Indian Journal of Experimental Biology
Vol. 38, February 2000, pp. 167-176

Development of a computer software for analysis of SDS-PAGE protein fingerprints of bacterial isolates

S K Saxena[1], A N A Ibrahim[2], Santanu Chaudhury[1*] & S S Thukral[2]

[1]Department of Electrical Engineering, Indian Institute of Technology, New Delhi, 110 016

[2]Department of Microbiology, V P Chest Institute, University of Delhi, Delhi, 110 007

Received 1 April 1997; revised 15 October 1999

Protein fingerprinting is a widely used technique in epidemiological studies for typing bacterial strains. This study reports the development of a computer based gel analysis system. The system has the capability to analyse SDS-PAGE whole-cell protein profiles using digital image processing techniques. The software incorporates spatial and frequency domain operators for image enhancement, support for geometric correction of images and new algorithms for identification of strain tracks and protein bands. The system also provides facilities for correcting imaging defects for inter-gel comparison, similarity analysis, clustering and pictorial representation of results as a dendrogram. The software is highly interactive, user-friendly and can produce accurate results for differentiation of bacterial strains with minimal overhead of time.

Image processing and analysis techniques have found applications in various fields of life sciences like epidemiology, taxonomy, cytogenetics, molecular biology etc., for providing quick, accurate and precise understanding of structural features of living organisms at macroscopic as well as microscopic levels. In epidemiological studies, digital image processing can be applied for analysis of protein fingerprints of bacterial strains obtained by using sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) technique. These protein fingerprints give an indirect reflection of bacterial DNA and are used for typing bacterial isolates[1]. A gel contains tracks, representing the protein profiles of different bacterial strains and the electrophoretically separated proteins constitute different bands in each track.

Visual analysis of gels for identification and comparison of bands is, in general, inaccurate and depends on human expertise and efficiency. Manual calculations of similarity values and cluster analysis is cumbersome, technically demanding and time consuming. Further, visual analysis cannot cater for inter-gel variations. This makes the task of creating and maintaining a large database, difficult and complex. Hence, there is a need for an interactive, semi-automated software for processing and analysis of these gels that can produce standard, accurate and rapid results. A number of computer based analysis methods have been developed[2,3]. Other than basic image processing functions, techniques used in these systems for track and band identification are proprietary and are not easily discernible. Commercially available software is integrated with gel documentation systems which are exorbitantly priced. The software that is available without gel documentation systems is also highly priced. Hence there is an urgent need for indigenous development of such a software. This would enable development of know-how for meeting typical requirements of a gel analysis system.

We describe here an indigenously developed gel analysis software for analysing images of protein fingerprints of bacterial isolates. This software incorporates new algorithms for identification and detection of protein-tracks and bands in a gel. These algorithms are significant achievements of this endeavour. Similarity coefficients between tracks can be calculated using either Dice Coefficient or Pearson Moment Correlation Coefficient[10]. For the purpose of clustering, the well known unweighted pair group method using averages (UPGMA) algorithm[8] has been provided in the system. The software is X Windows based and can be run on a personal computer under Linux operating system (a public domain version of UNIX for 80x86 family based systems). This is a unique feature of this software because all other commercially available software packages are in Microsoft Window's environment. Since, LINUX is a public domain free software, software environment of the proposed gel analysis system will not entail any additional cost. In this

study, we describe the basic features of the software and its application for analysis of electrophoretic protein fingerprints of bacterial isolates.

Materials and Methods

System Design—The gel analysis system has been designed as an interactive tool. The processing cycle envisaged by the system includes human operator as the most vital component. The design, therefore, has been guided by the user's perception of application. User is expected to manipulate and process scanned images of protein profiles for making necessary observations and similarity calculations. Consequently, user has been provided with a menu based interaction facility for initiating the processing steps. Operations have been grouped into different classes for ease of reference and selection.

Interactive system is expected to provide a collection of tools which would help the user to effectively visualize gel images. To meet this requirement, the software includes a set of image enhancement routines which can eliminate noise and improve the quality of image. User can select appropriate combination of tools depending on the nature of the image. These sets of enhancement routines are also useful for the operator to pre-process the image in order to extract optimal performance from the track and band detection utilities. User has also been provided with facilities for retracing the processing steps for ensuring error recovery. This offers an option to the user for experimenting with different combination of processing tools for obtaining the best performance.

Core processing utility for band and track detection has been designed to operate in semi-automated mode in which user plays the pivotal role. User is expected to judge the performance of the processing routines and accordingly user can intervene if it is required. For this purpose the user has been provided with tools for visualization of grey level patterns in the image and facilities for correcting the output. This feature has been incorporated in the software because it has been found that performance of the image processing routines varies depending upon quality of the gel image. User interaction facility also includes an option for the experienced user to modify different parameters of the algorithms for tuning the system for different image acquisition hardware.

The present system incorporates tools for- (i) processing the gel image to improve its quality in terms of enhancing the band contrast and also removing the unwanted background; (ii) analysis of gel image which includes routines for identification of bacterial strain tracks and their straightening, detection of bands in each track and their comparison; and (iii) calculation of similarity values between protein patterns and cluster analysis for identification, classification and studying the relatedness of strains. The molecular weights of proteins in each track can also be calculated if a standard molecular weight marker track is included in the gel.

