INTERNATIONAL WORKSHOP ON INDUSTRIAL APPLICATIONS OF
MACHINE INTELLIGENCE AND VISION (MIV-89), Tokyo, April 10–12, 1989

AUTOMATED READING OF DNA AUTORADIOGRAM IMAGES USING LANE PROFILING METHODS

K. PALANIAPPAN*
T. S. HUANG
H. LEE

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1101 West Springfield Avenue
Urbana, Illinois 61801 U. S. A.

Abstract

Currently there is a great deal of interest in automated high-speed DNA (deoxyribonucleic acid) sequencing for large scale genomic sequencing projects. Automated DNA sequencing involves computer interpretation of the chemical detection data which may be in the form of 2–D autoradiogram images, and 1– or 2–D fluorescence data. The analysis of autoradiogram images generated by the multiplex DNA sequencing method is considered. An overall approach to obtaining sequence data from the autoradiogram images is outlined and specific approaches for segmentation of the image into sets of four lanes, obtaining 1–D profiles of the lanes, detection of peaks in the 1–D profiles using a multiscale approach and aligning profiles across lanes using an interpolation method are discussed. Some intermediate results are also presented.

1. Introduction*

Currently there is a great deal of interest in automated high-speed DNA (deoxyribonucleic acid) sequencing for large scale genomic sequencing projects. Since the 1970's there has been an exponential increase in the number of nucleotides that have been sequenced each year (and is currently almost 20 million bases) [1], consequently collaborative efforts have been initiated among the United States', European and Japanese DNA Databanks to manage the expected tremendous increase in sequence data that will result from systematic genomic mapping programs [2]. Furthermore, Wada [3] estimates that automated sequencing techniques should reduce the cost from the present US $1 per sequenced nucleotide (approximate) by an order of magnitude given a potential throughput of 1 million bases per day.

One key step in automated DNA sequencing involves computer interpretation of the *chemical detection* data which may be in the form of 2–D autoradiogram images, and 1– or 2–D fluorescence data. Several commercial automatic film readers with varying accuracy and speed of reading have been developed [4, 3, 5]. We discuss approaches for analyzing 2–D autoradiogram images with consideration for reliability and computational requirements. Some of the approaches discussed are not only applicable to autoradiogram data but also one- and two-dimensional fluorescence data.

We briefly describe the experimental protocol used to generate the autoradiograms and introduce terminology. Church and Kieffer-Higgins introduced *multiplex* DNA sequencing [6] which mixes together different DNA fragments, with each fragment being flanked by two different oligonucleotide tags at the cloning stage. The mixed fragments are amplified , then undergo Maxam-Gilbert chemical sequencing to yield four sets of reaction products: G (guanosine), C (cytosine) + T (thymidine), G + A (andenosine), and C. The four reaction products are sorted by mobility in an electric field (which is proportional to size) in adjacent lanes of a sequencing gel and the result transferred to nylon membranes. The membranes are then probed with radioactively labeled complementary tag sequences; since each fragment has two unique tags it can be probed twice which allows for redundancy and error checking. Each probing produces an autoradiogram like the one shown in Figure 1. Distinct dark *bands* can be seen and these are equivalent to resolving a single nucleotide of the original DNA fragment sequence; the column location of the band (or corresponding bands for G or C) provides information for identifying the nucleotide type and the row location specifies the position of the nucleotide within the original DNA sequence fragment. The bands being approximately uniform in width, line up to form a *lane* or *track* and corresponds to one of the reaction products; so each group of four lanes represents the information for obtaining the complete sequence of a DNA fragment. The lanes, however, are not always vertically oriented and straight; the shape of the lanes is sometimes referred to as *well morphology*. The varying morphology of lanes, nonuniformity in band shape, size or spacing, and the shifting in alignment between lanes necessitates sophistication in automatic reading algorithms.

