Automatic Construction of Statistical Shape Models for Protein Spot Analysis in Electrophoresis Gels

Mike Rogers[a], Jim Graham[a] and Robert P. Tonge[b]
[a]Imaging Science and Biomedical Engineering,
University of Manchester, Manchester, UK.
[b]Protein Science, Enabling Science and Technology (Biology),
AstraZeneca, Alderley Park, Macclesfield, Cheshire, UK.
mike.rogers@man.ac.uk
http://www.isbe.man.ac.uk/∼mdr/personal.html

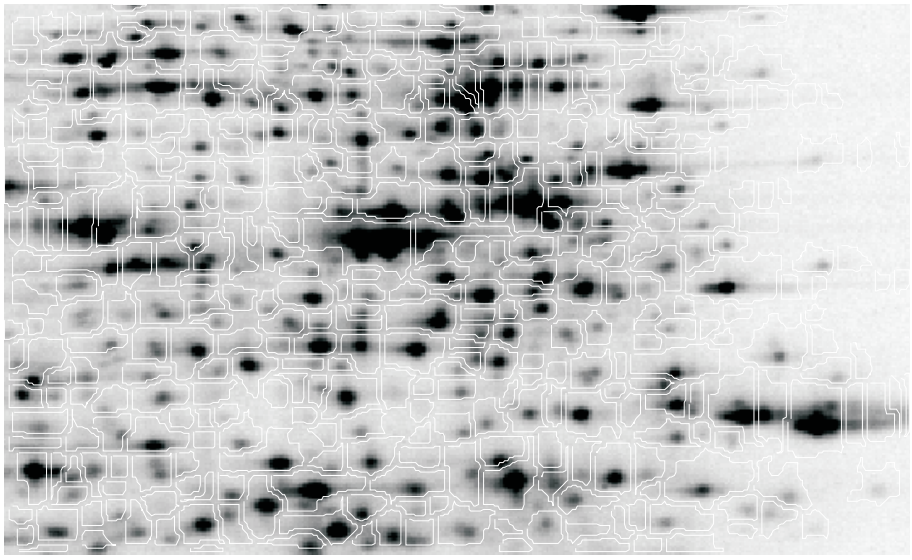
Abstract

Proteomics research relies heavily on electrophoresis gels, which are complex images containing many protein 'spots'. The identification and quantification of these spots is a bottleneck in the proteomics workflow. We describe a statistical model of protein spot appearance that is both general enough to represent unusual spots, and specific enough to introduce constraints on the interpretation of complex images. We propose a robust method of automatic model construction that is used to circumvent manual model construction which is subjective and time-consuming. We show that the statistical model of spot appearance is able to fit to image data more closely than the commonly used spot parameterisations which are based solely on Gaussian and diffusion formulations.

