Image Analysis For DNA Sequencing

K. Palaniappan *T. S. Huang*
Coordinated Science Laboratory &
Beckman Institute, University of Illinois
405 North Mathews Avenue
Urbana, Illinois 61801 U.S.A.

Abstract

There is a great deal of interest in automating the process of DNA (deoxyribonucleic acid) sequencing to support the analysis of genomic DNA such as the Human and Mouse Genome projects. In one class of gel-based sequencing protocols autoradiograph images are generated in the final step and usually require manual interpretation to reconstruct the DNA sequence represented by the image. The need to handle a large volume of sequence information necessitates automation of the manual autoradiograph reading step through image analysis in order to reduce the length of time required to obtain sequence data and reduce transcription errors. Various adaptive image enhancement, segmentation and alignment methods were applied to autoradiograph images. The methods are adaptive to the local characteristics of the image such as noise, background signal, or presence of edges. Once the two-dimensional data is converted to a set of aligned one-dimensional profiles waveform analysis is used to determine the location of each band which represents one nucleotide in the sequence. Different classification strategies including a rule-based approach are investigated to map the profile signals, augmented with the original two-dimensional image data as necessary, to textual DNA sequence information.

1. Introduction

Automated DNA (deoxyribonucleic acid) sequencing involves computer interpretation of the chemical detection data which may be in the form of two-dimensional (2-D) autoradiograph images and one- or two-dimensional fluorescence data [1]. Since the 1970s there has been an exponential increase in the number of nucleotides that have been sequenced each year; consequently, collaborative efforts have been initiated among the European, Japanese and United States DNA Databanks to manage the expected tremendous increase in sequence data that will result from systematic genomic mapping programs [2]. The current release of GenBank® contains about 35 million nucleotides and is being reorganized as a relational database to handle the complexity of sequence annotation and the increased rate at which data are being generated by advanced sequencing methods [3]. The amount of data expected by the year 2005 could be as high as 7 billion of which half would be the complete human genome (currently only about 0.02% of the human genome has been sequenced) [4].

An accelerated rate of sequencing over the next two decades assumes major improvements in technology [4] and, certainly, automation of the sequence reading stage will play a crucial role in increasing the sequencing volume [5]. Automated sequencing techniques should also reduce the cost per sequenced nucleotide. A great deal of effort has also been invested in analyzing, classifying, and comparing nucleic acid sequence data in order to elucidate genetic, structural and functional properties [6],[7].

One key step in automated DNA sequencing involves computer conversion of the *chemical detection* data, which may be in the form of two-dimensional autoradiograph images, and one- or two-dimensional fluorescence data to an ordered nucleotide sequence representation. Several commercial automatic film readers with varying accuracy and speed of reading have been developed [8],[9],[10]. Approaches for analyzing two-dimensional autoradiograph images with consideration for

® GenBank is a registered trademark of the U.S. Department of Health and Human Services.

 0-8194-0549-3/91/$4.00

reliability and computational requirements are discussed. Some of the approaches discussed are not only applicable to autoradiograph data but also to one- and two-dimensional fluorescence data.

The experimental protocol used to generate the autoradiographs is discussed and terminology is introduced. Church and Kieffer-Higgins introduced *multiplex* DNA sequencing [11] 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 related to the size of the DNA fragment) 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. An ideal representation of typical sequencing gels is shown in Figure 1 and is used to illustrate the terminology. Each probing produces autoradiographs such as those shown in Figures 2 and 3 depending upong the complementary probe sequence used to bind to the gel membrane; Figure 2 shows the standard (that is the sequence is known beforehand) and Figure 3 a corresponding probe autoradiograph. Distinct dark *bands*, which are equivalent to resolving a single nucleotide of the original DNA fragment sequence can be seen. The column location of the band (or corresponding bands in some cases such as for G and C in the right half of Figure 1) 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, which are approximately uniform in width, line up to form a *lane* or *track* that corresponds to one of the reaction products. Typically, each group of four lanes contains all the information necessary 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, the nonuniformity in band shape, size or spacing, and the shifting in alignment between lanes necessitate sophistication in automatic reading algorithms. Additional problems include missing lanes (and very low contrast regions) due to a poor binding of the probe to the membrane as in Figure 3.