The advantage of the multiplex approach over standard DNA sequencing methods is in reducing the number of separate chemical reactions by multiplexing early in the sequencing protocol and *demultiplexing* only prior to forming an autoradiogram; so the speedup over conventional approaches is proportional to the amount of multiplexing [6]. The multiplex approach also provides an *internal standard* whose sequence is known and hence can be used to estimate distortion parameters as well as speed up the reading of the probe autoradiograms. Once the multiplex sequencing method has been optimized it may be feasible to probe in parallel 100 membranes each day (on a twenty day cycle basis with twenty probes) with each membrane containing about 5000 resolvable nucleotides of information in twelve groups of lanes, to generate approximately 500,000 bases of data per day [7]. Processing such a large volume of data will inevitably require robust algorithms to read, assemble and analyze autoradiograms.

The multiplex sequencing method operates in a batch style in the sense that the complete autoradiogram must be developed before the sequence can be read. Continuous on-line sequencing systems that do not require radioisotopes and autoradiogram recording have been developed using fluorescence based detection. Fluorescence based methods may produce either one-dimensional traces for each nucleotide [8, 5, 9] or two-dimensional images [10] resembling autoradiograms and may be based on a single-dye four-lane sequencing format or a four-dye single-lane format. Fluorescence based methods, however, also have some disadvantages including lower sensitivity, spectral overlap in the emission of the fluorescence dyes, changes in the electrophoretic mobility of the DNA fragment to which the dyes are bound [8, 11], slower scanning, less reliability, and less flexibility [4], as well as higher cost [3, 4]. A novel method that uses a multi-wire proportional counter (MPWC) to reduce the èxposure time required to detect radioactivity and form

* This research was supported in part by National Institute of Health grant PHS AI-1016 awarded to J. Konisky.

TH0250-1/89/0000-0360$01.00 © 1989 IEEE

an image along with algorithms for automatically interpreting the coarser MPWC images to determine the sequence is described in [12]. This paper uses only autoradiogram images resulting from multiplex DNA sequencing.

2. Analysis of Autoradiogram Images

Although radiograph images, particularly for biomedical diagnosis applications, have been analyzed by computer since the early 1960's [13], two-dimensional gel electrophoresis autoradiogram images (of usually protein materials) have been investigated primarily in the 1980's [14, 15, 16]. DNA autoradiogram images of interest to us, however, have been analyzed by computer only recently [4, 17, 18, 19].

Radiographic image analysis usually consists of six general steps: (i) digitization of film, (ii) preprocessing for image enhancement, (iii) segmentation, (iv) extraction of size and shape characteristics, (v) extraction of texture features, and (vi) classification [13]. These six steps could also be applied to the analysis of DNA autoradiogram images. However, rather than taking a two-dimensional approach involving boundary and region detection, shape description, *etc.*, the autoradiogram image is converted to a set of one-dimensional signals which are then used to determine the DNA sequence. This reduction of dimension is possible due to the underlying nature of the data and the classification task which is to recover a linear ordered DNA sequence from the image. This approach also offers several advantages including speed of processing, robustness to distortions, and applicability to the analysis of DNA sequencing data based on other methodologies including the fluorescence based detection strategies described above.

Figure 1 shows a complete autoradiogram image of a standard membrane based on the multiplex sequencing method; the original image is 3691 × 1451 pixels with two bytes per pixel for gray level information. The membrane is 43 cm × 35 cm so the sampling rate is approximately 116 microns in the vertical direction and 241 microns in the horizontal direction (or 453 dots/inch × 218 dots/inch). Although higher resolution may be desirable the current images already require 10.7 Mb (megabytes) of storage which is equivalent to about forty-one 512 × 512 video frames; a 50 micron sampling rate would require about 120 Mb per image or equivalently 459 video frames. There are 48 lanes in Fig. 1 with each group of four lanes required to form a sequence. Since this is a standard each set of four lanes is from the same sequence and can be used to estimate information about band and lane distortions and variations in morphology. Figure 2(a) shows a 768 × 768 subimage of the image in Figure 1 that reveals more detail about the band patterns and reflects some of the detail available in the original image. The overlaid grid gives indications about the size of the features in pixel dimensions; width of bands range from 60 to 100 pixels and thickness of the bands varies from about 10 to 25 pixels. Furthermore, even in this seemingly distortion-free region of the entire autoradiogram the curvature of the lanes (vertical features) and bands (horizontal features) are visible.

