are available, which use different approaches, for example, a fixed laser beam[2], a digital-pad[3] or a dedicated image processing package[4]. These are either expensive, limited in what they do, or highly specialised. The goal of this project is to automatically read DNA sequence data from a gel autoradiograph using a general purpose image processing system. DNA sequence reading system described has been implemented using version 4 of VIPS (Vision Image Processing System) which currently runs on a MicroVAX under VMS, an IBM compatible PC under Windows, and a Macintosh[5].

After an image is captured from a DNA sequence gel autoradiograph, the background is normalized to remove the variation of the background by dividing the image by a background image. Contrast is then enhanced to make the faint bands more readable by stretching the intensity range of the bands using a series of local filters and point operations. Because the captured DNA sequence image consists of several lane sets, each individual lane set must be extracted into a subimage before it can be read. The gap between lane sets is used to separate subimages. The boundaries of each lane set are extracted for correcting geometric distortions. A linear convolution filter and local maximum filter extract the edges of the lane sets. A distance image is used to remove centre edges. The last stage is to detect the order that the bands occur in each of the four lanes and then to merge the subsequence where appropriate, to get a longer sequence.

The background normalization, contrast enhancement, boundary extraction and sequence detection algorithms are described separately in the following sections.

## BACKGROUND NORMALIZATION

An image is acquired from a DNA sequence gel autoradiograph (see Fig. 1). There is often an intensity variation from one side to the other of a captured image

Fig. 1 Image acquired from a gel autoradiograph.

because of uneven lighting. Fig. 2 shows part of a background image (that is without any autoradiograph) and its intensity profile along the line A-B.


Fig. 2 A part of background image with profile of A-B.

Background normalization is necessary to remove this variation. To normalize the background, it is necessary to acquire an image of the background. Any noise in the background image is removed by using a moving average filter with a 15 × 15 window. Each pixel on the object image is divided by the corresponding pixel in the background image. The equation of divide operation[6] is shown below:

$$P_{out}(x,y) = \frac{P_{in}(x,y)}{P_{bg}(x,y)} \times \text{constant} \qquad (1)$$

where, $P_{out}(x,y)$ is the resultant pixel value in output image, $P_{in}(x,y)$ is the pixel value in the object image, $P_{bg}(x,y)$ is the pixel value in the background image. The constant is set to 220. Any output pixel values which exceed 255 are clipped to 255. This process scales the image according to variations in the background intensity, removing any such variations from the object image.

## CONTRAST ENHANCEMENT

Some of the bands in the image are quite faint, and are not able to be detected reliably. A contrast enhancement step is required to make the faint bands more detectable. Fig. 3 shows the data flow diagram for contrast enhancement.

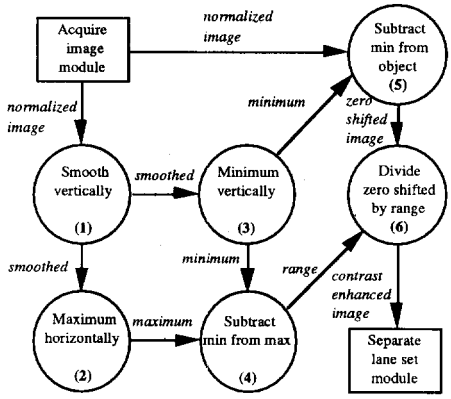


Fig. 3 Contrast enhancement data flow diagram.

1) The first step is to smooth the image vertically using a 15 × 3 moving average window. This smooths out any point noise which may be present in the image, and extends the bands vertically.
2) The contrast can be represented by the range of pixel values in the image. A maximum image is obtained using a maximum filter with a horizontal 1 × 8 0 window.

$$P_{out}(x,y) = Max(p_{11}, p_{12}, \ldots, p_{ij}, \ldots, p_{mn}) \qquad (2)$$

where the window has m × n pixels. A horizontal window is used to ensure that the maximum values come from the gap between the individual lane sets. A wide filter allows for considerable variation in lane set widths from image to image.

3) The minimum image is obtained using a minimum filter with a vertical 200 × 1 window.

$$P_{out}(x,y) = Min(p_{11}, p_{12}, \ldots, p_{ij}, \ldots, p_{mn}) \qquad (3)$$

A vertical window is used to enhance each lane set individually. The different lane sets can have different contrast, and require different degrees of enhancement. A tall window allows for lanes which have bands very widely spaced.

