Processing of DNA and Protein Electrophoresis Gels by Image Analysis

Donald G. Bailey and C. Bruce Christie

Image Analysis Unit and Plant Science Department
Massey University, Palmerston North
E-mail: {D.G.Bailey, C.B.Christie}@massey.ac.nz

Abstract

With recent legislation allowing for the registration of new cultivars, the analysis of DNA and protein electrophoresis gels is becoming increasingly important for cultivar identification. DNA fragments or proteins of different molecular weights are separated using electrophoresis, giving a series of bands with positions corresponding to the molecular weight. Image analysis of the gels removes much of the subjectivity of manual comparison of band position and intensity between samples.

The first step in processing corrects for geometric distortions, called smiling, caused by variations in conditions during electrophoresis. Most of the effects of smiling may be eliminated by detecting and straightening a pair of bands (one at each end of the lane) common to most of the lanes. Next, the individual lanes corresponding to each sample are automatically detected. The background fog is estimated and removed by subtracting it from the image. The bands are then detected and the positions and densities of each band are determined. By including a series of proteins or nucleotides of known molecular weights in one of the lanes, the unknown molecular weights of the bands in each sample are able to be estimated based on their position.

The use of image analysis techniques in this application provides a relatively quick and inexpensive method of objectively identifying differences between samples.

1. Background

The opportunity to reliably identify plant and animals including humans by means of a protein or genetic fingerprint is a recent development that has received a wide publicity in both the popular and scientific press. The application of these new methods of analysis in agriculture and horticulture has allowed the direct assessment of differences in protein and DNA in individuals, and this has increased the demand for more precise methods of characterising plants and animals. Naturally this is of interest to scientists but it also has economic importance in conferring a unique description on genetic resources that may be required for identification purposes in the case of disputes. This is made possible by examining the genotype rather than the phenotypic features that may be observed at the macroscopic level. Correct identification of species and hybrids can be difficult and frequently misleading if based solely on the morphological characteristics of individuals. To a purist, the analysis of the genotype has much to offer as it is not dependent on expression of genes that contribute to phenotypic variation. The emergence of techniques allowing rapid analyses of protein and DNA offers a new suite of information that may be used to provide a unique description of an individual without the interference of the environment or seasonal variation. The stability of these methods of assessment offer new possibilities to characterise individual animals or plants or specific strains of microorganisms. It is also possible to investigate probable lineage or parentage, an

area that has legal credibility and is used routinely in forensic cases as may be an issue in determining the purity of seed lines.

1.1 Limitations

The rapid protein and DNA analysis techniques are ideal for taxonomists. However standardisation of the reaction conditions is critical to obtain repeatable results, this is much more of problem between laboratories than within a laboratory [1].

Protein expression is a phenotypic character that is regulated by an interaction between the genotype and the environment, hence it is important to standardise conditions to increase the validity of any intended comparisons. In contrast the stability of the nuclear material in a particular genome provides an excellent material for comparisons between related plants or even between related animals. Provided the extracted DNA is of both high purity and quality representing the genotype being examined it is capable of being used to generate stable fingerprints of the organism using either restriction fragment length polymorphism (RFLP) or random amplified polymorphic DNA (RAPD) techniques.

1.2 Preparative methods

Preparation of protein profiles requires a sample of protein to be solubilised and then the extracted components separated by electrophoresis in a gel matrix of either agarose, poly acrylamide or starch. Samples are loaded into wells in the gel and then separated by a potential difference across the gel. The separation of whole proteins or their components depends on the physicochemical properties of the proteins including size, charge and isoelectric point. The migration of material in the gels may only become apparent after the gel has been stained to reveal a characteristic profile for the sample.

Any DNA analysis begins with the isolation of the DNA. This may be purified or if one of the amplification procedures like RAPD relatively crude preparations may be used with no loss in efficiency. The large size of DNA precludes simple separation by electrophoresis, this problem is often overcome by digesting the DNA with enzymes (known as restriction endonuclease) that split the DNA at specific sites into smaller fragments. A thermal stable polymerase is used to produce many identical copies of the small fragments that are separated by electrophoresis on the basis of their size.

1.3 Analysis

Simple profiles are frequently analysed by eye, but complex data requires computer analysis. The data still must be input into a programme for analysis and this depends on being able to assess the position or the migration distance of the bands (and in some cases the intensity of the bands generated). While it may be possible to assess the distance migrated it is not possible to reliably quantify the intensity by eye alone.