Figure 2(b) shows a histogram for the 768 × 768 region in the original image. It reveals the limited range of gray levels in this region, approximately half the 256 gray levels available. Contrast enhancement algorithms can improve the appearance of the image. For the image in Fig. 2(a) the gray levels have been linearly rescaled to the full range of 256 gray levels which was considered adequate and superior to histogram equalization. The histogram also shows that setting a threshold (even a locally adaptive one) for isolating the bands from the background would be difficult. Thresholding usually led to incomplete, missing or merged bands. One of the reasons for this is the variation in dynamic range of the band intensities. For example, the ratio of background intensity to band intensity for clearly visible dark bands ranges from 40 to 2; and for faint bands can be as low as 1.03, almost indistinguishable from the background. Furthermore, *companion* bands are also highly nonuniform in intensity which makes their detection and classification an even more difficult task. Companion bands are those bands that must appear in two lanes. For example, in the chemistry protocol used to generate the autoradiogram of Fig. 1, bands in the first lane indicate the presence and position of guanine (G) bases, in the second lane the pyrimidines (Y) which are cytosine and thymine (C+T), in the third lane the purines (R) which are guanine and adenine (G+A), and in the fourth lane C; so companion bands would appear in lanes one and three for each G or lanes two and four for each C. If a companion band is missed then the base could be mislabeled. The image intensity dynamic range for companion bands ranges from 2 to 0.25 so faint bands can be easily missed using simple thresholding or edge detection operators. In fact many edge operators (including popular 3 × 3 masks such as the Prewitt, Sobel, Frei-Chen, or moment-based as well as more sophisticated operators such as the zero-crossings in the Laplacian of the Gaussian, or maxima in the output of an oriented first derivative of a Gaussian operator proposed by Canny) gave unsatisfactory results due to spurious edge responses, missing edge boundaries, merged edges, and disconnected or shifted edge contours. These difficulties in edge detection would need to be overcome using more sophisticated post processing algorithms for linking, grouping and classifying edges. Consequently, neither the region detection based (using thresholding) nor boundary detection based (using edge detection) approaches was strictly followed.

The difficulties can be appreciated by the reader in trying to determine the sequence corresponding to the image in Fig. 2(a). There are 2 1/2 sets (of 10 lanes) corresponding to the same sequence. In the first four lanes for example one can resolve 40 bands which starting from the top reads (from left to right):

CATTGTTAGA TTTCATACAC GGTCCTGAC
tgCGTTAGCA

The lower case letters indicate a region of compression where two nucleotides should occur but are difficult to resolve in the image.

Since the objective is to process 100 autoradiograms (membranes) per day, this requires that each autoradiogram be analyzed in under 15 minutes (or about 21 seconds per video frame equivalent) from scanning to sequence recognition. Since computer analysis is decoupled from the electrophoresis equipment more time could be devoted to analyzing each film by using more computers. This can be readily accomplished by using one or more scanners to capture data and then using a group of locally networked computers to analyze each film (for n computers the available time for analysis would increase by 15n minutes, n < 100). In fact the same principle could be applied in order to distribute the work for analyzing *each* image by partitioning the image into blocks.

3. Image Analysis Methods

An overall paradigm for analyzing autoradiograms generated by the multiplex DNA sequencing approach is shown in Figure 3. It should be emphasized that the flow diagrams indicate a preliminary approach only portions of which have been tested. There is scope for optimization of the modules as well as improvements to the approach itself. The initial steps of digitization and quantization of scanner data, preprocessing (histogram modification, filtering and morphology operations for image enhancement), and identification of the film as being a standard or probe is shown in Fig. 3(a). This first stage also involves film registration to accommodate for the placement of the film with respect to the scanner and extraction of identification information (such as the numbers shown in Fig. 1 or possibly bar codes) to keep track of the image and the sequence in large sequencing projects. The processing steps for the standard are shown in Fig. 3(b) and for the probe in Fig. 3(c). The outputs are shown are dark shaded boxes and inputs as lighter shaded boxes. The outputs include the digitized (and possibly enhanced) image, registration information, lane geometry and distortion parameters. The inputs include sequence reading rules which reflect the sequencing protocol used and the sequence used for the standard. Based on the number of correctly identified bases in the standard, it may be rejected if there are too many errors due to distortions arising from experimental conditions. One of the advantages of the multiplex sequencing method is the availability of an internal standard for estimating lane boundaries, lane registration, band size, shape and spacing variations. Once these parameters are estimated then they can be used to speed up the reading of the probe sequences (up to 40 autoradiograms or more) and do not have to be recalculated each time as in the usual sequencing approach where an internal standard is not available.