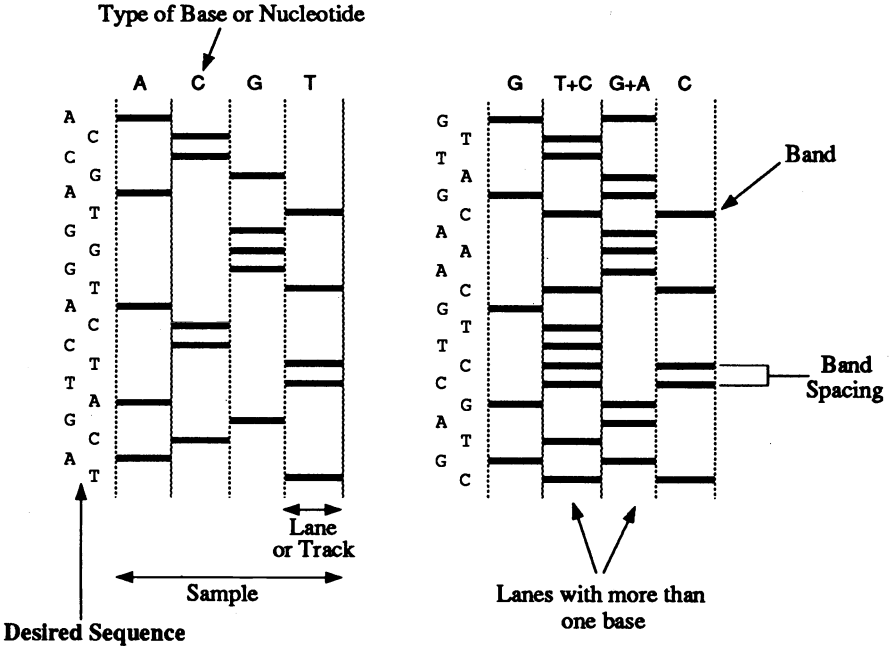


Figure 1 Ideal representation of typical DNA autoradiograph images.

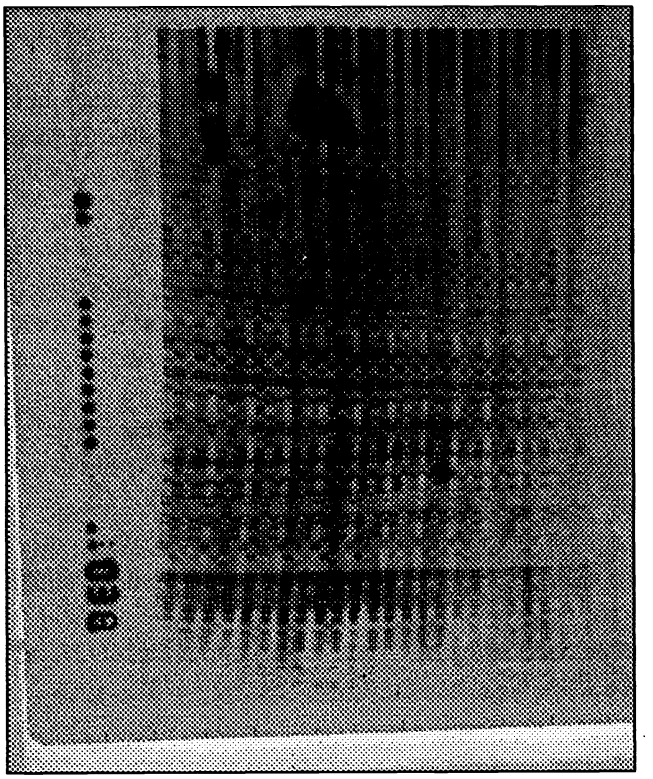


Figure 2 Entire digitized autoradiograph of a standard membrane produced using the multiplex sequencing method showing 12 sets of identical sequences (48 lanes). (Original data provided by Dr. George M. Church, Dept. of Genetics, Harvard Medical School, Boston, MA)

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 autoradiograph; thus, the speedup over conventional approaches is proportional to the amount of multiplexing [11]. 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 autoradiographs. 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. Processing such a large volume of data will inevitably require robust algorithms to read, assemble and analyze autoradiographs.

