Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images

In image analysis of two-dimensional electrophoresis gels, individual spots need to be identified and quantified. Two classes of algorithms are commonly applied to this task. Parametric methods rely on a model, making strong assumptions about spot appearance, but are often insufficiently flexible to adequately represent all spots that may be present in a gel. Nonparametric methods make no assumptions about spot appearance and consequently impose few constraints on spot detection, allowing more flexibility but reducing robustness when image data is complex. We describe a parametric representation of spot shape that is both general enough to represent unusual spots, and specific enough to introduce constraints on the interpretation of complex images. Our method uses a model of shape based on the statistics of an annotated training set. The model allows new spot shapes, belonging to the same statistical distribution as the training set, to be generated. To represent spot appearance we use the statistically derived shape convolved with a Gaussian kernel, simulating the diffusion process in spot formation. We show that the statistical model of spot appearance and shape is able to fit to image data more closely than the commonly used spot parameterizations based solely on Gaussian and diffusion models. We show that improvements in model fitting are gained without degrading the specificity of the representation.

Keywords: Spot modelling / Statistical models / Two-dimensional gel image analysis  PRO 0421

1 Introduction

Image analysis of 2-DE gels requires the identification of a large number of individual spots (possibly in excess of 3000 per gel). These must be characterized for further analysis of the sample, such as comparison across a set of gels. Currently, many commercial and academic 2-DE image analysis packages are available [1–7], each with an associated spot identification and characterization algorithm. 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 characterized and represented as a list of parameters over which further analysis can be carried out. Spot characterization algorithms can be divided into two categories: parametric and nonparametric. Nonparametric methods [6, 8–11] carry out various heuristic post-processing routines on the raw segmentation boundaries to delineate the spots. Spots are then represented by a set of measurements calculated over the detected spot regions. No explicit constraints on the shape of the boundaries or the appearance of the spot are imposed. The flexibility of this approach is outweighed by the relative lack of robustness when complex images are analyzed. Parametric methods use models to parameterize protein spots. The use of models to aid the interpretation of complex data is a well established technique in image interpretation [12–15]. Models represent prior knowledge which is used to impose constraints on the analysis procedure, which in turn improves the robustness of the solution.

Commonly, nonparametric methods are used to provide an initial spot segmentation, which can then be refined using model based, parametric quantification. In 2-DE gel analysis, the most commonly used spot model is a Gaussian function [7, 16]. Figure 1(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 [17]. The Gaussian formulation is insufficiently flexible to represent the true variation in spot appearance. Figure 1(b) shows an example of a high volume protein spot exhibiting a saturated, 'flat-top' shape. Bettens [17, 18] addressed this shortcoming by proposing a model

Correspondence: Dr. Mike Rogers, Imaging Science and Biomedical Engineering, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
E-mail: mike.rogers@man.ac.uk
Fax: +44-161-275-5145

Abbreviations: PDM, point distribution model; PCA, principal components analysis
Based on the physics of the spot formation. 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.

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 Fig. 1(c). Further examples are shown in Fig. 7(a). 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 a daunting task to directly estimate all the physical variables affecting spot formation. Therefore, our model is built using a statistical analysis of the resulting spot appearance, trained from examples. The model has two main parts, the first representing valid variation in spot shape, and the other pertaining to the diffusion process that forms spot appearance. In the following subsections we briefly describe the Gaussian and diffusion models.

### 1.1 Gaussian model

The most common protein spot model is based on the 2-D Gaussian function:

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

where \(B\) is background intensity and \(I\) is spot intensity, \(x_0\) and \(y_0\) control spot location and \(\sigma_x\) and \(\sigma_y\) control the spread of the Gaussian independently in \(x\) and \(y\) directions. Protein spot formation is a diffusion process. Under ideal conditions, the bivariate Gaussian model represents diffusion of an initial concentration focused at a single point in two independent directions.