Image analysis procedures allow some degree of automation in data processing and handling due to the filtering and computational options available. These are invaluable when the background on the gel requires correction to reveal the obscured data in the bands and also in the standardisation of the migration distance. In addition to swelling and shrinkage of the gel before and after staining, the conditions associated with sample preparation, gel loading with sample, gel composition, temperature combined with the voltage and current can all contribute to slight differences in the migration distance and anomalies such as band smiling that make unaided visual interpretation of data potentially very difficult.

2. Objectives

The objectives from an image analysis standpoint are therefore to facilitate the gathering of data from DNA or protein electrophoresis gels, and automate this process as much as possible. Specific objectives relating to the image analysis requirements are:

- To correct for severe geometric distortion.

- To remove the background fog from the image allowing the true densities to be estimated.
- To detect each lane and the positions and densities of the bands within each lane.

Most uses of the data involve making comparisons between two or more samples, or looking for features which several samples have in common. An appropriate output from the image analysis system is therefore to represent each band by a vector with components relating to the position and density of the band. The data for each sample would be the set of vectors representing the bands detected within the corresponding lane on the gel.

3. Image Processing Steps

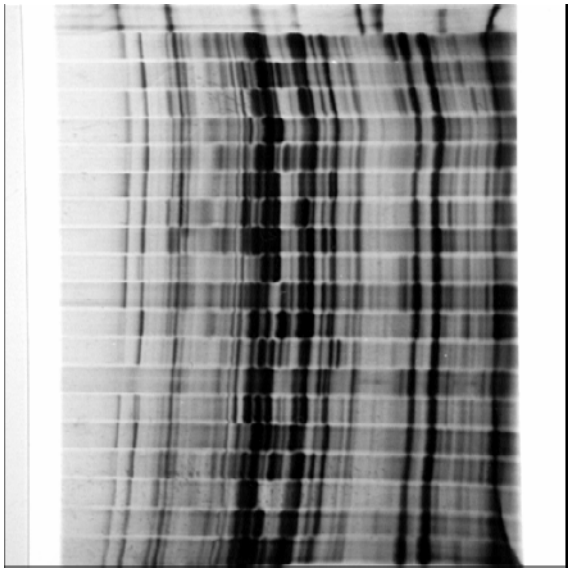

Figure 1: A typical electrophoresis gel.

3.1 Image Acquisition

Images of the DNA or protein gels are captured into the system from either photographs or dried gels using a video camera. Although it is possible to capture images of the gels directly, this in not the preferred option since the gels are fragile and may be easily distorted or damaged. In addition, radioactive samples and gels requiring UV fluorescence may need to be captured onto film to be visualised. In most circumstances, the gels are dried or photographed for archival anyway, so this does not introduce an unnecessary step. A 512 x 512 image of the gel is captured, with a typical image shown in figure 1.

3.2 Smile Correction

Geometric distortions often occur on the gel because of variations in the conditions during electrophoresis. The most common effect, caused by temperature variations on the gel, results in the central lanes running slightly faster than those at the edges. This gives a characteristic curved shape known as smiling. The first step is to correct for any such distortions if present.

Points within the image which are the local intensity minima (within a 1 x 9 pixel window) are detected. These represent the peaks associated with bands. For reliable correction, only the significant bands are of interest. These bands have a local range (within a 1 x 15 pixel window) greater than a preset threshold. The points detected in this way are overlaid on the original image in figure 2a.

The user is then prompted to select a significant band common to most lanes near the left edge of the region of interest. The detected point nearest to each selection is used as a seed to grow a reference line which represents the distortion in the image. To grow the reference line, the top and bottommost pixels of the selected line are found. If these are not near the top or bottom

