IMAGE PROCESSING IN DNA SEQUENCE READING

Baozhen Fan
Computer Science Department
Massey University

Donald Bailey
Image Analysis Unit
Massey University

John Hudson
Computer Science Department
Massey University

ABSTRACT

The generation and analysis of DNA sequence data has played a significant role in the elucidation of biological systems. DNA sequence reading is a bridge between the generation and the analysis of DNA sequence data. The authors of the paper have developed software for directly reading DNA sequences from autoradiographs. This has been implemented using version 4 of VIPS. The problems in DNA sequence reading that need to be solved are low contrast, irregular lanes, variation in contrast on the background, unexpected horizontal lines among bands, persistent secondary structure bands, sequence scanning and dealing with data uncertainty.

INTRODUCTION

DNA strands consists of a sequence of 4 different bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA sequencing consists of determining the sequence of bases along the length of the DNA strand, so is an important feature of the characterisation of newly isolated genetic elements. There are three main stages to DNA sequencing [1]:

1) prepare DNA and carry out the DNA reactions
2) run the reactants on an electrophoresis gel and obtain an autoradiograph
3) read the sequence from the autoradiograph.

Preparation consists of breaking the DNA strand into short fragments, with each fragment starting from the same place in the sequence. These fragments are separated depending on the type of base at the end of the fragment. Since the fragments are of different length (and weight) they may be separated by using an electrophoresis gel. The samples are placed at one end of the gel, using a different lane for each base, and diffuse across the gel under the influence of an electric field. The smaller, lighter fragments travel faster and move further than the larger, heavier fragments. The distance moved in a given time therefore depends on the length of the fragment, hence the position of the base from the starting place in the sequence. By labelling each fragment with a radioactive element, an autoradiograph taken after electrophoresis shows where each fragment has travelled to, giving a banded appearance. The sequence is then determined by reading the sequence of bands in the four lanes.

Currently most DNA sequences are read manually. The DNA sequence data is then typed into the computer to be added to or compared with a DNA sequence database which has been previously obtained. This process is prone to error. A few DNA sequence reading software packages have been developed, for example Digiseq and HelixScan. Digiseq [2] is a digital-pad based DNA sequence reading program. An operator manually uses the digital-pad to point to each band of the sequencing gel autoradiograph. Variations of the gel running conditions lead to bent lanes on the autoradiograph and the digital-pad method does not always recognise the sequence correctly. HelixScan [3] uses a hand-held scanner to obtain images of the autoradiograph, which are then processed on a Macintosh computer.

The method described here has been implemented using version 4 of VIPS [4]. The software currently runs on a MicroVax under VMS, an IBM compatible PC under Windows, and a Macintosh.

IMAGE PROCESSING IN DNA SEQUENCE READING

Like many other image processing applications, DNA sequence reading involves three stages: preprocessing, feature extraction, data reading. The software structure is shown in Figure 1.