We present some initial results and examine the methodology of the various steps for different stages of the approach shown in Figure 3. Methods for: (a) segmentation of the image into groups of four lanes, (b) one-dimensional profiling for each lane, (c) band (peak or valley) detection, (d) feature extraction, (e) inter-lane alignment, and (f) classification based sequence construction, are discussed.

3.1. Segmentation Into Lanes

In [18] lanes were detected using projections onto the horizontal axis, where pixels along each column for a given number of rows are summed together. Given an image $g(x,y)$ then a local x-projection, $C(x)$, is defined as $C(x) = \sum_{y \in W} g(x,y)$. The window size is influenced by the sampling rate as well as the size of the features to be detected. First derivatives in $C(x)$ are used to initially estimate the location of the lane boundaries which are refined by lane following. Figure 4(a) shows an example of an x-projection. Although the window covers 6 1/2 lanes only one can be detected. The reason the projection approach fails is because there are *no* distinct gaps between the lanes; a necessary requirement for the approach in [18]. Furthermore, as noted earlier there is a large variation in lane widths that needs to be determined from the image and not biased in terms of prior estimates.

A simple edge detector is first applied to find vertical edges in the image then maxima are sought in the x-projection to determine the initial location of the lane boundaries. Several 3 x 3 edge operators of the form

$$\frac{1}{s+2}\begin{pmatrix} -1 & 0 & 1 \\ -s & 0 & s \\ -1 & 0 & 1 \end{pmatrix} \tag{3.1}$$

were tried as well as the 2 x 2 difference operator $\begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$. The values of s that were tried are 1 (Prewitt), 2 (Sobel), 1.4142 (Frei-Chen), 2.0671 (moment) and 2.0668 (closed band). The 2 x 2 operator gave much fewer responses than the 3 x 3 operators all of which behaved quite similarly with the moment and closed band giving fewer spurious responses. Figure 4(b) shows the x-projection after applying the Prewitt operator from which the lane boundaries can be detected. The detection and localization can be improved by using long narrow operators such as 7 x 3, tailored to detect vertical edges, by suppressing nonmaxima in the horizontal direction (orthogonal to the edges), and by smoothing the x-projection profile to reduce noise. In order to refine the boundary estimates the correlation coefficient can be used to detect the transition from one lane to the next by searching in a small neighborhood around the maxima in the x-projection of the edge operator responses. Consider two adjacent columns $g(\mathbf{n})$ and $g(\mathbf{n}+\mathbf{k})$ in the image where $\mathbf{n} = (n_1, n_2)$ and $\mathbf{k} = (k_1, k_2)$ are integer coordinate pairs. Under the assumption of Gaussian noise in the two columns the pixel intensities can be considered to be samples from a bivariate Normal distribution. Then the maximum likelihood estimate of the correlation coefficient ρ is

$$\widehat{\rho} = \frac{\sum_{\mathbf{n} \in W} [g(\mathbf{n}) - \overline{g}(\mathbf{n})][g(\mathbf{n}+\mathbf{k}) - \overline{g}(\mathbf{n}+\mathbf{k})]}{\left(\sum_{\mathbf{n} \in W} [g(\mathbf{n}) - \overline{g}(\mathbf{n})]^2 [g(\mathbf{n}+\mathbf{k}) - \overline{g}(\mathbf{n}+\mathbf{k})]^2\right)^{1/2}} \tag{3.2}$$