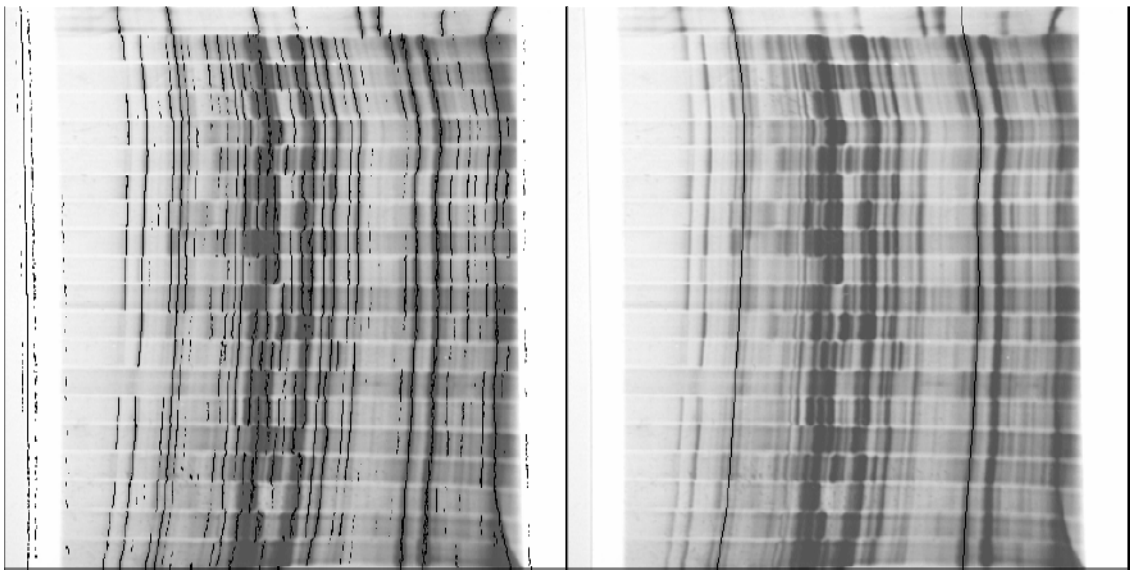

**Figure 2: a) Significant bands that may be potentially used for smile correction;
b) Smooth curve fitted through selected bands common to most lanes.**

edge of the image, the nearest band vertically is found and added to the seed line. This process is continued until all of the lanes containing the selected band have been added.

As the individual points which have been detected may contain noise, a smooth curve is fitted through the points. To accomplish this, the image is split into seven sections, each section 128 pixels high, with a 64 pixel overlap between the sections. A quadratic curve is fitted through the points in each segment using least squares. Where the sections overlap, the two polynomials from the overlapping sections are splined to give a cubic. This represents the broad shape of the distortion without being sensitive to noise in the image.

The user is prompted for a second significant band on the right hand side of the image. This is processed in the same way giving a second smooth curve representing the distortion. The lines detected in this manner are shown in figure 2b.

Warping the image to make these two lines straight is sufficient to correct most of the geometric distortion in the image. Each row of the distorted image is stretched linearly and repositioned to straighten both detected lines (figure 3a). This process will be most accurate

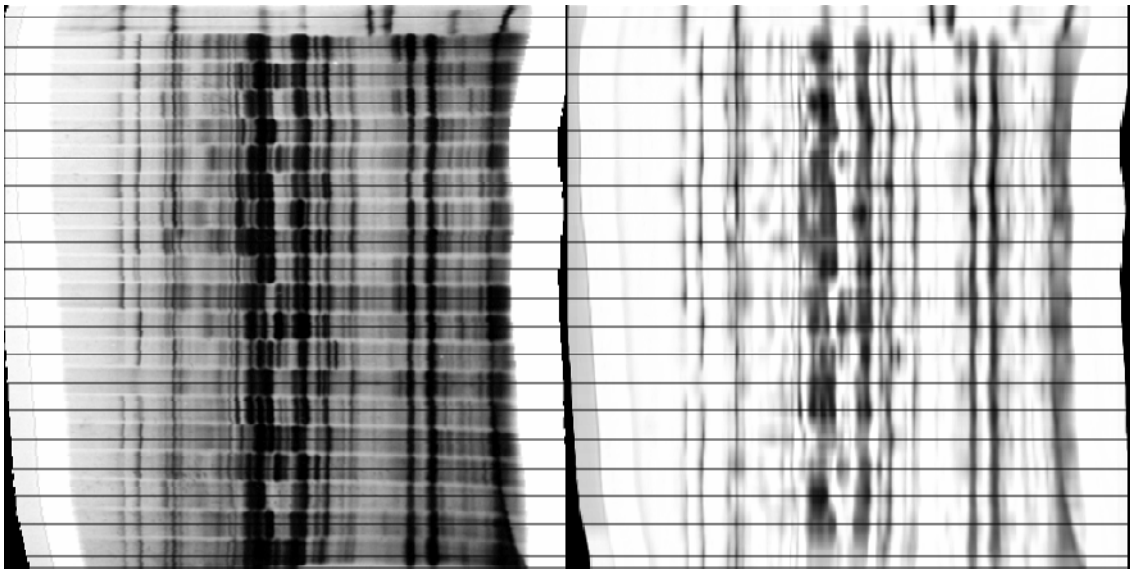

**Figure 3: a) After correcting for the smile in the image;
b) Image after removing background fog and smoothing vertically to remove noise.
Horizontal lines show the automatically detected lane centres.**

between the selected lines, requiring the selected bands to be as close to the left and right hand regions of interest as practical.