The multiplex sequencing method operates in a batch style in the sense that the complete autoradiograph must be developed before the sequence can be read. Continuous, on-line sequencing systems that do not require radioisotopes and autoradiograph recording have been developed using fluorescence-based detection. Fluorescence-based methods may produce either one-dimensional traces for each nucleotide [8] or two-dimensional images [12] resembling autoradiographs and may be based on a single-dye, four-lane sequencing format or a four-dye, single-lane format. In the fluorescence techniques, the vertical axis of separation is time rather than space. 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 especially if several fluorescins are used, slower scanning, less reliability, and less flexibility as well as higher cost [10]. One particularly difficult problem is that as the DNA fragments increase in length they pass the detector more slowly and the bands become wider but the distance between bands remains constant. A novel method that uses a multiwire proportional counter (MPWC) to reduce the exposure time required to detect radioactivity

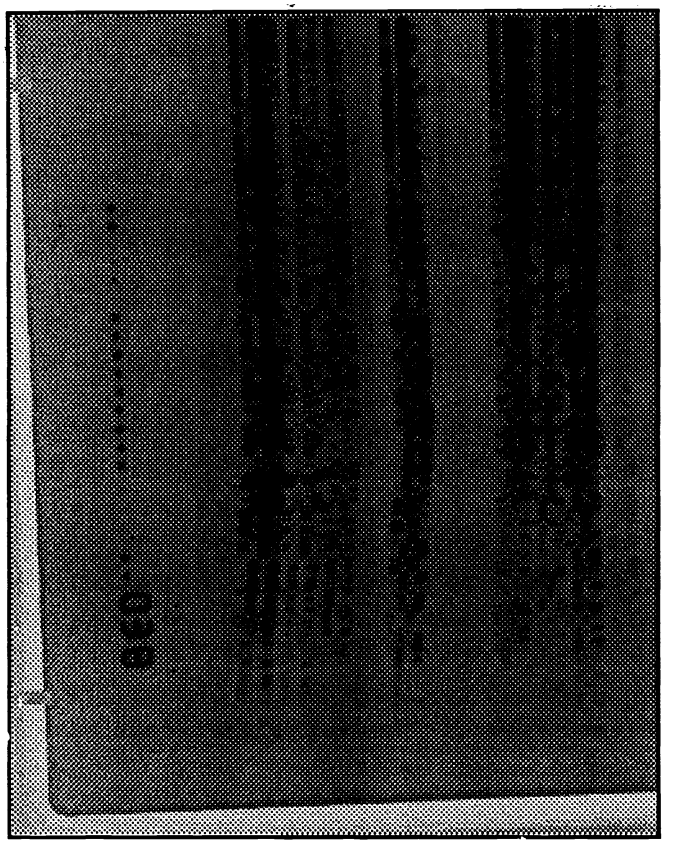


Figure 3 A probe autoradiograph produced from the same gel membrane as the standard shown in Figure 2.

and form an image along with algorithms for automatically interpreting the coarser MPWC images to determine the sequence is described in [13]. The algorithms developed have been applied only to the analysis of autoradiograph images resulting from the multiplex DNA sequencing technique.

2. Analysis of Autoradiograph Images

Biomedical radiograph images such as two-dimensional gel electrophoresis autoradiograph images (usually of protein materials) have been investigated primarily in the 1980s [14],[15]. DNA autoradiograph images of interest to us, however, have been analyzed by computer only recently [16],[17],[18],[19],[10].

Rather than taking a two-dimensional approach involving boundary and region detection, shape description, etc., the autoradiograph image is converted to a set of one-dimensional signals which are then used to determine the DNA sequence. The reduction to one dimension is possible due to the underlying nature of the data and the classification task which is to recover a linearly 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 2 shows a complete autoradiograph 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, whereas for the displayed image the gray level range has been rescaled to one byte per pixel. Since the membrane is 43 cm × 35 cm the sampling rate is approximately 116 μm (microns) in the vertical direction and 241 μm in the horizontal direction (or 453 dots/in × 218 dots/in). 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 μm sampling rate would require about 120 Mb per