Preprocessing of gel image

Scanning of gel—Wet/Dried gel or their photographs are scanned on a flat bed HP Scanjet 4C plus scanner with a resolution of 100 or 300 dots per inch (dpi) so that bands that are 1 or 0.08 mm apart can be resolved. The gel is scanned into a digital image where each pixel is allocated 8 bits. Of the various file formats such as PC Paintbrush (PCX), TIFF, GIF, JPEG, etc., the software accepts the image in PCX file. The size of the image depends on the resolution selected. The desired portion of gel can be selected during scanning which would be subsequently processed and analysed. However, care is to be taken to mark the top of gel as image top as this has significance for determination of molecular weights of the bands. A typical gel of size $16 \times 14 \times 0.3$ cm (this is the standard size using the electrophoresis apparatus - LKB, Sweden) is stored as 500 x 400 pixel image. A portion of a typical scanned image is shown in Figure 2. Here, the gel image orientation is such that long axis of tracks are vertical i.e. along y axis and bands in a track are horizontal i.e. along x axis.

Enhancement of gel image—A gel image consists of banding patterns of varying optical densities along each track due to differences in the content of each protein. These bands can be enhanced to assist visual inspection. Depending upon the quality of input gel image, a set of image operators can be applied in spatial or frequency domain.

Image enhancement in spatial domain—These methods include procedures that are based on direct manipulation of pixel by an operator defined over a neighbourhood of eight surrounding pixel in the gel image[4]. Following operators have been incorporated in the software.

Contrast enhancement—For improving the contrast of an otherwise dull gel image, histogram equalization method is used to achieve a uniform histogram for output image. An image with too low or

too high illumination of objects has grey levels occupying only a portion of the grey scale. By using a transformation function equal to cumulative distribution of image histogram, an output image with greater dynamic range can be achieved[4].

Sharpening—An edge is a transition area where pixel brightness value changes quickly. Such changes occur at lane and band boundaries which may need accentuation. This is achieved using a Sobel operator.

Noise reduction using averaging and median filters—The high frequency noises are eliminated using low pass filters like averaging and median filters. In averaging filter, each pixel is replaced by the weighted average of its neighbourhood pixels. In median filter, the input pixel is replaced by the median of the pixels contained in a window of size 3x3 around the pixel.

Image enhancement in frequency domain—This is achieved by transforming the input gel image into frequency domain[5], multiplying the transformed image by filter function and taking the inverse transform to produce the enhanced image.

Low pass filter—For smoothening the sharp transitions, such as noise (represented by high frequency contents in frequency domain), high frequency components of the gel image can be attenuated in the transform domain using a low pass filter.

High pass filter—Gel image sharpening can be achieved using high pass filtering in which the low frequency components are attenuated without disturbing high frequency information in Fourier transform.

Gel analysis

After preprocessing, the enhanced gel image is subjected to analysis process for identification of tracks and bands corresponding to individual bacterial strains. A threshold based algorithm is used for detection of tracks. During preparation and photography of gel, tracks may get distorted. Linear and non-linear distortions of these tracks are removed using geometric correction technique. Subsequently, bands are detected in these tracks using a novel technique.

Track/lane identification—Track boundaries occur where there is significant transition in grey levels along the horizontal direction. Approximate track boundaries are first calculated by finding average pixel intensity along each column and discriminating columns corresponding to tracks on the basis of an appropriate threshold. For each set of consecutive columns labelled as track, a medial axis is identified. Then tracks are straightened (explained later) using geometric correction method that includes spatial transformations and gray level interpolation. Exact track boundaries are, now, decided by taking a fixed deviation from the identified medial axis of the corrected lanes as these are expected to be of fixed width because a specific plastic comb is always used during gel preparation. Rest of the background is then removed to achieve a clear gel image with straightened lanes.

Geometric correction of tracks—Elimination of geometric distortions involves estimation of parameters of a geometric distortion model and rectification of image using interpolation.

Geometric transformation—This is achieved by spatial relocation of pixels using tie points. The locations of these pixels (tie points) in distorted and corrected gel image are known precisely. Geometric distortion can be modelled by the following pair of bilinear equations -

$$x_c = c_1 + c_2x + c_3y + c_4xy \quad \ldots (1)$$

$$y_c = c_5 + c_6x + c_7y + c_8xy \quad \ldots (2)$$

where x, y are pixel coordinates in distorted gel image and x_c, y_c are pixel coordinates in corrected gel image. Eight coefficients $c_i = 1, 2, 3, \ldots 8$ are solved with eight known tie points using Gauss-Jordan algorithm for solving simultaneous equations[4].

Gray level interpolation—Geometric transformation maps points to non-integer coordinates. Since, distorted gel image is digital and has pixel values defined only at integer coordinates, gray level values at the corrected coordinates are set to the intensity value obtained as weighted average of four surrounding pixels.