After smile correction, the user selects the boundaries of the region of interest. This allows regions which are off the end of the gel to be ignored in later processing.

3.3 Lane detection

As the lanes have a slightly lighter gap between them, the gaps, and hence the lane positions may be detected automatically. The image is averaged horizontally, giving a 512 x 1 average image. This is filtered to find the gaps between the lanes, and then the lane centres. The detected centres are shown overlaid on the straightened image in figure 3a. Occasionally the detected centre is not right, so the user is asked to add missing lanes or remove extra lanes by clicking with a mouse.

3.4 Background Removal

There is often a background "fog" on which the bands are superimposed. The density of this background varies across the image, and also from lane to lane depending on the presence of groups of dense bands. This background does not contain any information of interest, but instead obscures the true density of the bands. The level of the background fog is estimated by using a maximum filter along the rows (using a 1 x 39 window). This removes all bands within the image. The positions of the brighter regions are restored by following this with a minimum filter using the same window size. Sharp edges are then smoothed by averaging using a 1 x 19 window. This forms an estimate of the background fog which is subtracted from the straightened image. The effect of removing the fog is readily apparent in the two profiles shown in figure 4b.

3.5 Band Detection

Since each band is vertical in the image, any noise present may be reduced by averaging along the length of the bands (using a 19 x 1 window). This also smears the lanes together along their edges (see figure 3b), however this is not a problem since the lane centres have already been detected in a previous step. The 19 pixel window is sufficiently short that the blurring does not reach the centre of the lane distorting the value used. The final step is to detect the local minima (within a 1 x 7 window) along each row, as these represent the positions of the band peaks. All of the bands which have a density greater than a preset threshold level are kept. This prevents spurious noise from being detected, and also faint false bands which can occur as a

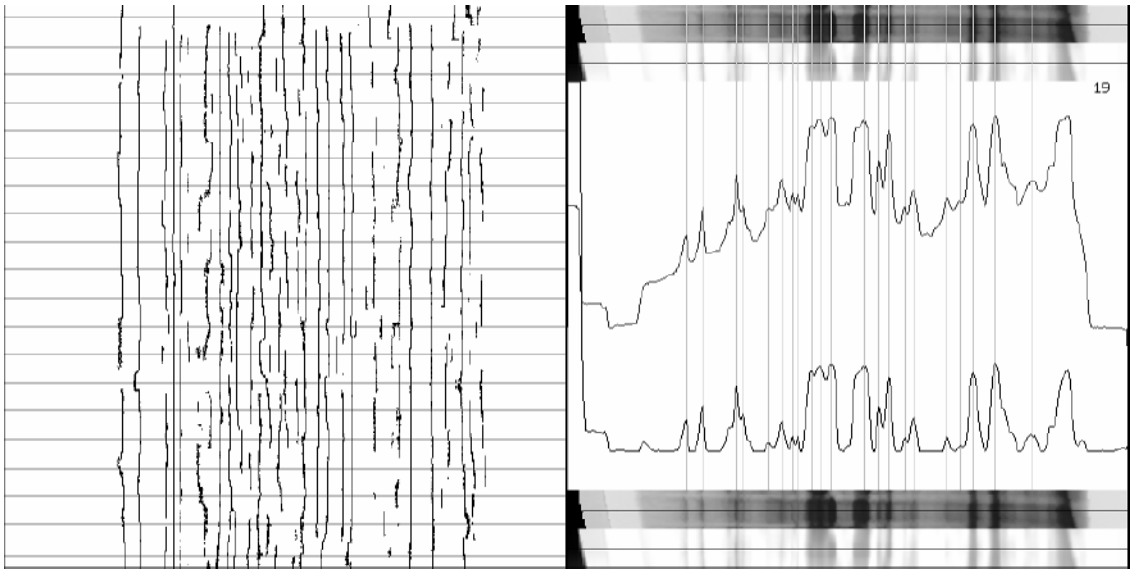

Figure 4: a) The local minima in the image corresponding to bands above the noise threshold; b) The original and filtered images of lane 19 (second lane from the bottom) showing the density profiles and positions of detected bands.

result of contamination of the sample. The detected bands which are in the region of interest are shown in figure 4a.