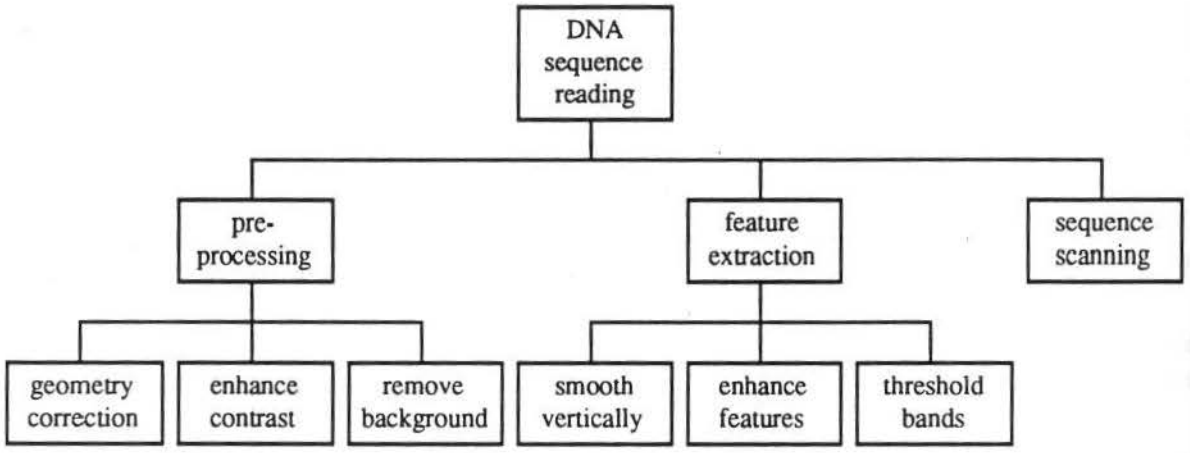


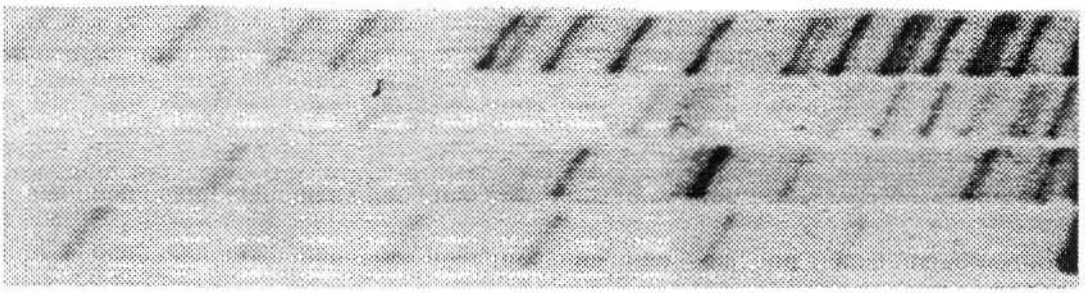
Figure 1. DNA Sequence Reading Software Structure.

GEOMETRY CORRECTION

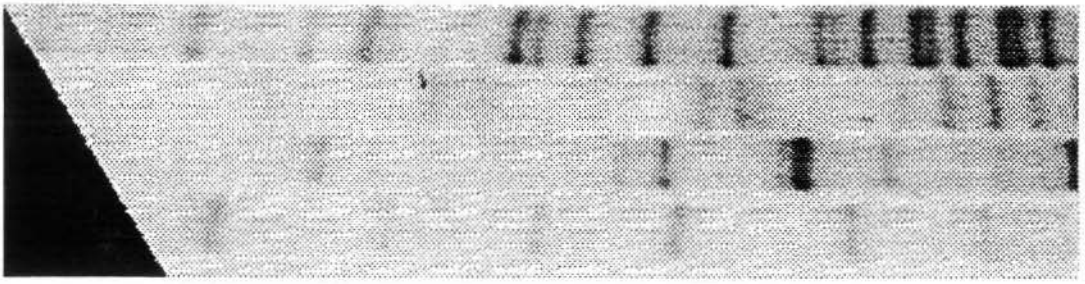
Variations of the gel running conditions lead to geometric distortions on the autoradiograph. In order to successfully process most autoradiographs, preprocessing is required to correct for these distortions.

Uneven gel running temperature and other gel running conditions may cause bent lanes on the autoradiograph. Lanes may bend from side to side, or the bands along the lanes may be skewed. In either case, the group of lanes can be broken into sections and each section is straightened individually. At present, this step is performed manually by selecting key points on each section and shearing the image to make the lanes horizontal and the bands vertical. Examples are shown in figure 2 a) - d).

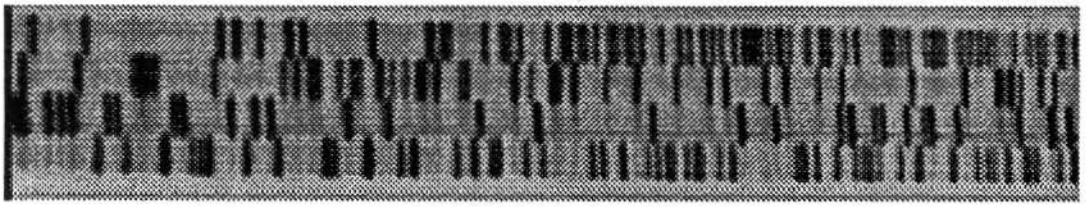
The lane and band straightening procedures may not be able to completely remove all of the distortion. Lanes may not be the same width over their complete length, and the gel running speeds may be slightly different for different lanes in the sequence (or even from one side of a lane to the other). These effects give a trapezoidal distortion which may be corrected by warping the image.