Segmentation of bands—An accurate identification of bands in each bacterial strain track is essential as this forms the basis for similarity analysis. Two different techniques have been incorporated in the system for the purpose of band detection. In the first technique the protein absorbance profile is used. The protein absorbance profile of each lane is plotted by calculating the average pixel intensity at each scan line along the track from the beginning of track boundary to the end of track boundary. The positions of bands are determined by locating the peaks of the absorbance profile because bands correspond to higher grey level than remaining parts of a track. For

this purpose, a peak detection filter is applied to the absorbance profile. For generating the peak detection signal, the absorbance profile is first smoothed and then differentiated. In Fig. 1 we have shown an example peak detection signal generated from an absorbance profile. A zero crossing of the detection signal to negative values, denoted by s_i indicates start of the peak. A zero-crossing of the detection signal to positive following a negative cross-over, denoted by m_i indicates the local maximum of the absorbance profile, i.e. actual location of the peak. The location between two successive negative crossovers at which the detection signal attains its local maximum, denoted by e_i, indicates end of the peak region in the the absorbance profile. In the second technique, a band is identified at both the boundaries using the gradient change along the track. Suitable positive and negative gradients are selected. A band begins when there is a transition from lighter region to a darker region and gradient exceeds the positive gradient threshold. Similarly a band ends when transition takes place from a darker to a lighter region. Thus, all the bands in a track are selected and the boundaries are marked at the valley points of the profile.

Manual selection, deletion and aligning of bands— An interactive facility has been incorporated to add or delete bands in any particular lane through visual comparison between the tracks and their respective protein absorbance profile placed adjacent to each track. This process enables manual intervention if desired, before finally subjecting the image to similarity analysis.

Similarity analysis—Similarity analysis is the process of calculating the similarity or correlation coefficient between two protein patterns based on comparison of their respective bands:

Dice coefficient—The software incorporates routines for similarity calculation on the basis of Dice coefficient[10]. Similarity calculation is done on the basis of band positions only. Grey levels have not been considered because the intensity levels are not always determined by the amount of protein only and can vary from sample to sample and gel to gel.

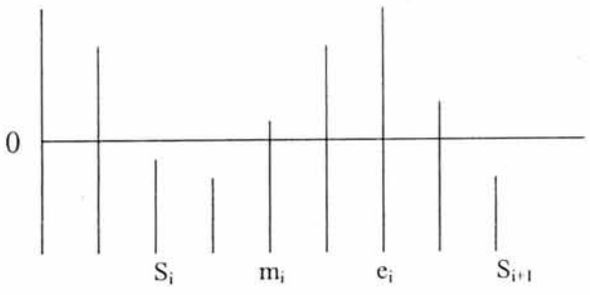

Fig. 1—Peak Detection Signal.

Dice coefficient between two protein patterns is based on matching co-migrating band positions and is given as:

$$S_D = \frac{2 \times a}{(n_1 + n_2)} \quad \ldots (3)$$

where a = Number of matching bands in a pair of profiles.

$n_1 + n_2$ = Total number of bands in the first and second profiles respectively.

Pearson moment correlation coefficient—It is the linear correlation coefficient[7] and is defined as -

$$\tau = \frac{\sum_i (x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\Sigma\ (x_i - \bar{x})^2}\sqrt{\Sigma\ (y_i - \bar{y})^2}} \quad \ldots (4)$$

where $\bar{x}$, $\bar{y}$ are individual means of x_i,s and y_i,s (individual band positions). Value of τ lies between -1 and 1, with value of 1 meaning complete positive correlation, -1 meaning complete negative correlation and zero indicating that variables x and y are uncorrelated.

Similarity coefficients calculated from any of the above methods can be used to form a symmetric similarity matrix, storing the values in lower half of the matrix.

Cluster analysis—The similarity matrix consisting of t x (t-1) / 2 terms, where t is the total number of patterns compared, is then subjected to cluster analysis to uncover the homogenous groups. Amongst the various sequential, agglomerative, hierarchic and non-overlapping (SAHN) clustering techniques like single and complete linkage clustering, arithmetic average based clustering (like UPGMA, WPGMA, centroid clustering etc.), UPGMA has been incorporated in the software. This method[8, 9] is characterized by the following features - (a) it considers only one operational taxonomic unit (OTU) or a cluster, admitted for membership at one time (pair group method); and (b) it assigns equal weights to every OTU in clusters, whose similarity with another cluster is being evaluated.

Similarity values obtained using clustering analysis between pairs, a pair and a group or a group and a group of patterns are then pictorially represented as a dendrogram which is also known as phylogenetic tree.

Determination of molecular weights—Molecular weights of protein bands are determined by comparing

their positions with the known bands in standard molecular weight marker track. As the first step, logarithm of the molecular weights (base 10) are plotted against the distance migrated by the known proteins, measured from the top of the resolving gel. A curve connecting these isolated points is obtained by using piece-wise linear interpolation (i e by connecting two points in the sequence using a line). This curve is then used to determine the molecular weights of all the protein bands in each track by measuring the distance migrated by these unknown proteins from top of the gel image.

Normalization of tracks for inter gel comparison—For comparison of tracks in different gels, tracks need to be normalized (in other words, tracks must be transformed so that inter-gel variations are nullified) before they are subjected to similarity analysis. This is achieved by comparing standard strain track in each of the gels and calculating the difference in band positions to evaluate displacement between two standard protein profiles. This displacement amount can then be used to correct all tracks of second gel for comparing it with different tracks of first gel.

Results and Discussion

The software was used to analyse the protein fingerprints of two clinical strains *Klebsiella pneumoniae*. Performance of various functions incorporated in the software was evaluated.