The density profile for each lane is then presented in turn to the user in the manner of figure 4b. The user is shown the original lane, and the processed lane along with the corresponding density profiles. The positions of the detected bands are marked for verification. The user is able to add any bands that were missed by the computer, or to remove any false bands that may have been detected. Most bands are detected reliably by the computer, although sometimes a band needs to be added if two very close bands are unable to be resolved automatically. The only time that bands need to be removed is when noise or a dust speck causes an extra band to appear.

Finally, the horizontal position of each band in the lane is converted to molecular weight, and that weight and the density are written to a data file for later analysis.

3.6 Calibration

The positions of the bands within the image are able to be measured directly by the system. While this is adequate for making comparisons between samples run on the same gel, the positions of the bands on different gels are cannot readily be compared because variations in the data capture process mean that a particular band will not always be in the same position within the image.

Every gel is unique in its influence on the migration distance for calibration samples, therefore any methods that allow comparison between bands with enhanced precision would be useful. Variables may be introduced during electrophoresis (the temperature, operating voltage or other conditions may be different); during the processing of the photographs (varying camera to gel distances, or enlargement factors); or during image capture (different positioning of the photographs or enlargement factors).

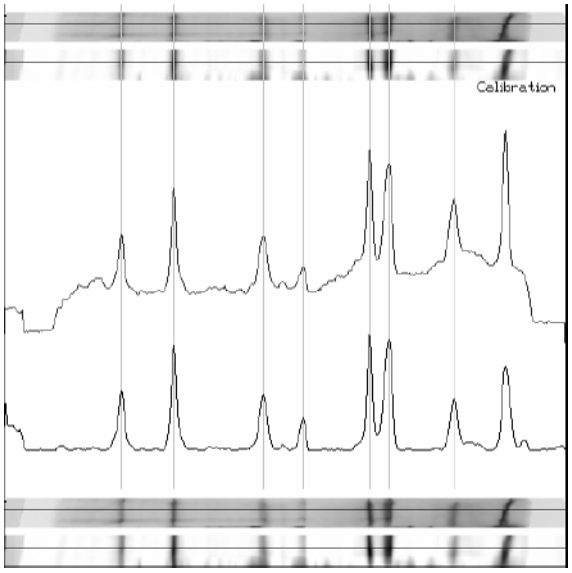

Figure 5: The images and the density profiles for the calibration lane.

To allow the results obtained for different gels to be compared, the results from each gel must be normalised to a uniform standard. Since the final position of a band in the image depends very strongly on molecular weight of the corresponding protein or DNA fragment, this is the obvious choice for our standard. This then becomes of problem of estimating the unknown weights associated with each of the detected bands.

If the molecular weights of unknown fragments separated by electrophoresis are to be estimated then this requires a precise estimate of the migration distance for a mixture of fragments of known size. Errors introduced here will reduce the validity of any interpolation of fragment size of the unknowns. Image analysis techniques provide greatly enhanced resolution of migration distance, thus minimising inaccuracies in the data collection stage. Simple

mathematical models have been developed to allow the sizes of unknown fragments to be estimated based on a standard curve. The log of the molecular weight of small proteins from 15,500 to 200,000 was linearly related to the relative migration distance [2,3]. An identical model has been applied to DNA fragments [4], while other models include logarithm of the molecular size versus the logarithm of the migration distance [5] and molecular size versus the reciprocal of mobility [6]. More complex models have been developed for estimation of protein and DNA fragment molecular weight after electrophoresis [7] and some methods are reviewed by Owen and Beck [8].

All of the methods require that a calibration lane be used. The calibration lane contains a standard set of proteins or DNA fragments of known molecular weight. By relating the measured positions of the bands within the calibration lane (see figure 5) to the known molecular weights of the corresponding components, the various parameters associated with the model may be determined for each gel. We have used the model

$$position = a \log weight + b \qquad (1)$$

where the parameters *a* and *b* are determined by performing a least squares fit between the positions and molecular weights.

4. Accuracy

A number of factors affect the accuracy of the results obtained using image analysis.

The camera resolution limits the amount of fine detail that may be obtained from the gel. This in turn limits the size of gel that may be reliably processed. The position of each band is determined only to the nearest pixel, and bands must be separated by a minimum of 3 pixels to be resolved by the algorithm described above. If a gel contains many closely spaced bands then it is necessary to capture and process only a portion of the gel.