image or equivalently, 459 video frames. There are 48 lanes in Figure 2 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 4 shows a histogram for a portion of the original image. The range of gray levels in this region is limited to approximately half of the total 256 gray levels available. Contrast enhancement algorithms can improve the appearance of the image such as local intensity rescaling. The histogram reflects the gray scale rescaling and also shows that setting a threshold (even a locally adaptive one) for isolating the bands from the background would be difficult. Thresholding usually leads to incomplete, missing or merged bands. One of the reasons for this is the variation in the 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, whereas, for faint bands, it can be as low as 1.03, almost indistinguishable from the background. *Companion* bands are also highly nonuniform in intensity which makes their detection and classification an even more difficult task; companion bands are bands that are appear in roughly the same horizontal position but are present in two or more lanes. For example, in the chemistry protocol used to generate the autoradiograph of Figure 2, bands in the first lane indicate the presence and position of guanine (G) bases, in the second lane they indicate 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; thus, 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. Because the dynamic range of the image intensity for companion bands ranges from 2 to 0.25, faint bands can be easily missed using simple thresholding or edge detection operators. In fact, many edge operators gave unsatisfactory results due to spurious edge responses, missing edge boundaries, merged edges, and disconnected or shifted edge contours. The edge operators that were tried included 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 a Gaussian, maxima in the output of an oriented first derivative of a Gaussian operator proposed by Canny, or the oriented masks of Nevatia-Babu. These difficulties in edge detection would need to be overcome using more sophisticated postprocessing algorithms for linking, grouping and classifying edges. Consequently, neither the region detection-based (using thresholding) nor boundary detection-based (using edge detection) approaches were strictly followed.

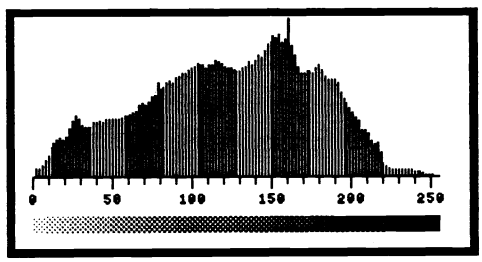


Figure 4 Histogram for a portion of the DNA autoradiograph shown in Figure 2 after gray level rescaling (with bands being light in color).

3. Image Analysis Methods

The initial steps are 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 are shown. This first stage also involves film registration to accommodate the placement of the film with respect to the scanner and extraction of identification information (such as the numbers shown in Figure 2 or possibly bar codes) to keep track of the image and the sequence in large sequencing projects.

The outputs of different image analysis modules include the digitized and enhanced image, registration information, lane geometry and distortion parameters. The input requirements 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, they can be used to speed up the reading of the probe sequences (up to 40 autoradiographs or more) and do not have to be recalculated each time as in the usual sequencing approach where an internal standard is not available.

New techniques for image segmentation, obtaining profile features, correcting nonlinear geometric distortions, and an evaluation of pattern classification using a statistical and rule-based approach are the major contributions. Some initial results 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) interlane alignment, and (f) classification-based sequence construction are discussed.

3.1. Segmentation into Lanes

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 3×3 vertical edge operators were tried as well as the 2×2 Roberts difference operator. The 2×2 operator gave many fewer responses than the 3×3 operators (including Prewitt, Sobel, Frei-Chen) all of which behaved quite similarly. Figure 5 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 7×3, 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

$$\widehat{\rho} = \frac{\sum_{\mathbf{n}\in W} [g(\mathbf{n}) - \overline{g}(\mathbf{n})][g(\mathbf{n}+\mathbf{k}) - \overline{g}(\mathbf{n}+\mathbf{k})]}{\left(\sum_{\mathbf{n}\in W} [g(\mathbf{n}) - \overline{g}(\mathbf{n})]^2 \sum_{\mathbf{n}\in W} [g(\mathbf{n}+\mathbf{k}) - \overline{g}(\mathbf{n}+\mathbf{k})]^2 \right)^{1/2}} \tag{3.1}$$

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 columns 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*,

$$z = \frac{1}{2}\log_e\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right) \tag{3.2}$$

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

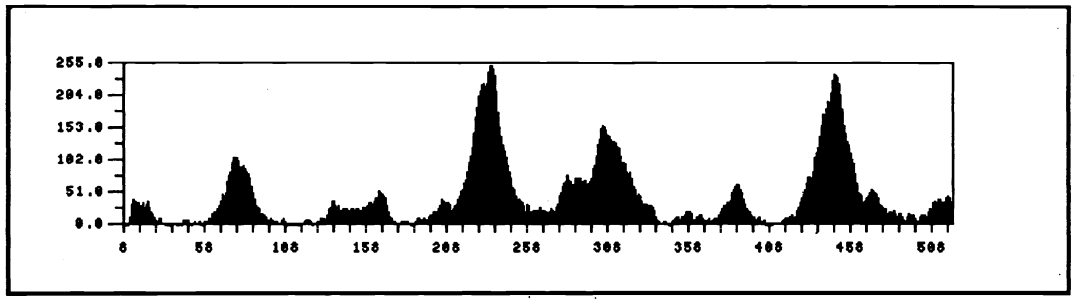


Figure 5 Projection onto the horizontal axis of a small region, from the image in Figure 2 after processing using a 3×3 Prewitt edge operator for vertical edges. Although every lane is not accurately localized, all are detectable.