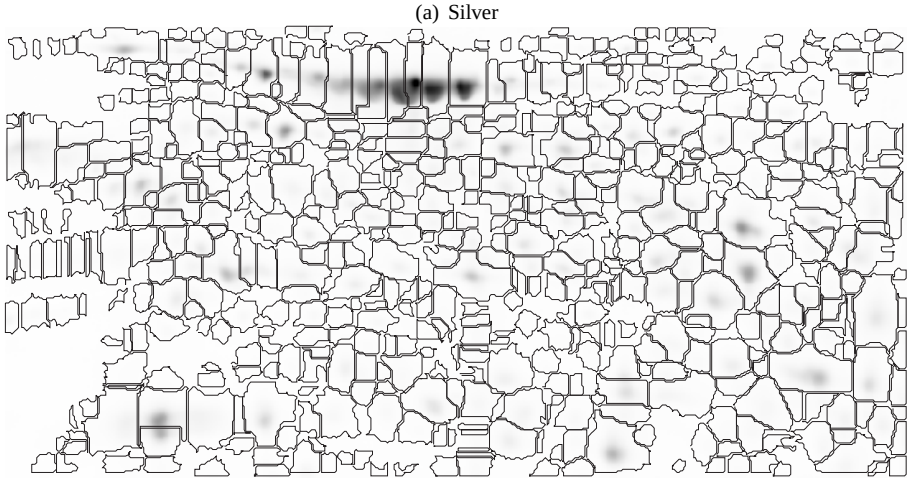
1 Introduction

Proteomics is the study of the complete set of proteins in a cell or organism throughout the entire life-cycle. It is hoped that this research will enhance understanding of cell function in general and, more specifically, it will also identify proteins that can be used as drug targets and disease markers. The main barrier to proteomics research is complexity. It is estimated that total number of proteins in a human cell could be as large as 500,000. Key to any analysis are separation and detection technologies. A well-established and widely used technology is 2-Dimensional Electrophoresis (2-DE). This process separates protein mixtures by iso-electric point (pI) and molecular weight (MW). Separation results from two separate diffusion processes which are driven along orthogonal axes in a polyacrimide gel, resulting in a grid of protein strains. The separated proteins are visualised by pre or post staining, yielding an image, containing protein 'spots'. Figure 1 shows two segments of 2-DE gel images stained using different techniques. In practice, 3,000-4,000 spots can be visualised on a single gel image, each representing an individual protein strain. The analysis of these complex gel images is a significant bottleneck in the proteomics research workflow [6].

BMVC 2003 doi:10.5244/C.17.36

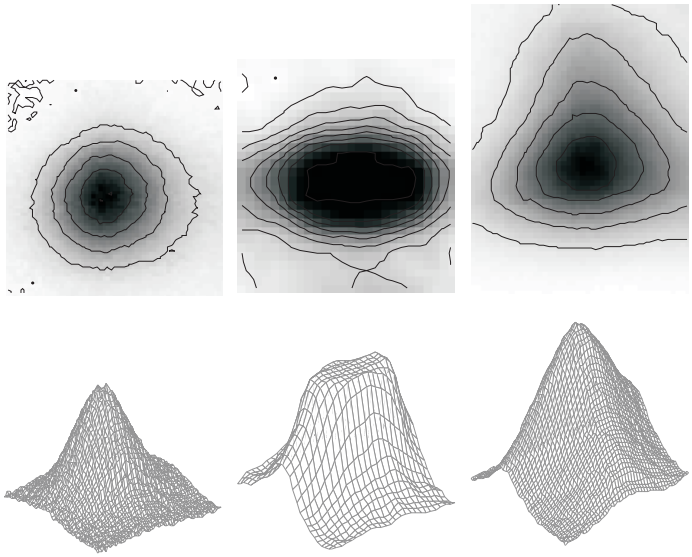


(a) Silver



(b) Fluorescent

Figure 1: Example electrophoresis images with watershed boundaries. (a) A sliver stained image with 403 delineated fitting regions. (b) A fluorescent dye image with 573 fitting regions.



(a) Gaussian (b) Flat-Top (c) Irregular

Figure 2: Examples of electrophoresis gel spots. The top row shows the appearance of the spot in the image with contours of constant gray-level overlayed. The bottom row shows a 3D mesh representation of the same data. (a) Gaussian, (b) 'Flat-top', (c) Irregular.

Image analysis of 2-DE gels requires the identification of a large number of individual spots. These must be characterised for further analysis of the sample. One of the first steps in any spot detection algorithm is the segmentation of individual spots from the background. After the segmentation step, spots are quantified and represented as a list of parameters over which further analysis can be carried out. Commonly, protein spot models are used to aid quantification by imposing constraints, which in turn improves the robustness of the solution. The most commonly used spot model is a Gaussian function [4]:

$$S(x,y) = B + I \exp\left(-\frac{(x-x_0)}{2\sigma_x^2}\right) \exp\left(\frac{(y-y_0)}{2\sigma_y^2}\right) \tag{1}$$