The type of gel used may also affect the capacity to resolve closely related peaks. Acrylamide gels usually give better resolution than agarose but are more prone to smiling.

The identification and separation of closely spaced bands is a difficult problem. If the bands are obscured, the dips within the banding pattern may not be readily visible. The image analysis approach presented here aids in this situation by plotting the density profiles along each lane. Merged peaks will not be detected automatically by this system unless there is a distinct local maximum associated with each peak. However, a smaller peak can often be identified in the profile as a shoulder on the side of another peak and can be picked out manually. In spite of this, very faint bands will be obscured by the presence of dense bands.

There may be other reasons for ignoring weak bands as they may be more artefacts of the extraction procedure or they could represent contamination of the equipment. In DNA research employing multiplication techniques, a single foreign DNA fragment can be amplified to produce a false positive reaction. Faint false positive bands have also been observed with the Taq polymerase from certain suppliers.

The quantification of density and position of the bands is made difficult by the dispersion or spreading of the components as the migrate along the gel. This problem is worse for lower molecular weight bands as the travel greater distances on the gel. Determining the position is compensated to a certain extent since the relative spacing (as a function of the molecular weight) is greater for these bands. However, the dispersion lowers the effective resolution since adjacent bands may merge. A more serious problem is that bands which disperse more will have lower peak densities for two reasons. First, as the material is spread out, the amount actually at the peak will be lower. Second, if there are several adjacent bands merging together, this raises the level of the background fog which is taken out. This reduction in contrast should not present too many problems provided that samples on different gels are processed in a similar manner.

The dynamic range of the video camera used to capture the images, or the film and printing process if photographs are used for input, provides limits on the quantification of band density. A particular problem is the saturation of films with very dense bands.

Finally, the model used to calibrate between distances and molecular weights does not provide an exact fit. The actual calibration will depend on the weights used in the calibration sample, and which of the bands are included in the model. Since most of the errors introduced are systematic, comparisons between samples using the same calibration weights should be straight forward. Using a different calibration series may result in different calibration errors, making the problem of matching corresponding bands in different samples more difficult.

5. Conclusions

Image analysis procedures provide an excellent opportunity to greatly enhance the reliability of detecting and using information revealed by the DNA and protein analyses. The algorithm described is able to identify the bands automatically in most cases. Where adjacent bands merge, the results may be easily edited to add missing bands. Advantages of the technique described over visual comparison are:

- its objectivity, as the positions and densities of the bands are calculated automatically;
- the ability to remove the background, enabling true densities to be obtained; and
- the ability to compare results on separate gels, because of calibration of band positions to molecular weight units.

References

1. **Williams C.E. and St. Clair D.A.** Phenetic relationships and levels of variability detected by restriction fragment length polymorphism and random amplified polymorphic DNA analysis of cultivated and wild accessions of Lycopersicon esculentum. *Genome*, **36**:619-630. (1993)

2. **Shapiro A.L., Vineula E., Maizel J. V.** Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochem. and Biophysical Research Communications* **28**:815-20 (1967)

3. **Weber K. and Osborne M.** The reliability of molecular weight determinations by dodecyl sulphate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry* **244**:4406-12 (1969)

4. **Aaij C. and Borst P.** The gel electrophoresis of DNA. *Biochimica et Biophysica Acta* **269**:192-200 (1972)

5. **Meyer J.A., Sanchez D., Elwell L.P. and Falkow S.** Simple agarose gel electrophoretic method for the identification and characterisation of plasmid deoxyribonucleic acid. *Journal of Bacteriology* **127**:1529-37 (1976)

6. **Southern E.M.** Measurement of DNA length by gel electrophoresis. *Analytical Biochemistry* **100**:319-23 (1979)

7. **Plikaytis B.D., Carlone G.M., Edmonds P. and Mayer L.W.** Robust standard estimation of standard curves for protein molecular weight and linear duplex DNA base pair number after gel electrophoresis. *Analytical Biochemistry* **152**:346-64 (1986)

8. **Owen R.J. and Beck A.** Evaluation of three procedures using a laser densitometer and microcomputer for estimating molecular sizes of restriction endonuclease digest fragments and application to Campylobacter jejuni chromosomal DNA. *Letters in Applied Microbiology* **4**:5-8 (1987)