where $\overline{g}(\mathbf{n})$ and $\overline{g}(\mathbf{n}+\mathbf{k})$ are estimates of the mean and the window W is a thin (one or few column wide) strip. When, N_W, the number of pixels in W is large or moderate the transformation of $\widehat{\rho}$ known as *Fisher's z*,

$$z = \frac{1}{2}\log_e\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right) \tag{3.3}$$

has an asymptotic Normal distribution with mean $z = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$ and variance $\frac{1}{N_W - 1}$. So the hypothesis $H_o : \rho = \rho_o$ can be tested using tables of the standard Normal distribution. The location in the search area where ρ_o drops below 0.75 for example can be used to delineate a lane boundary.

3.2. Lane Profiles

Most methods of analyzing DNA autoradiograms proposed to date have relied upon converting the two-dimensional image to a set of one-dimensional profiles or densitometric traces. The profiles are obtained with or without the use of column-to-column correlations within each lane. Some type of correlation or registration analysis is necessary when the bands are curved and non-horizontal in order to obtain a high resolution profile that reduces the effects of noise. Rather than using the conventional correlation function $\sum_{\mathbf{n} \in W} g(\mathbf{n})\, g(\mathbf{n}+\mathbf{k})$ which requires multiplications and is sensitive to brightness changes across the image, the following morphological correlation is used,

$$\mathrm{M}(\mathbf{k}) = \sum_{\mathbf{n} \in W} \min\left(g(\mathbf{n}+\mathbf{k}), g(\mathbf{n})\right) \tag{3.4}$$

Maximizing $\mathrm{M}(\mathbf{k})$ can be shown to be equivalent to minimizing $\sum_{\mathbf{n} \in W} |g(\mathbf{n}) - g(\mathbf{n}+\mathbf{k})|$ the sum of the absolute values of the differences [20]. In order to reduce the sensitivity of $\mathrm{M}(\mathbf{k})$ to absolute brightness levels the local means can be first subtracted from each column. One key advantage of using $\mathrm{M}(\mathbf{k})$ is that it is fast since each term requires only a comparison operation. The number of calculations can be further reduced by evaluating only partial sums of $\mathrm{M}(\mathbf{k})$ with thresholds to reject poor matches early in the matching process. Determining the displacement $\mathbf{k}$ with respect to the center of the lane is usually sufficient to follow gradually sloping bands. Figure 5 shows the profiles for the first four lanes of the image in Fig. 2(a) as an *inverse* image (that is the peaks correspond to the dark bands).

For geometric accuracy at the sub-pixel level some form of interpolation is required. For a fixed column, k_2, determine the optimal value of k_1,

$$k_1^{\max} = \max_{k_1} M(k_1, k_2) \tag{3.5}$$

Then sub-pixel accuracy can be achieved by fitting a quadratic function to several values of k_1 around $k_1^{\max}$ and selecting k_1^* as the location of the maximum of the function. A simple alternative that examines just one value of the correlation on either side of $k_1^{\max}$ is

$$k_1^* = \left(k_1^{\max} - \frac{1}{2}\right) + \frac{M(k_1^{\max}, k_2) - M(k_1^{\max} - 1, k_2)}{2M(k_1^{\max}, k_2) - M(k_1^{\max} - 1, k_2) - M(k_1^{\max} + 1, k_2)} \tag{3.6}$$

which shifts the value of $k_1^{\max}$ by up to half a pixel if there is asymmetry in the values of $\mathrm{M}(\mathbf{k})$ to either side of $k_1^{\max}$ [21].