where B is background intensity, I is spot intensity, x_0 and y_0 control spot location and σ_x and σ_y control the spread of the Gaussian independently in x and y directions. Figure 2(a) shows an example of a typical protein spot with a Gaussian profile. This model is assumed to provide a good representation of most spots present in most gel images. However, it has been shown that Gaussian models produce an inadequate fit to some protein spots, most notably large volume, saturated spots [1]. Figure 2(b) shows an example of a high volume protein spot exhibiting a saturated, 'flat-top' shape. Bettens [1] addressed this shortcoming by proposing a model based on the physics of the spot formation process. Protein spots are formed by a diffusion process, which is only adequately represented by a Gaussian when the initial concentration distribution occupied by the sample has a small area. Bettens' diffusion model more adequately represents spots in the gel when

this assumption is not met:

$$\begin{aligned} S(x,y) = \quad & B + \tfrac{C_0}{2}\left[\operatorname{erf}\left(\tfrac{a'+r'}{2}\right) + \operatorname{erf}\left(\tfrac{a'-r'}{2}\right)\right] \\ & + \tfrac{C_0}{r'\sqrt{\pi}}\left[\exp\left(-\left(\tfrac{a'+r'}{2}\right)^2\right) - \exp\left(-\left(\tfrac{a'-r'}{2}\right)^2\right)\right] \end{aligned} \tag{2}$$

where $r' = \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}$, B is background intensity, C_0 is initial concentration, D'_x and D'_y are related to the diffusion constants in the two main directions of diffusion, x_0 and y_0 control location and a' is the area of the disc containing the protein material. As $a \to 0$ equation 2 reduces to the bivariate Gaussin (eqn. 1).

Both the Gaussian and diffusion models assume perfect diffusion across the gel medium. Spots created by a perfect diffusion process will be regular and symmetric. In practice, the diffusion process is not perfect and spots can be formed with unpredictable, unusual shapes. An example of such a spot is shown in Figure 2(c). To represent more adequately the full range of observed spot shape, we have developed a new protein spot model that is both flexible enough to represent irregular shape variation and specific enough to retain usable constraints on the interpretation of gel images. The physical process by which irregular spots are formed is extremely complex. It would be daunting task to directly estimate all the physical variables affecting spot formation. Instead, we have used a Point Distribution Model (PDM) [3] to represent observed variation in spot shape. Gaussian convolution simulates the diffusion process and forms a full model of spot appearance. In section 2 we describe the model, together with an automatic method for model construction. Results of an evaluation of the model and a discussion are presented in sections 3 and 4.

2 Modelling Protein Spot Shape and Appearance

To represent observed variation in protein spot shape we have used a PDM trained with a set of protein spot boundaries. The PDM only represents shape, but we require a full model of spot appearance. Protein spot formation in 2-DE gels is a diffusion process which is equivalent to convolution of an initial concentration distribution with a 2-D Gaussian kernel. We have assumed the initial concentration distribution can be represented as a flat 2-D shape within the boundary represented by the shape model. This flat shape is convolved with a bi-variate Gaussian kernel giving a full model of spot appearance. Figure 3 shows an example of the full spot appearance model. We define our model using the parameter vector $\vec{p} = (B, I, x_0, y_0, \sigma_x, \sigma_y, s, \vec{b_s})$, where B is an additive background term, I is spot intensity, x_0 and y_0 control location, σ_x and σ_y control the spread of the Gaussian along the two directions of diffusion, s is a scaling for the spot shape (from the alignment procedure) and $\vec{b_s}$ is a vector of PDM shape parameters. This model is equivalent to the bi-variate Gaussian when $s = 0$, and is equivalent to the diffusion model when the shape parameters, $\vec{b_s}$, represent an elliptical shape.

2.1 Automatic Spot Model Construction

Section 2 described the basis of the models we use. Here we address the practical issue of building the model: determining the training shapes from spot images and calculating

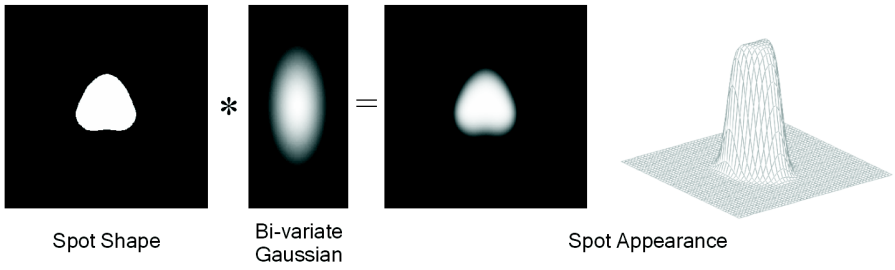


Figure 3: Spot model formation. A flat shape is convolved with a bi-variate Gaussian kernel, which is equivalent to a diffusion process.