4) A range (or contrast) image is obtained by subtracting the minimum image from the maximum image.
5) The minimum image is also subtracted from the normalised image to give the zero shifted image. A constant value of 5 is added to this difference image so that the gaps between the lane sets remain white. This effectively prevents enhancement of any noise in the gaps.
6) The zero shifted image is then divided by the range image according to equation (1). This stretches the image on the basis of contrast (giving Fig 4). Where the contrast (range) is low in the image, it is stretched more than where the contrast is high.

Fig. 4 Contrast enhanced image from Fig. 1.

## BOUNDARY EXTRACTION

The next step in processing the autoradiograph image is to separate the lane sets into individual subimages (Fig. 6a). This segmentation step is reasonably straight forward and is detailed in Fan and Bailey[7]. Geometric distortions often occur in DNA sequence images. Most of these distortions may be corrected by finding the boundaries of the lane set in the subimage and making left and right boundaries vertical. Fig 5 shows the data flow diagram for extracting the boundaries.

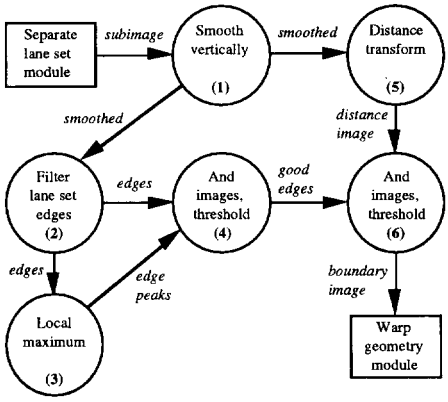


Fig. 5 Boundary extraction data flow diagram.

1) The first step is to smooth the image vertically using a 40 × 3 moving average window. This makes the vertical features (the bands) in the image more visible (Fig 6b).

2) A linear convolution filter extracts the vertical edges of the lane sets. This may be represented by the following equation.

$$P_{out}(x,y) = \frac{1}{m * n} \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} * p_{ij} \qquad (4)$$

where $w_{ij}$ are the kernel weights within an m × n window. In this case a 3 × 3 window is used with the weights {1 0 -1; 1 0 -1; 1 0 -1} to give the edges as in Fig 6c.

3) A local maximum detection filter detects the maximum pixels in a moving window from the edges image.

$$P_{out}(x,y) = \begin{cases} 1 & \text{if } P_{in}(x,y) = Max(p_{11}, p_{12}, \ldots, p_{mn}) \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

If the pixel at (x,y) in the input image is unchanged by filtering with a maximum filter (equation 2), it is a point of local maximum. A 1 × 15 window is used to detect maxima corresponding to each of the detected edges along each row.

4) False edges (where the local maxima is 0) are removed by logically *and*ing the two images pixel by pixel.

$$P_{out}(x,y) = P_{in1}(x,y) \text{ and } P_{in2}(x,y) \qquad (6)$$

This is then thresholded (equation 7), so that only the edges that are not 0 in both images are kept.

$$P_{out}(x,y) = \begin{cases} 1 & \text{if } P_{in}(x,y) \geq threshold \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

The resulting edges are shown in Fig 6d.

5) Only the left and right boundaries of the lane set are required. Approximate boundary positions may be found by thresholding the vertically smoothed image from step 1. This is then distance coded to determine the distance of each pixel from the boundaries.

$$P_{out}(x,y) = \|(x,y) - (\text{nearest black pixel})\| \times \text{constant} \qquad (8)$$

The distance coded image is then inverted to make the central region darker (Fig 6e).

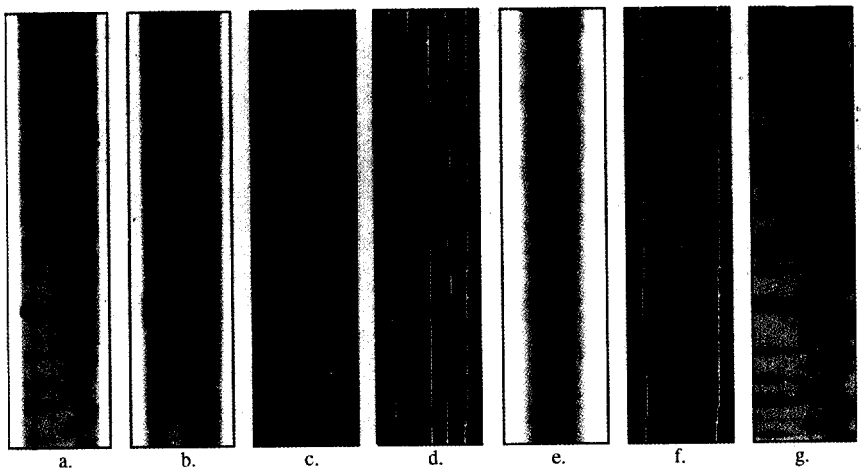


Fig. 6 Boundary extraction: a. a subimage; b. vertically smoothed image; c. filtered edges; d. maximum extreme; e. distance image; f. lane set boundaries; g. geometry corrected image.