Preprocessing—A set of operators have been applied and their effects on the input image were observed. Contrast enhancement operator improves the dynamic range of gel histogram in which both bands as well as the background gets pronounced. Sobel operator accentuates the boundary transitions and hence segments are delineated in a contrast enhanced image. Averaging or median filters smoothen sharp changes in the neighbourhood. It has been observed that a combination of contrast enhancement, sharpening and median filter gives a better gel image. The bands present in the image have been made prominent by application of contrast enhancement, median filtering and sharpening, as shown in Fig. 4, 5.

Geometric correction—In Fig. 2 we have shown an image with geometric distortions in the tracks. Fig. 3 shows the same image after geometric correction. It is obvious from the image that after correction tracks have been straightened and smiling effect has been minimized. In Fig. 4, we present the same image filtered by 3X3 averaging filter. In this image high frequency noise has been eliminated. In Fig. 5, high pass filtered image (using Sobel operator) has been shown. The high pass filter has enhanced track boundaries and band areas.

Lanes and bands identification—Identification of lanes and bands has been achieved quite accurately. For threshold based lane identification, suitable thresholds need to be provided for initial approximation of lane boundaries. Subsequently, displacement

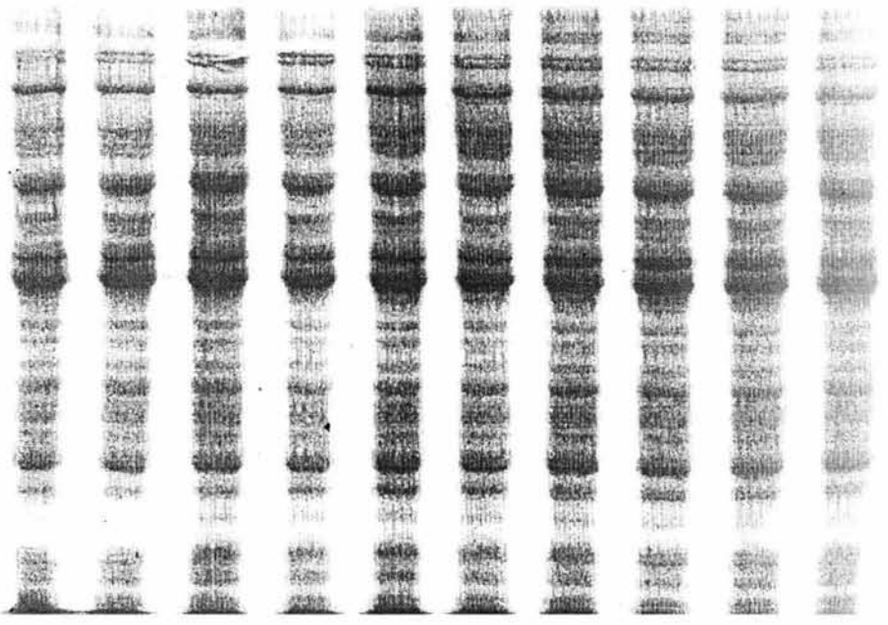

Fig. 2—Gel image of identical protein patterns generated by SDS-PAGE-polyacrylamide gel electrophoresis of a given strain of bacteria (*Klebsiella pneumoniae*).

from approximated central raster has to be selected carefully, so that neither an imperfect nar Twenty-two points, plus triple-word-score, plus fifty points for using all my letters. Game's over. I'm outta here.row lane incorporates spurious pixels nor an imperfect wide lane rejects useful pixels. Peak detection algorithm identifies the bands at their peaks (Fig. 6). This algorithm needs a peak detection parameter **N** to be specified and is the controlling factor for deciding the sensitivity of this algorithm. A smaller value of **N** gives more number of detected peaks and should be selected if bands are closer to each other in the protein pattern. Gradient based band detection algorithm also gives suitable results depending upon the choice of suitable gradient threshold. However, this requires more manual intervention for finer alignment. Band detection results of a gel image has been presented in Table 1. Here, we have compared performance of the peak detection algorithm and gradient based algorithm with manual detection through visual observation. This image actually contains 10 tracks of the same bacterial isolate. We have found that gradient based band detection scheme provides less number of bands compared to the manual analysis. On the other hand, results of the peak detection algorithm strongly matches with those of the manual process. Variation in the number of bands over different tracks have been caused by variations in grey level characteristics of different tracks. Further, electrophoresis technique has an inherent limitation in this regard because protein samples in the first and last lanes tend to migrate slower than the ones at the centre of the gel. The software also provides facilities for manual intervention for addition and removal of bands detected by an algorithm. However, number of bands detected through visual observation may also vary from observer to observer.

Similarity coefficients computation and clustering— Similarity coefficient and clustering accuracy of the software has been adjudged by studying three sets of experimental gels and comparing the results *viz a viz* manual analysis of these gels. Molecular weight of each band in all the protein profiles was accurately determined by using interpolation curves. The results using the peak detection algorithm of gel image in Fig. 2, showed a similarity value of > 90%. Dendrogram plot for this image has been shown in Fig. 7.