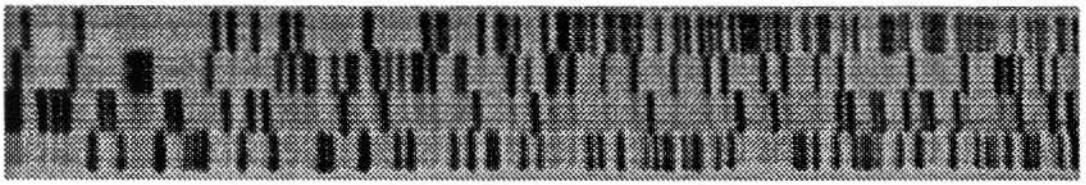
a) Part of a DNA sequence with skewed bands.



b) The image from a) after straightening the skew.



c) A DNA sequence image which is not straight horizontally.



d) The image from c) after straightening.

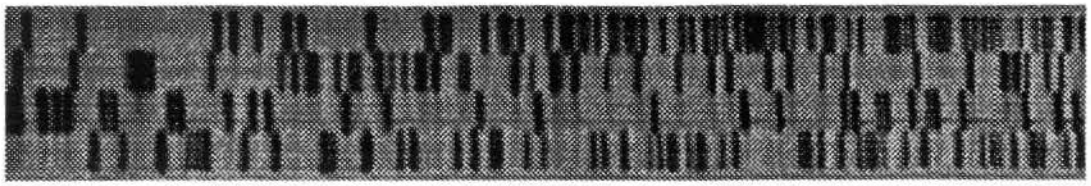
Figure 2. Straightening lines and lanes in a geometrically distorted image.

BACKGROUND PROCESSING

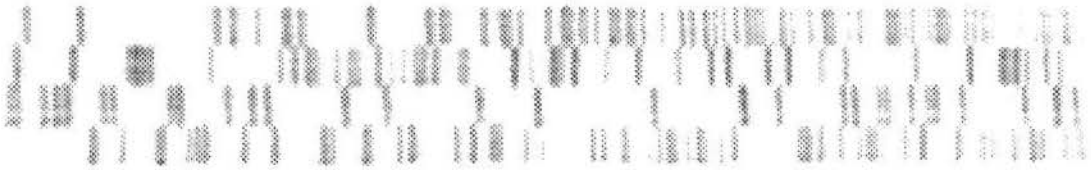
Varying the gel running times makes different parts of the DNA sequence readable on the autoradiograph since the smaller bases run faster within the gel. Therefore it is possible to get a very long DNA sequence by joining different parts of the sequence that are obtained from different gel running times, varying from short times for the smallest bases, to long times for the largest bases. However, if the gel running time is very short, the smallest bases often give low contrast on the autoradiograph. A local intensity stretch operation is used to enhance the contrast, making even faint bands readable.

Global thresholding methods are adversely affected by the presence of a linear background gradient. The background may be estimated by using a box filter, selecting the maximum value within a rectangular box. The variation in background intensity may be removed by subtracting the estimate of the background from the input image.

Uneven gel running conditions often cause unexpected horizontal lines among the bands (see Figure 3). By making the box only a few pixels high, these horizontal lines may be removed along with the background. A box size of 2 pixels high by 20 pixels wide was found to give the best results.



a) Input DNA sequence



b) The image from a) after subtracting the background.

Figure 3. Background processing.

FEATURE EXTRACTION

The main features of the DNA sequence image are the vertical bands. Keeping and extracting these features is the key problem in this application. The bands at one end of the sequence are usually closer than the bands at the other end, making it difficult to determine the band spacing. Bands resulting from secondary structures on the DNA strand may be bunched together on the autoradiograph [5], and need separating.

Vertically smoothing the image before feature extraction is necessary to reduce the noise and enhance the separation, especially where the bands are very close. Such a smoothing step increases the accuracy and extends the range of the DNA sequence that is readable from the autoradiograph.

A 3 x 3 linear convolutional filter is used to enhance the vertical bands of the DNA sequence. Since the bands are narrow and peaky, this filter detects even the wider bands. Where there are bands bunched together as a result of the secondary structures, it is able to separate these into individual bands a lot of the time. The following filter weights were used

-1	1	-1
-1	4	-1
-1	1	-1

After the individual lanes are separated a local edge enhancement filter [6] further enhances edges and separates close bands. Figure 4 shows a filtered and enhanced DNA sequence.