The size of the window to use for the x-projection needs to be locally adaptive. Using a constant size window gives unsatisfactory results since some regions of the image are more highly distorted than other regions, and some regions may have a scarcity of detail which results in poor localization. Adjusting the size of the window, using a heuristic which measures band density so that each nonoverlapping window contains approximately the same number of bands, automatically maintains a given level of localization performance and follows distortions. Since the window size is changing, the size of the edge operator must also be locally modified. Consequently, edges at varying scales are being detected in different regions of the image.

3.2. One-Dimensional Lane Profiles

Some of the methods proposed to date for analyzing DNA autoradiographs 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 nonhorizontal in order to obtain a high-resolution profile that reduces the effects of noise. Rather than using the conventional correlation function $f_2(\mathbf{k}) = \sum_{\mathbf{n}\in W} g(\mathbf{n})g(\mathbf{n}+\mathbf{k})$ which requires multiplications and is sensitive to brightness changes across the image, the morphological correlation is used,

$$\mathrm{M}(\mathbf{k}) \equiv f_3(\mathbf{k}) = \sum_{\mathbf{n}\in W} \min(g(\mathbf{n}+\mathbf{k}), g(\mathbf{n})) \tag{3.3}$$

Maximizing $\mathrm{M}(\mathbf{k})$ can be shown to be equivalent to minimizing

$$f_1^{(1)}(\mathbf{k}) = \sum_{\mathbf{n}\in W} |g(\mathbf{n}) - g(\mathbf{n}+\mathbf{k})| \tag{3.4}$$

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; however, this was not found to be necessary as the improvement in performance was only marginal. Several advantages to using $\mathrm{M}(\mathbf{k})$ are that it is fast, since each term requires only a comparison operation, and it performs almost as well as the correlation function $f_3(\mathbf{k})$ in terms of the behavior of the mean and variance as the SNR is increased; in fact, the mode for the morphological correlation tended to be slightly sharper than for the correlation function. 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. In addition, the number of shifts evaluated need not be restricted to a fixed set, S, but can be made dependent on the local value of the criterion. The steepest descent algorithm for image matching as described in [21] has reasonable computational cost and converges with high probablility. A modified steepest descent scheme that searches for a local maximum of the criterion function in an incremental stepwise fashion

allowing for local maxima within a search window is given below. The algorithm assumes that a reference column of the image $g(\mathbf{n})$ is being matched to a neighboring column.

Modified Steepest Ascent Image Profile Matching Algorithm

START: Set $MATCH \leftarrow TRUE$. Set τ_m the matching threshold. Set τ_s the window for the search neighborhood. Select the reference profile $g(\mathbf{n})$ and k_2 (the profile or column to be matched). Evaluate an initial match value, $M(k_1, k_2)$ using say $(k_1, k_2) \leftarrow (0, 1)$.

LOOP: Evaluate $M(k_1^{(s)}, k_2)$ for $k_1^{(s)} \in \tau_s$ and select the maximum $M(k_1^{(max)}, k_2)$
If $M(k_1^{(max)}, k_2) > M(k_1, k_2)$ then Go To **UPDATE**.
Otherwise Go To **TEST**.

UPDATE: Set $k_1 \leftarrow k_1^{(m)}$. Go To **LOOP**.

TEST: If $M(k_1^{(m)}, k_2) < \tau_m$ then set $MATCH \leftarrow FALSE$.
Exit.

Determining the displacement k with respect to the center of the lane is usually sufficient to follow gradually sloping bands. However, this can lead to severe problems with impulsive noise or *spots* close to the center of the lane which can then manifest themselves as artifact bands in the center profile. Choosing a *median* profile based on several columns near the center of the lane is usually sufficient to overcome this problem. Figure 6 shows the profiles for the first four lanes for part of the image in Figure 2 (the peaks correspond to the dark bands).