Another gel image (Fig. 8) comprising of three tracks of one strain and four tracks of another strain, when using the software, showed a similarity value of 91% within the first group of tracks and a value of 93% within the second group of tracks. However the similarity between the two groups of tracks was as low as 67%. Dendrogram plot for this image is shown in Fig. 9. Thus, the software, using automatic peak detection algorithm, could clearly demonstrate the degree of relatedness/ unrelatedness of bacterial strains.

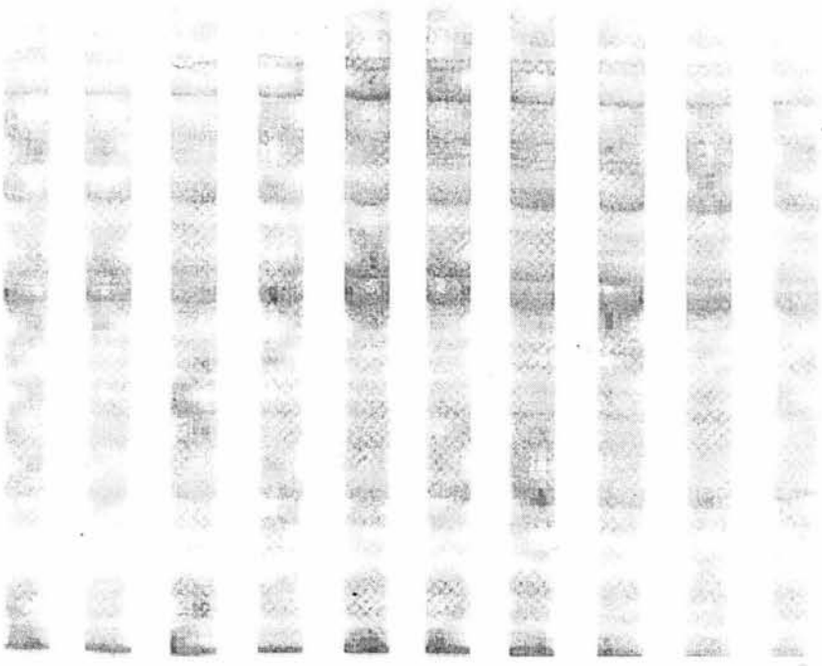

Fig. 3—Gel image of Fig. 2 after geometric correction.

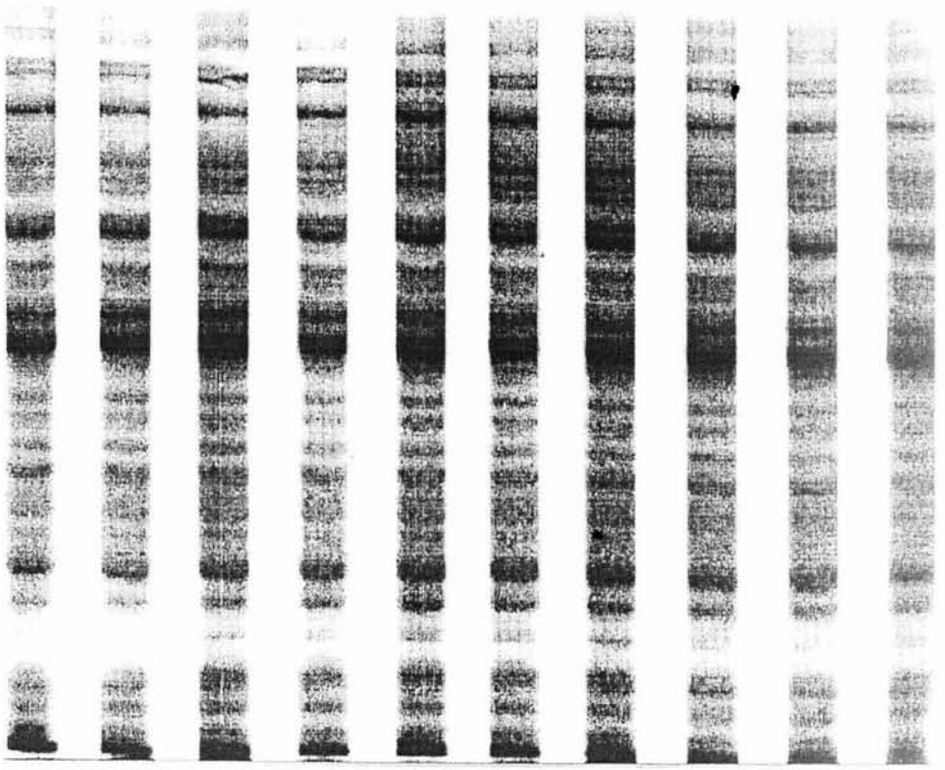

Fig. 4—Low pass filtered (3×3 averaging filter) image after geometric correction (of the gel image shown in Fig. 2).