the distributions of parameter values. In many applications of PDMs, manual marking of landmark points has been used. Due to the complexity of the images, and the number of spots required to build a model, this is an impractical strategy in this case. We proceed by segmenting the spots in the training images, smoothing the boundaries obtained using a general shape representation and making the landmark points evenly spaced round the resulting boundary. As the boundaries are extracted from real image data, a number of overlapping spots will be represented. These need to be detected and excluded from the training data, as their inclusion would bias the model and result in reduced specificity.

2.1.1 Generating the Training Set

Raw spot boundaries are obtained by thresholding the Laplacian of Gaussian transform of the training gel images. The resulting boundaries are smoothed using a Fourier shape descriptor [5] resulting in a parametrisation of the spot shape by the Fourier coefficients (5 harmonics). Spot appearance is modelled by convolving this smoothed shape with a Gaussian kernel, in the same way described in section 2. The parameters of this spot appearance model are then optimised to improve the fit to the original image data using a Levenberg Marquardt gradient descent algorithm. This provides an adjusted parametrisation of the shape matched to the image data. In this way the shapes used to build our statistical model are derived from our model of spot appearance, rather than the somewhat arbitrary data-driven segmentation. Using a Fourier representation in this strategy does not impose any explicit shape constraints on the boundaries extracted. The PDM landmark representation is obtained from the resulting spot shapes by placing 25 evenly spaced points around the boundary.

2.1.2 Robust Model Building

Automatic generation of training shapes will include incorrect shapes in the model. These shapes are the result of unseparated overlapping multi-spot groups. The Fourier shape representation imposes no explicit shape constraints, other than smoothness, so it is not possible to filter these incorrect segmentations at that stage. We could filter the resulting shapes by hand, but this would be a highly time consuming and subjective process. Rather, we have chosen to reduce the influence of such shapes by using Robust Principal

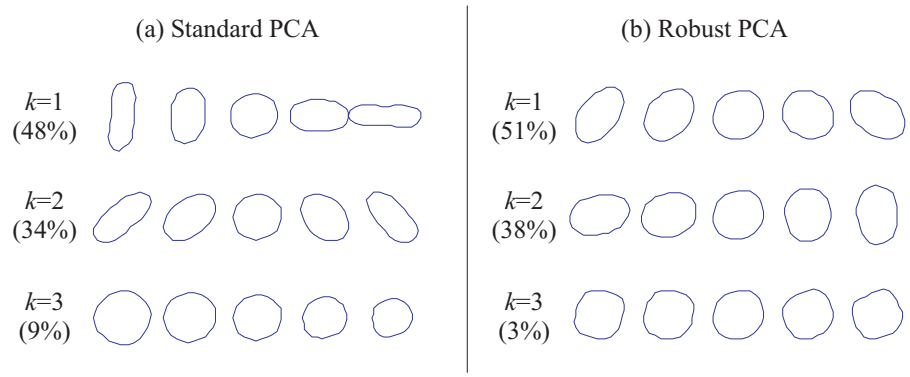


Figure 4: Robust PCA. (a) The first 3 of 10 modes (± 2 std.dev.) PDM built using standard PCA. (b) The first 3 modes of a PDM built using Robust PCA. Both models were trained with the same data.

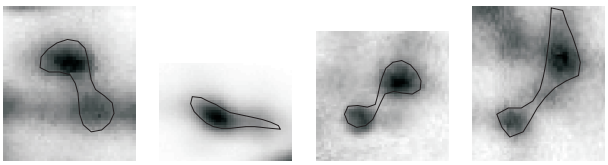


Figure 5: Four examples of shapes that have been downweighted by robust PCA. Each shape is superimposed over the image patch used in its generation.