### 1.2 Diffusion model

In practical situations, the assumptions that must hold for a Gaussian model to accurately represent protein spots are often broken. Bettens [17, 18] addressed this with a more detailed theoretical model of anisotropic diffusion. Again, the model assumes perfect diffusion characteristics. In this case the initial concentration is assumed to be uniformly distributed across a circular disk, leading to the formula:

\[
S(x,y) = B + \frac{C_0}{2}\left[\operatorname{erf}\left(\frac{a' + r'}{2}\right) + \operatorname{erf}\left(\frac{a' - r'}{2}\right)\right] + \\
+ \frac{C_0}{r'\sqrt{\pi}} \left[\exp\left(-\left(\frac{a' + r'}{2}\right)^2\right) - \exp\left(-\left(\frac{a' - r'}{2}\right)^2\right)\right]
\]

with

\[
r' = \sqrt{\frac{(x-x_0)^2}{D_x} + \frac{(y-y_0)^2}{D_y}}
\]

where \(B\) 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.
Note that as \( a' \to 0 \), Eq. (2) reduces to the bivariate Gaussian. The model becomes more accurate than a Gaussian model when the initial single point assumption is invalid. This is most likely for proteins with relatively high concentrations, which tend to appear in gel images as spots with large volumes exhibiting a ‘flattened’ profile.

2 Materials and methods

2.1 Statistical model

The diffusion model of spot formation is combined with a model of spot shape defined by a training set. The shape representation is compact, yet sufficiently comprehensive to represent the full range of observed spot shapes.

2.1.1 Modelling shape

Methods of representing shape variation have received much attention in machine vision in the past [15, 19–21]. Cootes et al. [12] introduced one of the most widely used techniques called point distribution models (PDMs). A PDM represents the statistics of the observed variation in a training set of shapes, and is constructed in three steps: first parameterizing the shapes by placing landmark points on object boundaries in a training set of images, then aligning the landmarks, and finally analyzing the remaining variation amongst the aligned training data.

2.1.1.1 Landmarking the training set

The landmark points provide a vector which represents the shape of a spot: \( \mathbf{x} = (x_1, x_2, x_3 \ldots, x_n, y_1, y_2, \ldots, y_n)\). Details of how we determine the landmark positions are described in Section 2.2. For the moment, note that we use 25 points to represent the spot shape.

2.1.1.2 Aligning the landmarks

As we wish to model shape variation, it is first necessary to remove other sources of variation. Spot shapes are aligned with respect to their centre of gravity and scale. In the general case of shape modelling the Procrustes alignment method [22] is commonly applied to exclude orientation variation (for details see [12]). In the case of gel spots, we already incorporate orientation into the landmarking process so this step is unnecessary.

2.1.3 Modelling the shape variation

The shapes in the training set are represented by a set of aligned shape vectors, \( \mathbf{x}_i \) with dimensionality \( 2n \) (in our case 50). The number of degrees of freedom with which the shapes can vary is typically much less than \( 2n \). This is because the variation in landmark position between examples is usually highly correlated. PDMs use principal component analysis (PCA) to capture these correlations and therefore reduce the number parameters required to represent the shape. The approach is as follows. The \( 2n \times 2n \) covariance matrix, \( \mathbf{S} \), of the data is:

\[
\mathbf{S} = \frac{1}{s-1} \sum_{i=1}^{s} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T
\]

where \( \bar{\mathbf{x}} \) is the mean of the vectors. The aligned training data, \( \mathbf{x}_i \), forms a cloud of points around \( \bar{\mathbf{x}} \) in a \( 2n \)-D space. The eigenvectors, \( \mathbf{p}_j \), and corresponding eigenvalues, \( \lambda_j (j = 1, \ldots, 2n) \), of \( \mathbf{S} \) represent a set of orthogonal axes which are aligned with the principal modes of variation of the cloud. The eigenvectors corresponding to the largest eigenvalues represent the most significant modes along which the shape can vary. \( \lambda_j \) gives the variance along the \( j \)th component. Most of the shape variation can be represented by selecting a smaller number, \( n_s < 2n \), of these axes which explain a large proportion of variation. Often \( n_s \) is chosen so the selected axes, or modes, explain at least, say, 95% of the variance exhibited in the training set.