3.3. Multiresolution Peak Detection

The peaks (which correspond to the dark bands) in Fig. 5 need to be reliably detected and accurately located. Peaks are modeled as being Gaussian shaped, $A \exp\left(-\frac{1}{2}\left(\frac{x-m}{b}\right)^2\right)$. The following three steps are used to initially locate and characterize the peaks:

1. Convolve the image I with the second derivative of the Gaussian, $\frac{\partial^2 G_\sigma}{\partial^2 x^2}$, for several values of the scale parameter σ (the range of σ is governed by the largest and smallest peaks to be detected). It should be noted that the result of the convolution does *not* shift the location of the peaks.
2. Mark the maxima in each $\frac{\partial^2 G_\sigma}{\partial^2 x^2} * I$ image where the dark bands in the image correspond to peaks in the profile.
3. For each maxima marked in Step 2, calculate $\frac{\partial}{\partial \sigma}\frac{\partial^2 G_\sigma}{\partial^2 x^2} * I$ using which the peak's scale size b can be estimated as

$$b = \sigma\left(\frac{3}{1-\sigma B_\sigma} - 1\right)^{\frac{1}{2}} \quad , B_\sigma = \frac{\frac{\partial}{\partial \sigma}\frac{\partial^2}{\partial x^2} G_\sigma * I}{\frac{\partial^2}{\partial x^2} G_\sigma * I} \tag{3.7}$$

and the peak strength A as

$$A = \frac{\left(b^2+\sigma^2\right)\frac{\partial^2 G_\sigma}{\partial^2 x^2} * I}{\sqrt{2\pi}\, b\, \sigma} \tag{3.8}$$

Since $B_\sigma = 0$ ideally at the center of a peak the estimate of b should be close to $\sqrt{2}\sigma$. So estimates for b that differ greatly from $\sqrt{2}\sigma$ should be rejected as candidate peaks at that scale.

The justification for this approach follows that of the multiscale region detector proposed in [22] for fitting disk shaped regions to textured images; in the above algorithm we use the model of Gaussian shaped peaks.

Since neighboring peaks interact the above estimates for the peak location, size and strength are iteratively refined using a maximum likelihood updating scheme [23].

3.4. Feature Extraction

A number of features that would be useful for constructing the sequence are extracted from the image. For the bands these would include: (i) location of the peak, (ii) location of the band centroid, (iii) band area (strength) and ratio of peak strengths between the lanes for a given row position, (iv) width and height of band estimated using a best fitting ellipse, (v) orientation of the ellipse, (vi) elongatedness , and (vii) irregularity of shape in comparison to an ellipse which can be estimated as the difference in area between the fitted ellipse and the actual band. Within each lane useful features are: (i) average spacing between bands, (ii) average height of bands, (iii) a model function for describing variation in band spacing from the bottom to the top of each group of four lanes, and (iv) a model for describing variation in band height within each lane.

These features along with rules for dealing with merged bands that appear as plateaus in the lane profile, compressed bands, faint closely spaced bands that often appear as shoulders around a peak are all used in the classification stage.

3.5. Inter-lane Alignment

Even in regions of the image where the bands appear relatively straight some type of registration correction between lanes is necessary. For example, in Fig. 2(a) for the leftmost four lanes the third lane is shifted downwards by around 14 pixels with respect to the first lane near the bottom of the image and by around 19 pixels near the top of the image. This misregistration between lanes is evident in Fig. 5 where for example the peaks in lane one and three corresponding to a G around pixel 90 do *not* coincide. Similarly, the bands in lane four are shifted downwards by around 14 pixels with respect to lane two in the lower portion of the image and around 24 pixels in the upper portion of the image. So the lanes need to be aligned with respect to each other before the classification stage.

Suppose locally the geometric distortions can be represented by a bilinear transformation. Let P_1, P_2, P_3, and P_4 be a set of control points representing the local warping as shown in Fig. 6. If we wanted to produce an image with straightened tracks and bands then we would transform the quadrilateral to a rectangle. A bilinear transformation of each coordinate accomplishes this. The parameters of the transformation can be determined by solving two systems of four linear equations each. Once the transformation is available then the output image can be determined quite efficiently (that is determining which pixel(s) of the input image fills a given pixel of the rectangular output image), requiring just two additions per pixel. However, obtaining a straightened image is not necessary for aligning the lane profiles with respect to each other.