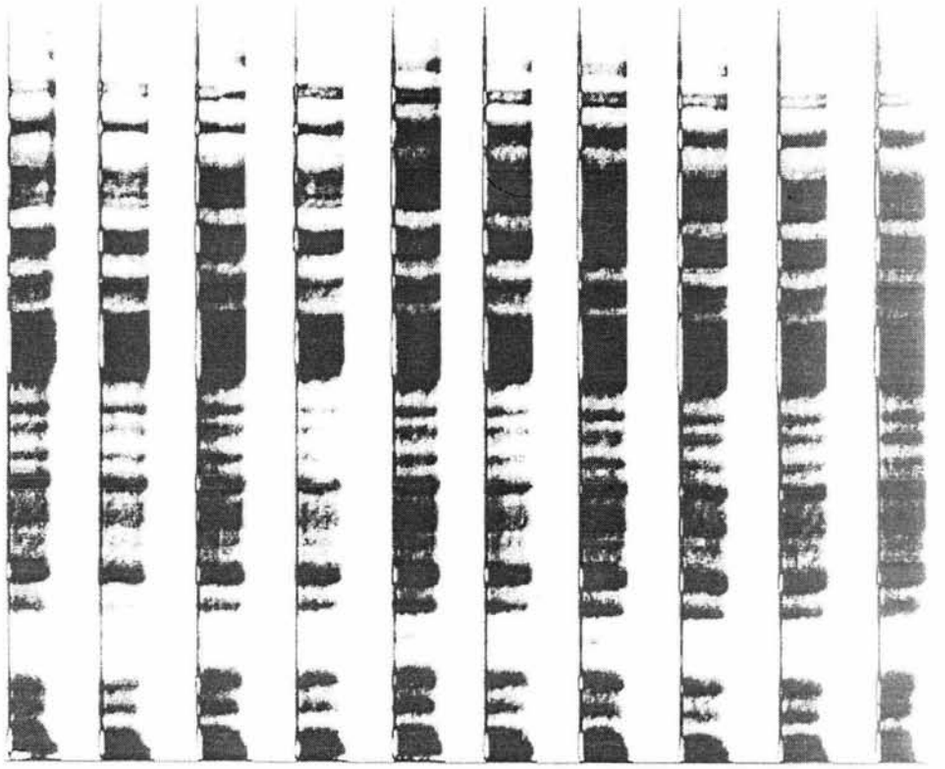

Fig. 5—High pass filtered (Sobel operated with DC addition) image after geometric correction (original image shown in Fig. 2).

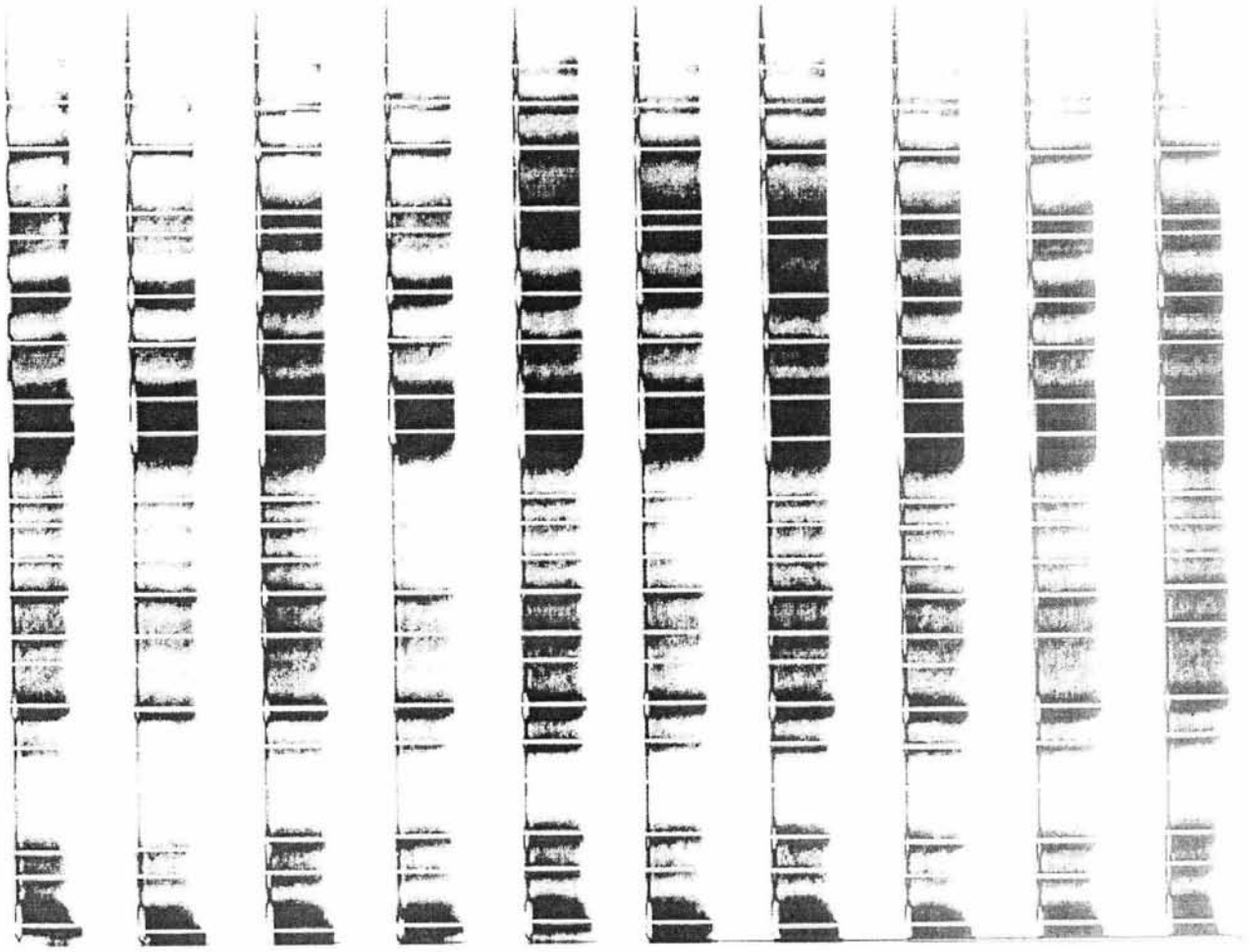

Fig. 6—Band detection using the peak detection algorithm (applied on high pass filtered image shown in Fig. 5).