Neglecting any alignment steps, any shape, \( \mathbf{x}_i \), in the training set can then be approximated by a weighted sum of the first \( n_s \) eigenvectors and the mean shape:

\[
\mathbf{x}_i \approx \bar{\mathbf{x}} + \mathbf{P} \mathbf{b}_s
\]

where \( \mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n) \) is the matrix of the first \( n_s \) eigenvectors, and \( \mathbf{b}_s \) is an \( n_s \) dimensional vector of weights, normally referred to as shape parameters. PDMs are generative models. New examples can be constructed by varying the values of the shape parameters, \( \mathbf{b}_s \), in Eq. (4). The shape parameters, \( \mathbf{b}_s \), which best match the model to a particular shape vector, \( \mathbf{x}_i \), can be calculated as follows:

\[
\mathbf{b}_s = \mathbf{P}^T (\mathbf{x}_i - \bar{\mathbf{x}})
\]

2.1.2 Modelling appearance

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 bivariate Gaussian kernel giving a final model with the form:

\[ S(x,y) = B + F(x + Pb) \cdot G(x,y,\sigma_x,\sigma_y, I) \]  

(6)

where \( \cdot \) is the image convolution operator,

\[ G(x,y,\sigma_x,\sigma_y, I) = I \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2}\right) \exp\left(-\frac{(y-y_0)^2}{2\sigma_y^2}\right) \]  

(7)

and \( F(x + Pb) \) is an image with value 1 within the shape and 0 outside. We define our model using the parameter vector \( p = (B, B, x_0, y_0, \sigma_x, \sigma_y, s, b_0) \), where \( B \) is an additive background term, \( I \) is spot intensity, \( x_0 \) and \( y_0 \) control location, \( \sigma_x \) and \( \sigma_y \) 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 \( b_0 \) is a vector of shape parameters.

Figure 2 shows an example of the full spot model. A flat spot shape generated using the statistical shape model is convolved with a bivariate Gaussian function resulting in an irregularly shaped ‘flat-top’ protein spot. This model is equivalent to the bivariate Gaussian when \( s \to 0 \), and is equivalent to the diffusion model when the shape parameters, \( b_0 \), represent an elliptical shape.

### 2.2 Building a spot model

Section 2.1 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 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 around 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.

#### 2.2.1 Automatic generation of training shape vectors

Raw spot boundaries are obtained by thresholding the Laplacian of Gaussian transform of the training gel images (Gaussian \( \sigma = 5 \)). The resulting boundaries are smoothed using a Fourier shape descriptor [15] resulting in a parameterization of the spot shape by the Fourier coefficients. The Fourier coefficients represent spot shapes using five harmonics. Spot appearance is modelled by convolving this smoothed shape with a Gaussian kernel (Eq. (6)). The parameters of the joint model are then optimized to improve the fit to the original image data using a Levenberg-Marquardt gradient descent algorithm. This provides an adjusted parameterization 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.

The PDM landmark representation is obtained from the resulting spot shapes by placing 25 evenly spaced points around the boundary. The position of the first point is the topmost intersection of the boundary with the line \( x = x_0 \), where \( x_0 \) is the \( x \) coordinate of the shape’s centre of gravity. The number of points and their placement are defined rather arbitrarily. It is, in principle, possible to optimize the landmark positions [23], but this was judged unnecessary for this study, as the spot shapes are consistently smooth and relatively featureless.

#### 2.2.2 Robust model building

Automatic generation of training shapes carries the danger that incorrect shapes may be included in the model. These shapes may be the result of incorrect data-driven segmentation or the inclusion of overlapping spots as
single objects. 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 component analysis [24] in the model building. We expect the number of incorrect shapes to be small, 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 Fig. 3. The figure shows two PDMs, one built using standard PCA (Fig. 3(a)) and one built using robust PCA (Fig. 3(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 ten modes, whereas the robust model requires only six modes. The contribution of each mode to the total variance of the training set is shown for each model. The centre of each row shows the mean

![Figure 3](image1.png)  
**Figure 3.** Modes of variation of PDM models. The proportion of total model variance each mode represents is given. Each row is formed by varying the appropriate shape parameter ± 2σs by from the model mean, whilst keeping all other parameters fixed. (a) The first six of 10 modes PDM built using standard PCA, retaining 99% of total training set variance. (b) All six modes of a PDM built using robust PCA. Both models were trained with the same data.

model shape; on either side of the mean are shown the change in shape obtained by varying the corresponding shape parameter $b_k$ by ± 2σs about the mean. 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’. The first robust mode corresponds to mode 2 of the standard model, and represents shape ‘skew’, although with a somewhat less extreme trend than the standard model. Aspect ratio variation is represented by robust mode 2. The rest of the robust modes represent less significant shape variation, including some squaring-off and trends towards triangular spot shape. Figure 4 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.