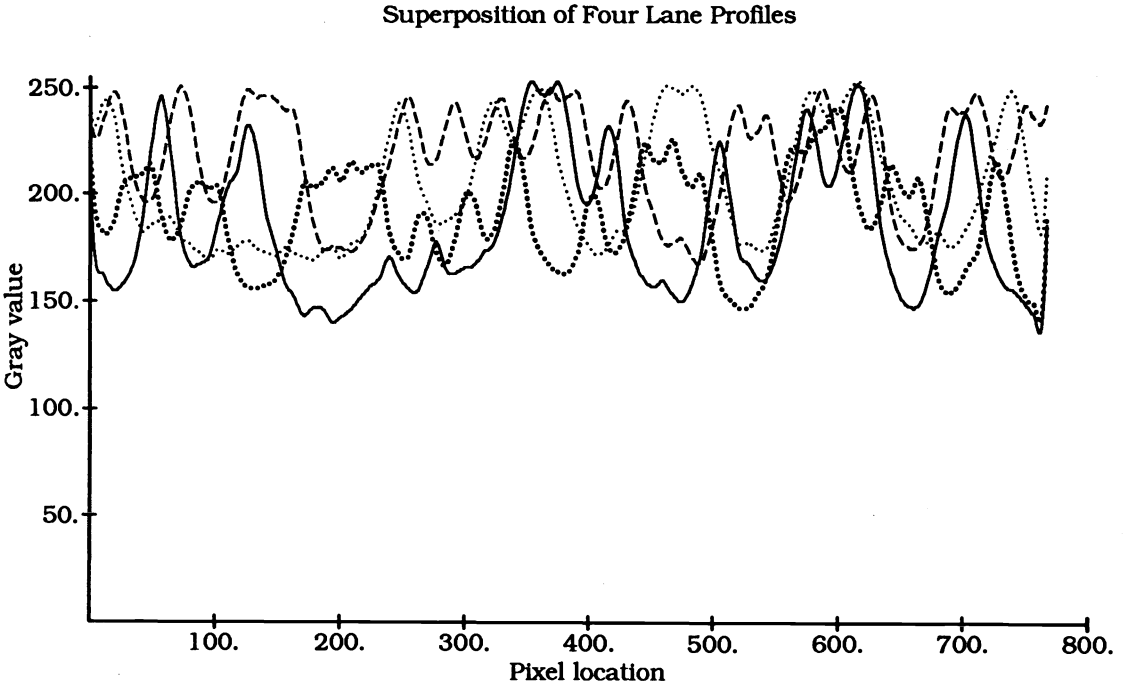


Figure 6 Profiles for a portion of the image shown in Figure 2. The solid and dashed lines are Lanes 1 (G) and 3 (G+A), respectively, and the thick and thin dotted lines are Lane 2 (C+T) and 4 (C), respectively. The profiles are plotted prior to interlane alignment.

3.3. Detecting Individual Bands

Individual bands are modeled as Gaussian-shaped peaks in the one-dimensional signal since this provides a simple description of the profile that leads to tractable analysis and, with three adjustable parameters, can be fit to a variety of peaks in real data. A multiscale approach is used to automatically detect peaks of different sizes and does not require knowledge about the total number of peaks nor an initial segmentation. The multiscale approach is approximately equivalent to bandpass filtering the image at several frequency ranges and combining the filter outputs algorithmically; note that the term *image* will be used interchangeably to refer to either 1- or 2-dimensional signals as appropriate. Multiscale approaches have been applied to a variety of image understanding problems including feature detection, curve analysis, region interpolation, texture segmentation and object recognition [22].

The peak detection algorithm involves four steps. First, the image is convolved with a series of Laplacian of a Gaussian filters, $\nabla^2 G_\sigma * I$ (which is just the second derivative of the Gaussian in 1-D, that is $\frac{\partial^2 G_\sigma}{\partial x^2} * I$) over a range of σ that depends on the size of the peaks to be detected. This generates a multiple scale representation of the original image with the scale parameter being σ. Depending upon the size of the mask and image as well as the computer or special purpose architecture involved, it may be more efficient to implement this set of filtering operations in either the spatial domain or the frequency domain. Second, local maxima (minima) in the convolution output for each filter size are extracted as candidate peak (valley) locations. Third, each maximum (minimum) marked in Step 2 is convolved with a filter that is the variation of the scale of the Laplacian of a Gaussian, $\frac{\partial}{\partial \sigma} \nabla^2 G_\sigma * I$. The result of this convolution combined with the results of the first step are used to estimate the parameters of a Gaussian-shaped peak at each candidate peak point. Finally, the results of the previous step are used as the initialization for an iterative refinement of the detected peaks and their associated parameters. A theoretical justification for the algorithm is given.