Component Analysis [2] in the model building. We expect the number of incorrect shapes to be small and their shape to be unusual, and therefore they can only influence the model as outliers in the shape distribution. Robust PCA iteratively reduces the influence of outliers on the resulting model. The effect of the robust PCA can be seen in Figure 4. The figure shows two PDMs, one built using standard PCA (Figure 4(a)) and one built using robust PCA (Figure 4(b)). The models were generated from the same training data. Both models represent the spots by principal components that retain 99% of the observed variance, in the robust case this is 99% of the variance remaining after the iterative weighting procedure. The standard model represents the retained variance in the training data using 10 modes, whereas the robust model requires only 6 modes. The contribution of each mode to the total variance of the training set is shown for each model. The first mode of the standard model represents a large variation in aspect ratio with an apparent 'waist' becoming visible at the extremes of the mode. This mode would allow the model to represent multiple overlapping spots, which is undesirable. There is no mode in the robust model that allows shapes with 'waists'. Figure 5 shows examples of shapes that have been treated as outliers by the robust analysis. They all represent highly uncharacteristic shapes and several are clearly multiple spots.

3 Evaluation of Models

We have compared the results for fitting the statistical spot model to image data with those achieved using the Gaussian and diffusion models. The experimental procedure was as follows. Spot regions were detected in a test image using a watershed algorithm. Each of the spot models was fitted to each spot region using a Levenberg-Marquardt non-linear optimisation algorithm to determine the best model parameters, minimising the following residual: $r = \sum_{x,y \in R} \left[(S(x,y|\vec{p}) - I(x,y))^2 / \left(n_R (I_R^{max} - I_R^{min}) \right) \right]$ where R is the region of the image over which fitting takes place, $x,y \in R$ are the coordinates of the pixels within the fitting region, $I(x,y)$ are image values, $S(x,y|\vec{p})$ are the model values given the parameter vectors, I_R^{max}, I_R^{min} are the maximum and minimum image values within the region, and n_R is the number of pixels within the region. This residual provides a measure of model fit error that is normalised with respect to the intensity of the spot (which we have approximated as $I_R^{max} - I_R^{min}$) and the size of the fitting region (the number of pixels n_R). This residual form allows direct comparisons of fit quality to be made between high and low volume spots. The three models were fitted to 403 watershed delineated spots from a silver stained E.coli gel (375x228 pixels, 8 bit) and 573 spots from a gel stained with a fluorescent dye (2896x2485 pixels, 24 bit). The silver image is low-resolution and contains many saturated and overlapping spots, whereas the fluorescent image is much higher quality and contains fewer saturated or overlapping spots.

The mean residuals $\bar{r}$ for each model after fitting to all regions in both images are shown in Table 1. In general the fitting results for the fluorescent image are better due to the higher resolution of the image data. The statistical model results in the smallest average residual after fitting for both images. Figure 6 shows the mean residual for each spot model and image, grouped by volume. Group one contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume. In both cases, the largest improvements in fit made by the statistical model are associated with the largest spot volumes. We have assumed that high volume spots are more likely to produce unusual spot shapes, which, we have argued, are the best represented by the statistical model. For the silver image, small and medium volume spots (groups 1-6) give fits for the Gaussian, diffusion and statistical diffusion models that are almost equivalent. However, the statistical model results in reductions in residual for all volume groups of the fluorescent image. This suggests that in the fluorescent image all spot groups contain shape variation away from Gaussian assumptions, even the smallest spots by volume. This trend is not visible in the silver image data and this may be due to the low-resolution of the image preventing full convergence. For all spot volume groups the statistical model results in fits that are better than or equivalent to the fits of the other two models. This is achieved in both images despite large visual and resolution differences.