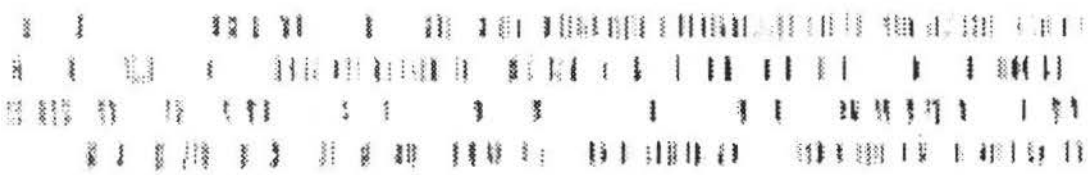


Figure 4. Feature enhanced image from Figure 3 b).

After feature extraction, the darkest pixel in each band is extended though the whole band to make detection more reliable. The image is then thresholded to detect the bands. Figure 5 shows the binary image resulting from this process.

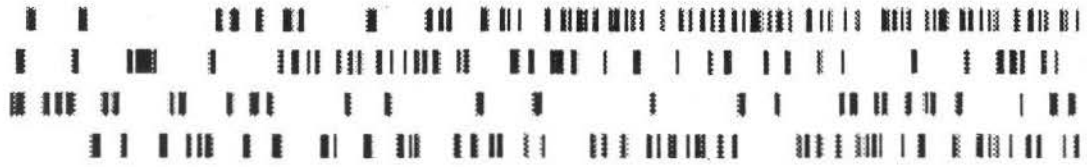


Figure 5. Image from figure 4 after thresholding.

SEQUENCE SCANNING

Each lane represents a different base in the DNA strand. The lane sequence is defined by the user (or is assumed to be T, C, G, A by default). The four lanes are scanned in parallel to determine the order in which the bands from the different lanes occur, hence the order of the bases in the DNA sequence. As the data is scanned, the sequence is saved to a file for later comparison with a database.

Uncertainty is a problem during DNA sequence reading. If a band in one lane is very close to one in another lane, it is hard to determine which one base comes first. Occasionally, two bands from different lanes are in the same column, even though neighbouring bands are far away. A second case of uncertainty is where the band is wide. There is uncertainty in determining whether it is a single band or two (or more) bands which have not been separated by the processing. In each case, an uncertainty code (*) is used to indicate that the result is uncertain. Figure 7 shows the DNA sequence extracted from this example.

```
G C G*T G G G C T A   G G A C C C*C A G G   A A A C T G T A G T
G A C T C T C C A A   C G C C A T C G C A   C A C A C T C T T A
C C A G T A A T T C   T A C G A T C C T T   C T T A T A C T T A
T C T A G A A A C T   A T A T A C T A T C   T A*G T T T C G T C
T A T A C T T A G C   T G A A G A G T T A   T G T*C A G T G T T
T A G T C T A T A T   T A C A C T G C A T   A T C T G G C T A G
T A*
```

Figure 6. DNA sequence from the image in figure 5.

DISCUSSION AND CONCLUSION

At this stage, the image processing algorithm is still being refined. Although it has been shown here that this approach works for the example given, it has not yet been tested on a wide range of samples to determine its weaknesses. The algorithm, as it stands, relies on the geometry correction stage to ensure that the lanes are horizontal and the bands are vertical. This stage is still performed manually, and needs to be automated to make the algorithm independent of any operator bias. The algorithm also needs to be extended to improve the performance in the regions in the input where the sequence is uncertain.

ACKNOWLEDGMENTS

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, Oxford, pp 117-172 (1982).

[2] Jensen H.B., "Instructions for Typeseq and Digiseq", Software information file.

[3] HelixScan sales brochure, Helix, PO Box 85608, San Diego, California 92186-9874.

[4] Bailey D.G. and Hodgson R.M., "VIPS - a Digital Image Processing Algorithm Development Environment", *Image and Vision Computing*, Vol 6, pp 176-184 (1988).

[5] Brown N.L., "DNA Sequencing", Methods in Microbiology, Academic Press, London, Vol 17, pp 259 - 313 (1984).

[6] Bailey D.G., "A rank based edge enhancement filter", *5th NZ Image Processing Workshop*, pp 42-47, Palmerston North (1990).