Extension of the above approach to general two-dimensional images is not straightforward when ridges/valleys as well as peaks/pits in the gray level surface need to be detected because the *orientation* and size of the ridge/valley also needs to be determined. A peaked surface has negative mean curvature and positive Gaussian curvature whereas a ridge has zero Gaussian curvature and negative mean curvature. Similarly a pit or cupped surface has positive mean and Gaussian curvatures and a valley surface has zero Gaussian curvature and positive mean curvature. The main point is that for ridges and valleys (of the $\cup$ and $\cap$ type, respectively) one of the principal curvatures is zero. There are four other surface types (plane, saddle ridge, saddle valley and minimal) but these are not considered in this thesis. Peak and pit detection in 2-D can be directly generalized from the 1-D results. However, detecting ridges and valleys requires a more local approach. The axis along which the projection profile appears Gaussian-shaped is defined to be orthogonal to the *ridge/valley orientation* axis (the orientation axis is unique only up to the addition of an integer multiple of π). For the biomedical images we are using we usually have a priori information about the orientation of the ridges/valleys which is normally along the horizontal axis, as a result the 1-D profile shows Gaussian-shaped peaks along a vertical projection axis. When the ridge/valley orientation information is available, then the four steps of the algorithm may be applied locally to determine 1-D peaks/pits in the direction orthogonal to the orientation axis, to extract the extent, location and other features of the ridges/valleys. Alternatively, the ridge/valley orientation could be estimated using moment-based methods that determine the axis of symmetry or by an iterative search about an initial direction determined by a local planar fit. The locally detected properties can be used in a grouping step to form 2-D peak, pit, ridge and valley-shaped regions.

The peak detection algorithm was motivated by a similar approach for extracting homogeneous image regions based on fitting constant intensity disks of various sizes for constructing texture elements [23]. In the analysis of DNA autoradiograph images, peak regions need to be detected with accurate localization so that a Gaussian peak model has been used rather than a disk or bar as in [23]. In [24] a multiscale approach for analyzing histograms was presented and though a Gaussian model was used, the analysis was primarily concerned with fingerprints of 1-D signals, that is, zero-crossings of $\frac{\partial^2 G_\sigma}{\partial x^2} * I$ as σ is varied; we use information about extrema in the variation with respect to σ, $\frac{\partial}{\partial \sigma} \frac{\partial^2}{\partial x^2} G_\sigma * I$. The filters used for extracting Gaussian peaks have been theoretically shown to exhibit a number of desirable properties.

The peaks (which correspond to the dark bands) in Figure 6 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, $\frac{\partial^2 G_\sigma}{\partial x^2}$, 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 all the maxima in each $\frac{\partial^2 G_\sigma}{\partial x^2} * I$ image where the dark bands in the image correspond to peaks in the profile.
3. For each maximum marked in Step 2, calculate $\frac{\partial}{\partial \sigma}\frac{\partial^2 G_\sigma}{\partial x^2} * I$ from which the peak's scale size b can be estimated as

$$b = \sigma\left(\frac{3}{1-\sigma B_\sigma} - 1\right)^{\frac{1}{2}}, \quad B_\sigma = \frac{\frac{\partial}{\partial \sigma}\frac{\partial^2}{\partial x^2} G_\sigma * I|_{\text{local maxima}}}{\frac{\partial^2}{\partial x^2} G_\sigma * I|_{\text{local maxima}}} \tag{3.5}$$

Several other expressions for estimating b can also be obtained. The peak strength A can be estimated as

$$A = \frac{\left(b^2+\sigma^2\right)^{3/2}\frac{\partial^2 G_\sigma}{\partial x^2} * I|_{\text{local maxima}}}{\sqrt{2\pi}\, b\, \sigma} \tag{3.6}$$

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

Methods for dealing with several merged peaks which then appear as a plateau in the profile, accurately resolving shoulders of peaks and extending the approach to deal with general 2-D data, need further investigation. Based on the initial estimates a maximum likelihood iterative updating scheme that is also known as the Expectation-Maximization (EM) algorithm is used to refine the parameter estimates.

3.4. Feature Extraction and Classification

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, and faint closely spaced bands that often appear as shoulders around a peak, can all be used in the classification stage.

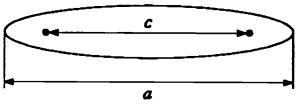
A set of forty features has been defined and some of them have been used in the classification step. For example, a set of local features describing properties of the shape and contrast of bands is given below. Other features extract information about band characteristics within and between tracks and a set of reliability indices for identifying problem areas.