These results demonstrate that the statistical model is able to fit well to a wide variety of gel image types. This is to be expected, as the model has the most degrees of freedom. An important question is whether the reduction in residual corresponds to a increase in model specificity. Both images contain watershed fitting regions with multiple spots. A specific model should not represent these regions well. We have carried out the following experiment to quantify the specificity of each type of model. Our aim is to determine the relative ability of the models to distinguish between single and multiple spots, using their model fit residual value. We have manually classified each fitting region in the fluorescent image (Figure 1(b), 573 regions) into one of two classes: single spot regions (472 regions)

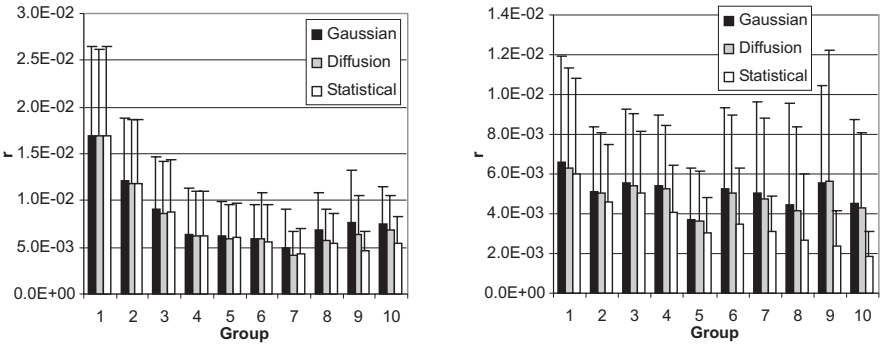
Model	Silver $\bar{r}$	Fluorescent $\bar{r}$
Gaussian	8.3×10^{-3}	5.11×10^{-3}
Diffusion	7.83×10^{-3}	4.94×10^{-3}
Statistical	7.49×10^{-3}	3.63×10^{-3}

Table 1: Mean residual after model fitting to 403 spots in the silver image and 573 spots in the fluorescent image.

or multiple spot regions (101 regions). Figure 7 shows five examples of the single spot region class, containing irregular, single spots and five examples of regions containing multiple spots, together with the fits and residuals of each model. For each of the single spot regions, the lowest residual is achieved with the statistical model. The fits of all models to multi-spot regions are visually poor (Figure 7(b)). Examination of the residuals of these 10 regions illustrates that, in general, it is not possible to define a threshold on residual value that perfectly discriminates between the two groups. This is the case for all the models. Figure 8 shows the estimated discrete probability distributions for each model for each region class. The separation of the class distributions is not good for any of the models, However, a more specific model will increase the separation between the two distributions. The distributions are non-normal, so to quantify the difference between each class we have chosen to use the non-parametric Kolmogorov-Smirnov (K-S) test. The K-S test measures the similarity between two datasets by finding the maximum discrepancy between their cumulative frequency distributions, which is called the d-statistic. The d statistic ranges between 0 and 1, the smaller the value of d, the more similar the two distributions. The discrete probability distributions (using 75 bins) and K-S distance measures for the class distributions of each model are given in Figure 8. The statistical model results in a K-S distance of $d = 0.672$, indicating that the distributions of single and multiple spot residuals are more distinct than those of the Gaussian and diffusion models ($d = 0.536$ and $d = 0.515$ respectively). This results shows that, as well as giving a more accurate quantification of 2-DE protein spots, the statistical model is more specific than the other models. The careful training and robust construction of the model results in a representation that is specific to single spots, and therefore that can not represent multiple spot regions significantly better than the other models. These selective fitting improvements lead to an increase in the separability of the two types of fitting regions.

4 Concluding Remarks

In this paper, we have described a statistical model of protein spot appearance, together with a automatic construction algorithm which takes into account the complexity of the image data. This model is both flexible and specific enough to represent the true range of protein spot appearance found in complex 2-DE gel images without the need to develop a sophisticated theoretical model of the physical processes driving irregular spot formation.



(a) Silver (b) Fluorescent

Figure 6: Mean residual $\bar{r}$ of model fit with error bars showing $+1$ std. err., plotted by increasing spot volume for each model. Spot volume group 1 contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume.