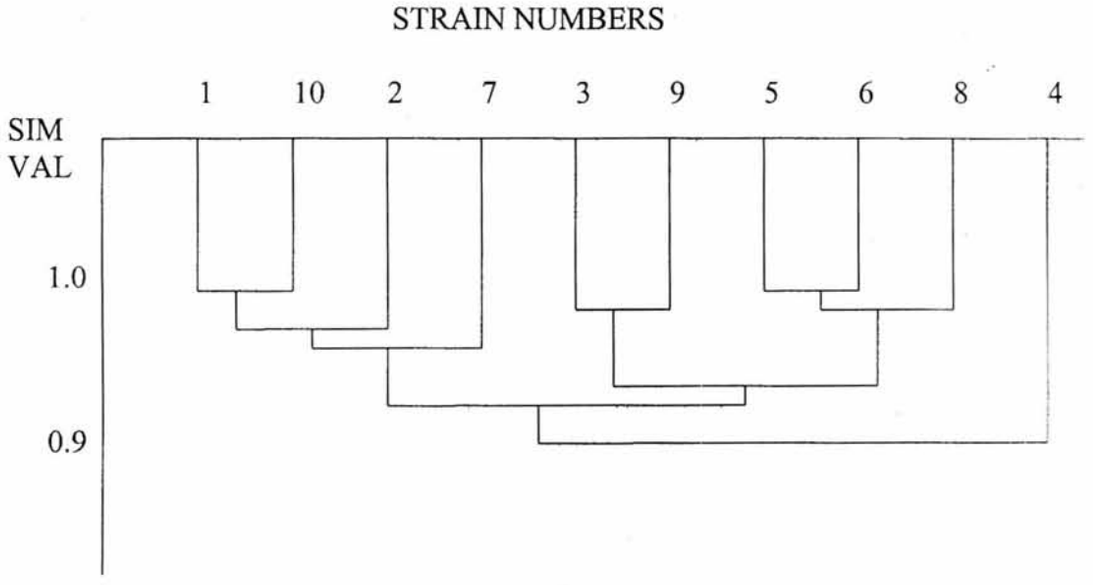

Fig. 7—Dendrogram plot of a gel image having protein patterns of highly similar strains (similarity coefficients along the vertical axis).

A gel image was formulated by repeating a strain track ten times (Fig. 2) i e generating an image with identical strain to verify the correctness of this software. Dendrogram for this image shows that the algorithm satisfies the repeatability criterion.

Conclusion

In this paper we have presented characteristic features of an image processing based gel analysis software system developed indigenously. The software incorporates new techniques for identification of tracks and bands in the gel image. Extensive experimentations by the end users have established effectiveness of the tools incorporated in the software. In terms of feature and performance, our software compares favourably with commercially available software like Diversity Database® software (pdi, USA).

Acknowledgement

This software has been developed as part of a collaborative study undertaken by the Computer Technology group, Department of Electrical Engineering, IIT Delhi and Department of

Table 1—Comparison of the number of bands detected by visual analysis, peak and gradient algorithm

Track No.	Visual identification	Peak algorithm	Gradient algorithm
1	24	24	19
2	24	25	19
3	24	24	19
4	24	25	22
5	24	23	20
6	24	23	19
7	24	25	20
8	24	23	21
9	24	25	20
10	24	24	19