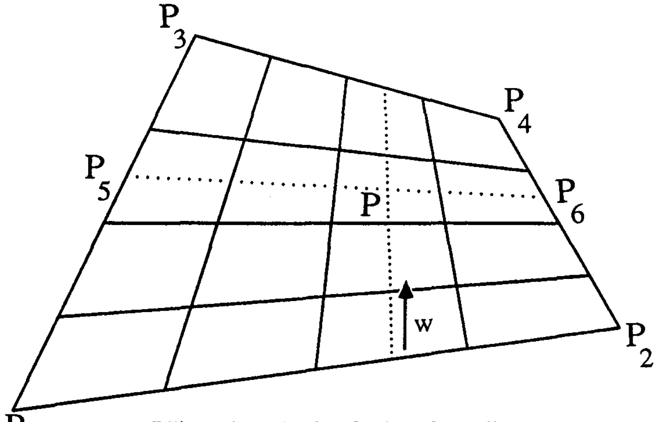

Figure 6 Bilinear interploation for inter-lane alignment

Given the control point coordinates the coordinates of P in the relative coordinate system which is defined as the fractional distance along opposite sides of the quadrilateral are desired. The relative coordinate w is the one required for alignment; the profiles are defined along the center-line of each lane and for inter-lane alignment the profiles need to be translated within the control quadrilateral to a common line. In order to determine w we introduce points P_5 and P_6 then,

$$\begin{aligned} P_5 &= w\,(P_3 - P_1) + P_1 \\ P_6 &= w\,(P_4 - P_2) + P_2 \end{aligned} \tag{3.9}$$

Using (3.9) and the expression for the slope of line P_5P_6 results in the following quadratic equation for w

$$a_1 w^2 + a_2 w + a_3 = 0 \tag{3.10}$$

where

$$\begin{aligned} a_1 &= ((x_3 - x_1)(y_4 - y_2) - (x_4 - x_2)(y_3 - y_1)) \\ a_2 &= (x_p - x_2)(y_3 - y_1) - (x_4 - x_2)(y_1 - y_p) - \\ &\quad (x_3 - x_1)(y_p - y_2) + (x_p - x_1)(y_2 - y_4) \\ a_3 &= ((x_p - x_2)(y_1 - y_p) - (x_p - x_1)(y_2 - y_p)) \end{aligned}$$

with the coordinates of P_i being (x_i, y_i) and of P being (x_p, y_p). The desired root for $w \in [0,1]$. Knowing w (for each point in the profile) then all the points of a profile can be translated to any other location in the direction defined by P_5P_6 simply by scaling each of the values of w. Suppose the line P_5P_6 is translated to $P_5'P_6'$ then each value of w in the original line is scaled by the factor $||P_5'P_6'|| \,/\, ||P_5P_6||$ along the new line. In this manner each of the four profiles can be translated to a common line (or aligned with respect one of the profiles) in order to achieve inter-lane alignment.

The difficult part of the alignment problem is to locate the control points. Since lane 1 and 3 share common bands for each G and lane 2 and 4 share common bands for each C each of the two lanes can be registered with respect to its companion using correlation methods. The result would be *two* control quadrilaterals which need to be merged into one set of four control points. Although the warping may not always be locally bilinear it is expected to be continuous and smooth. So in the process of merging control quadrilaterals the continuity of slopes can be used as constraints. When a bilinear interpolant does not allow for continuity of slopes then a higher order polynomial curve may be used as the interpolating function instead.

Once the profiles are aligned then classification can be performed using the features previously extracted to determine the ordered sequence. Currently we are improving the performance of a combined statistical and rule-based approach to the classification task.

Acknowledgements

The original images were digitized and provided by George Church at the Harvard Medical School. Discussions about DNA sequencing methods and difficulties with Carl Woese and Jordan Konisky of the microbiology department at the Univ. of Illinois at Urbana-Champaign were very helpful.

References

[1] C. DeLisi, "Computers in molecular biology: Current applications and emerging trends," *Science*, vol. 240, no. 4848, pp. 47–52, 1988.

[2] D. Soll, R. L. Kirschstein, L. Philipson, and H. Uchida, "DNA databases monitored," *Science*, vol. 240, p. 375, 1988.