6) The distance image is then *anded* (equation 6) with the edge image to code the edges with their distance. This reduces the strength of the edges within the lane set.

   The image can then be thresholded (equation 7) to extract only the lane set boundaries. These boundaries are shown in Fig 6f.

The left boundary is used to straighten the left side of the lane set and the right boundary is used to stretch the lane set to make the right side straight. This geometry correction procedure is described in detail in Fan and Bailey[7]. Fig. 6g shows the image after geometry correction.

## DNA SEQUENCE DETACTION[7]

The background does not contain any information of interest, and background is lighter than the bands, so any variation in the background caused by noise may be removed by clipping (see Fig. 7a). A 3 × 3 linear convolution filter (equation 4) with kernel weights of {1 1 1; 2 2 2; -3 -3 -3} is used to detect the horizontal edges associated with the bands. A non-linear edge enhancement filter[8] (with a window size of 3 × 13) then further enhances the detected edges and separates close bands (Fig. 7b). Only the central portion of each lane is kept to avoids problems where the ends of bands extend into adjacent lanes. The maximum value in each row of each band is extended across the complete band using a maximum filter (equation 2). This is then thresholded to give the individual bands (Fig 7c).

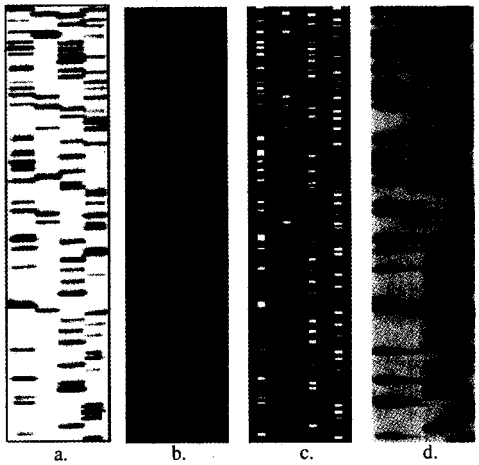


Fig. 7 Sequence detection: a. background removed image; b. extracted bands; c. thresholded bands; d. detected bands

The bands then are detected and scanned to determine the order in which the bands appear in the different lanes, hence the order of the bases Adenine, Cytosine, Guanine and Thymine in the DNA sequence. Fig. 8 is the DNA subsequence extracted from the above example. Different parts of a DNA sequence may be joined from different subimages or even from different images to obtain a longer sequence.

```
ATGAAATTGG  TAGAATGAGA  GACTTGAGAG
TGAATGTTAA  CACTATAAGG  TCGTTGTTAG
TTACAGAGCT  ACTATATGAG  TGAGTGTGTA
CTAGATGCAT
```

Fig. 8 DNA sequence of the example.

## DISCUSSION

At this stage, the image processing algorithms are still being refined. The algorithms work well for most of the examples tried, and they are now being tested on a wider range of autoradiographs. For the example shown in Fig 1, processing the image takes 7 minutes for the eight lane sets. This does not include the time required for the user to set up and acquire the image. The current algorithms require interaction with the user to verify the result and correct it if necessary.


## ACKNOWLEDGMENT

We would like to thank Dr Nick Ellison, Grasslands Research Centre, AgResearch (NZ) Limited, for providing invaluable technical information on DNA sequencing and for the provision of the autoradiographs used to develop the algorithms.


## REFERENCES


1. Davies R.W., DNA Sequencing, in "Gel Electrophoresis of Nucleic Acids: a practical approach", edited by Rickwood D. and Hamer B.D., IRL Press (1982), 117-172.
2. Pharmacia LKB Biotechnology, Automated Laser Fluorescent A.L.F. DNA Sequencer, S-75182 Uppsala (1989).
3. Jensen H.B., Instructions for Typeseq and Digiseq, Software information file.
4. HelixScan sales brochure, Helix, PO Box 85608, San Diego, California 92186-9874.
5. Bailey D.G. and Hodgson R.M., VIPS - a Digital Image Processing Algorithm Development Environment, Image and Vision Computing, 6(1988), 176-184.
6. VIPS Reference Manual and Users Guide, Image Analysis Unit, Massey University (1992).
7. Fan B. and Bailey D.G., Algorithms for DNA Sequence Reading by Image Processing, New Zealand Journal of Computing (1993), in press.
8. Bailey D.G., A rank based edge enhancement filter, 5th NZ Image Processing Workshop, Palmerston North (1990), 42-47.