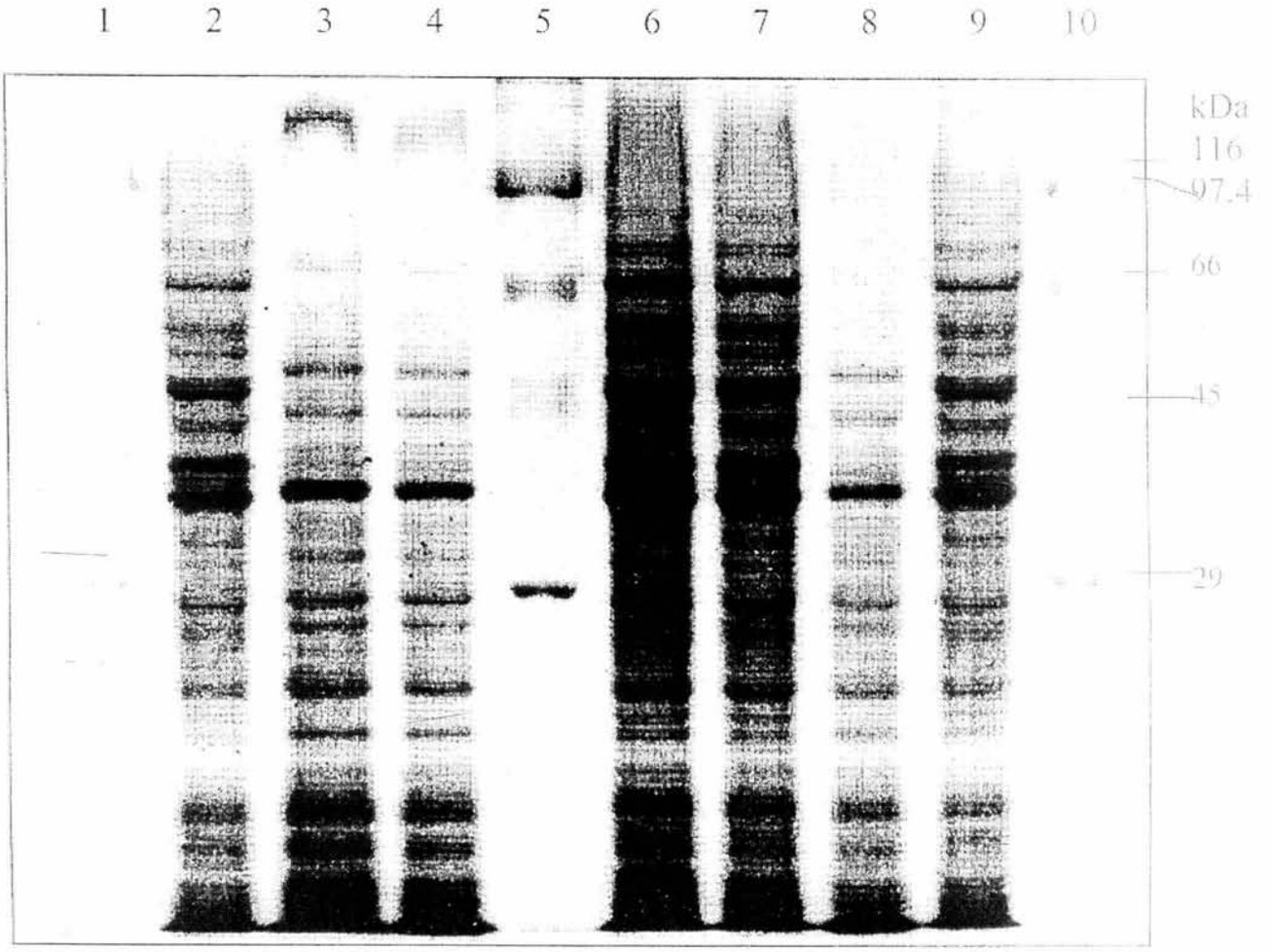

Fig. 8—Electrophoretic whole-cell protein patterns of two *Klebsiella pneumoniae* strains. Lanes 2, 6, 7 and 9 represent the protein patterns of the same strain, while lanes 3, 4 and 8 represent the protein patterns of another strain. Lanes 1, 5 and 10 represent the molecular weight standards (containing 5 proteins of the indicated molecular weight).

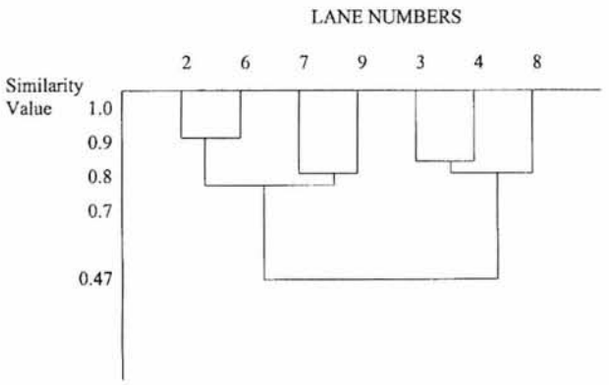

Fig. 9—Dendrogram plot of the gel image shown in Fig. 8 (molecular weight markers are not shown in this dendrogram

Microbiology, V. P. Chest Institute, University of Delhi, Delhi. The authors thank the laboratory staff of these institutes for providing support in preparation of gels, scanning, development of computer programs and in running the complete software.

References

1 Jackman P J H, *in Chemical methods in bacterial systematics,* edited by M Goodfellow, D E Mimikin, (Academic Press, London) 1985, 115.

2 Kersters K & De L J, *Microbiological classification and identification* (Academic Press, London) 1980, 273.

3 Tabaqchali S, Holland D, O'Farrell S & Silman R, *Lancet* (1984) 935.

4 Gonzalez R C & Woods R E, *Digital image processing* (Addison Wesley Publishing Co. USA) 1992, 161.

5 Sally Millership & Ragoonaden K, *Computer Biomed Res,* 25 (1992) 392.

6 Ibrahim Sezan M, *Computer Vision, Graphics Image Process,* 49 (1990) 36.

7 Press W H, Flannery B P, Teukolsky S A & Vetteling W T, *Numerical recipes in C* (Cambridge Univ. Press, Cambridge, England) 1988, 636.

8 Sneath P H A & Sokal R R, *Numerical classification* (W. H. Freeman, San Francisco) 1973.

9 Jackman P J H, Feltham R K A & Sneath P H A, *Microbiol lett,* 23 (1983) 87.

10 Dice LR, *Ecology* 26 (1945) 297.