[3] A. Wada, "Automated high-speed DNA sequencing," *Nature (London)*, vol. 325, pp. 771–772, 1987.

[4] J. West, "Automated sequence reading and analysis," *Nucleic Acids Research*, vol. 16, no. 5, pp. 1847–1856, 1988.

[5] U. Landegren, R. Kaiser, C. T. Caskey, and L. Hood, "DNA diagnostics – Molecular techniques and automation," *Science*, vol. 242, pp. 229–237, 1988.

[6] G. M. Church and S. Kieffer-Higgins, "Mulitplex DNA sequencing," *Science*, vol. 240, pp. 185–188, 1988.

[7] G. M. Church, "Harvard Medical School and Howard Hughes Medical Inst.." Personnel communication.

[8] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis," *Nature (London)*, vol. 321, pp. 674–679, 1986.

[9] W. Ansorge, B. Sproat, J. Stegemann, C. Schwager, and M. Zenke, "Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis," *Nucleic Acids Research*, vol. 15, no. 11, pp. 4593–4602, 1987.

[10] J. A. Brumbaugh, L. R. Middendorf, D. L. Grone, and J. Ruth, "Continuous on-line DNA sequencing using oligodeoxy nucleotide primers with multiple fluorophores," *Proc. National Academy of Sciences USA*, vol. 85, no. 15, pp. 5610–5614, 1988.

[11] L. M. Smith, R. J. Kaiser, J. Z. Sanders, and L. E. Hood, "The synthesis and use of fluorescent oligonucleotides in DNA sequence analysis," in *Methods in Enzymology* (R. Wu, ed.), vol. 155 Recombinant DNA Part F, ch. 19, pp. 260–301, New York: Academic Press, 1987.

[12] D. Q. Xu, M. K. S. Tso, and W. J. Martin, "Automatic interpretation of digital autoradiograph of DNA sequencing gels," in *Image Analysis and Processing II* (V. Cantoni, V. D. Gesu, and S. Levialdi, eds.), New York: Plenum Press, 1988.

[13] R. W. Conners, C. A. Harlow, and S. J. Dwyer, III, "Radiographic image analysis: Past and present," in *Sixth Int'l Conf. on Pattern Recognition*, pp. 1152–1169, 1982.

[14] R. C. Mann, B. K. Mansfield, and J. K. Selkirk, "Automated analysis of digital images generated by two dimensional gel electrophoresis," in *Pattern Recognition in Practice II* (E. S. Gelsema and L. N. Kanal, eds.), New York: Elsevier Science Publisher B. V., 1986.

[15] M. M. Skolnick, "Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological materials," *Comput. Vision Graph. Image Proc.*, vol. 35, pp. 306–332, Sept. 1986.

[16] G. Vernazza, S. B. Serpico, D. Guisto, and A. Caredda, "Computerized analysis of two-dimensional electrophoresis images," in *Medical Imaging* (P. Suetens and I. T. Young, eds.), vol. 593, pp. 154–162, Proc. SPIE, 1986.

[17] S. Nyberg, "Datoranalys av elecktroforesbilder for DNA-molekyler," Tech. Rep. D 30363-E1, National Defence Research Institute, Linkoping, Sweden, Dec. 1984. In Swedish.

[18] J. K. Elder and E. M. Southern, "Automatic reading of DNA sequencing gel autoradiographs," in *Nucleic Acid and Protein Sequence Analysis* (M. J. Bishop and C. J. Rawlings, eds.), (Oxford), pp. 219–229, IRL Press, 1987.

[19] T. P. Keenan and S. A. Krawetz, "Computer video acquisition and analysis system for biological data," *Computer Applications in the Biosciences*, vol. 4, pp. 203–210, Mar. 1988.

[20] P. Maragos, "Optimal morphological approaches to image matching and object detection," in *Second Intl. Conf. on Computer Vision*, pp. 695–699, Dec. 1988.

[21] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[22] D. Blostein and N. Ahuja, "A multiscale region detector," *Comput. Vision Graph. Image Proc.*, vol. 45, pp. 22–41, Jan. 1989.

[23] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley, 1973.