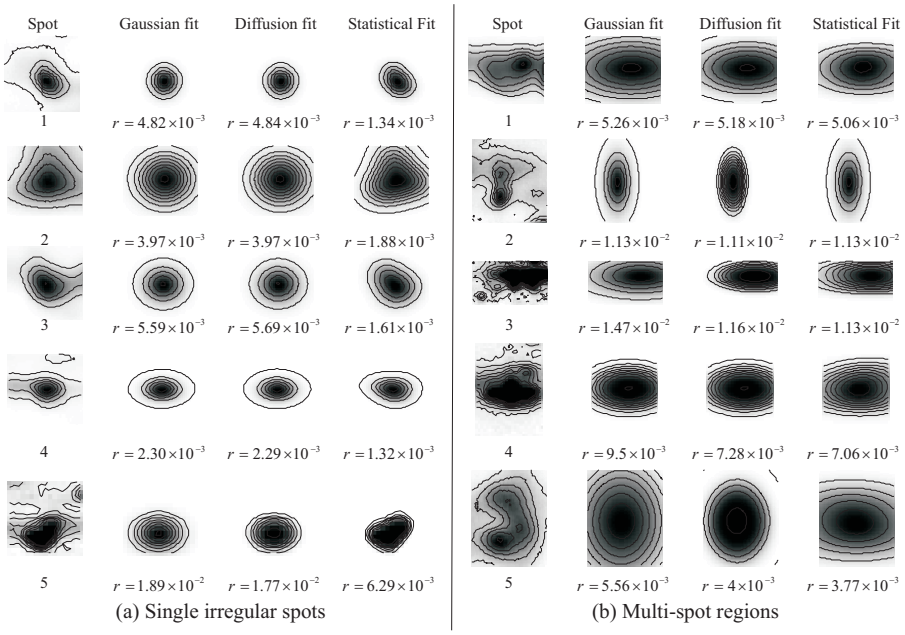
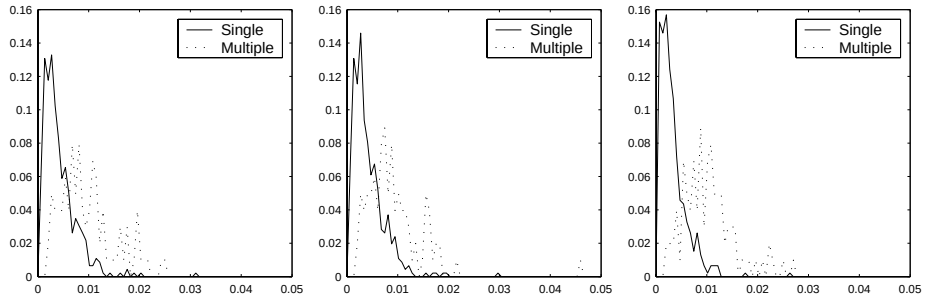


Figure 7: Example fits of each model to spot regions from the images shown in Figure 1, with resulting fit residuals for each model. (a) Shows examples of regions containing single spots with irregular shape. The improved fit of the statistical model is clear in each case. (b) Shows regions containing multiple spots. None of the models generate an adequate fit to these spots.



(a) Gaussian ($d = 0.536$) (b) Diffusion ($d = 0.515$) (c) Statistical ($d = 0.672$)

Figure 8: Discrete probability distribution (75 bins) of fit residual for single and multiple spot fitting regions with K-S distance measure. (a) Gaussian model (b) Diffusion model and (c) statistical model.

Acknowledgements

The authors are grateful for the financial support for the Biotechnology and Biological Sciences Research Council.

References

[1] E Bettens, P Scheunders, D Van Dyck, L Moens, and P Van Osta. Computer analysis of two-dimensional electrophoresis gels: A new segmentation and modelling algorithm. *Electrophoresis*, 18:792–798, 1997.

[2] N A Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, 29(3):231–237, 1980.

[3] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active Shape Models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[4] J I Garrels. The QUEST system for quantitative anaylsis of two-dimensional gels. *Journal of Biological Chemistry*, 264(9):5269–5282, March 1989.

[5] L H Staib and J S Duncan. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, November 1992.

[6] T Voss and P Haberl. Observations on the reproducability and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis. *Electrophoresis*, 21:3345–3350, 2000.