- m_1 Peak location (centroid)
- m_2 Peak amplitude
- m_3 Variance in peak strength within band region
- m_4 Peak width of current band (near center of lane)
- m_5 Band major axis length
- m_6 Band minor axis length
- m_7 Band orientation with respect to x-axis in degrees

m_8 Eccentricity or elongatedness

$$m_8 = \frac{c}{a} \tag{3.7}$$

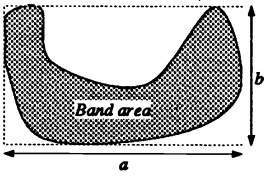
Note that m_8 is zero for a circle and one for a line.



m_9 Band curvature which is zero for an ellipse and increases for *smiling* or *frowning* bands

m_{10} Irregularity with respect to an ellipitical shape

$$m_{10} = \frac{\text{Band area}}{\pi ab/4} \tag{3.8}$$



Many of the features listed above need to be updated for each new observation. Rather than keep all the observations in memory to recalculate the feature we can use a sequential updating procedure to save memory requirements. For example, given a new observation the updating of the mean value can be done sequentially

$$\mathbf{m}_n = \frac{n-1}{n}\mathbf{m}_{n-1} + \frac{\mathbf{x}_n}{n} \tag{3.9}$$

The nth observation can be weighted in order to reduce the effect of variations due to local compressions and expansions in which case a weighted update can be similarly defined.

Rules were directly coded in software rather than using an expert system development environment. The results of an experiment for a very clean sequencing set with highly resolvable bands is given below. Classification rates of 91% (including ambiguities when no decision is made) can be achieved by a priori constraining the region of interest; that is not reading near the very top or bottom of images. The discarded region can be large; in the autoradiograph of Figure 2 the discarded data is equivalent to approximately 100 possible bases. On low contrast regions of the same image the performance drops to around 80%.

	T	C	A	G	N
T	40	1	1	0	2
C	2	43	1	0	1
A	1	1	49	2	1
G	0	1	0	34	2

Figure 7 Classification results for a portion of the autoradiograph shown in Figure 1 and for which profiles were also plotted in Figure 5. The letter N represents any base.

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] C. Burks et al., "Genbank: Current status and future directions," in *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences* (R. F. Doolittle, ed.), vol. 183 of *Methods in Enzymology*, pp. 3–22, New York: Academic Press, Inc., 1990.

[4] C. Burks, "How much sequence data the databanks will be processing in the near future," in *Biomolecular Data: A Resource in Transition* (R. R. Colwell, ed.), pp. 17–26, Oxford: Oxford University Press, 1989.

[5] B. Barrell and J. Sulston, "Gels to sequences," in *Computational Molecular Biology* (A. M. Lesk, ed.), pp. 131–136, Oxford: Oxford University Press, 1988.

[6] A. Lapedes, C. Barnes, C. Burks, R. Farber, and K. Sirotkin, "Application of neural networks and other machine learning algorithms to DNA sequence analysis," in *Computers and DNA* (G. Bell and T. Marr, eds.), vol. vii of *SFI Studies in the Sciences of Complexity*, pp. 157–182, New York: Addison-Wesley, 1990.

[7] M. S. Waterman, ed., *Mathematical Methods for Nucleic Acid Sequences*. Boca Raton, FL: CRC Press, 1989.

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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 of *Proc. SPIE*, pp. 154–162, 1986.

[16] R. S. Anbalagan, G. Hu, and A. K. Jain, "A segmentation and object extraction algorithm with linear memory and time constraints," in *Proc. Int. Conf. on Pattern Recognition*, pp. 596–600, Nov. 1988.

[17] 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.), ch. 9, pp. 219–229, Oxford: IRL Press, 1987.

[18] 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.

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

[20] P. Maragos, "Optimal morphological approaches to image matching and object detection," in *2nd Int. Conf. on Computer Vision*, pp. 695–699, Dec. 1988.

[21] A. K. Dewdney, "Analysis of a steepest-descent image-matching algorithm," *Pattern Recognition*, vol. 10, pp. 31–39, 1978.

[22] C. R. Dyer, "Multiscale image understanding," in *Parallel Computer Vision* (L. Uhr, ed.), pp. 171–213, New York: Academic, 1987.

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

[24] M. J. Carlotto, "Histogram analysis using a scale-space approach," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 121–129, Jan. 1987.