![Figure 4](image2.png)  
**Figure 4.** Four examples of shapes that have been downweighted by robust PCA. Each shape is superimposed over the image patch used in its generation.

### 2.3 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 [18] (Fig. 5). Each of the spot models was fitted to each spot region using a Levenberg-Marquardt nonlinear optimization algorithm to determine the best model parameters, minimizing the following residual:
Figure 5. Example training images with watershed boundaries. (a) A silver stained image with 403 delineated fitting regions, downloaded from http://www.2dgels.com/benchmark/. (b) A fluorescent dye image with 573 fitting regions.

$$r = \arg \min_p \left\{ \frac{\sum_{x,y \in R} (S(x,y|p) - l(x,y))^2}{n_R (l_{\text{max}}^R - l_{\text{min}}^R)^2} \right\}$$ (8)

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, $l(x,y)$ are image values, $S(x,y|p)$ are the model values given the parameter vectors $p$ (Section 2.1.2), $l_{\text{max}}^R$, $l_{\text{min}}^R$ 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 normalized with respect to the intensity of the spot (which we have approximated as $l_{\text{max}}^R$, $l_{\text{min}}^R$) 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 shown in Fig. 5(a) and 573 spots shown in Fig. 5(b). Fig. 5(a) is a section from a silver stained E. coli gel under standard conditions. (The image is available for download at http://www.2dgels.com/benchmark/.) The section is 375 × 228 pixels in size with a bit depth of 8. Fig. 5(b) is a section from a gel stained with a fluorescent dye. The section is 2896 × 2485 in size with a bit depth of 24. The silver image contains many saturated and overlapping spots, whereas the fluorescent image contains a much higher proportion of regular Gaussian-like spots with fewer saturated or overlapping spots.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Silver $\bar{r}$</th>
<th>Fluorescent $\bar{r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$8.3 \times 10^{-3}$</td>
<td>$5.11 \times 10^{-3}$</td>
</tr>
<tr>
<td>Diffusion</td>
<td>$7.83 \times 10^{-3}$</td>
<td>$4.94 \times 10^{-3}$</td>
</tr>
<tr>
<td>Statistical</td>
<td>$7.49 \times 10^{-3}$</td>
<td>$3.63 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

3. Results and discussion

#### 3.1 Accuracy

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 1 contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume. For the silver image, each group contains 40 spots (with 43 in group 10); for the fluorescent image, each group contains 57 spots (with 60 in group 10). In our evaluation images, spots correspond to dark regions, so volume was defined as the sum of the inverse pixel intensities over the watershed region:

$$V = \sum_{x,y \in R} (1 - l(x,y))$$ (9)

where spot intensity $l(x,y)$ is in the range (0, 1). Figure 6 shows that, 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 groups 9 and 10, with highest spot volume, statistical models resulted in average decrease in fitting residuals of 33% and 23.9% over the diffusion and Gaussian models respectively for the silver image, with 57.9% and 57.3% improvements for the fluorescent image. 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. The statistical model is able to fit to these subtle spot shape changes in the higher resolution fluorescent image. This trend is not visible in the silver image data and this may be because the lower resolution of the silver image data introduces local minima into the model fitting search space, which prevents full convergence. Under-converged results will not show a quantitative difference between the models' final residuals for small to medium spots.

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.

### 3.2 Specificity

The results show that our statistical model fits more closely to the image data than the other models. 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 decrease in model specificity. Both images contain watershed fitting regions with multiple spots. A specific model should not represent these regions well. We have argued that the images also contain regions with singular, but irregularly shaped spots. Figure 7 shows five examples of regions 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. On average, the residual of the statistical model is 65% and 63.9% lower than the fits of the Gaussian and diffusion models respectively. The fits of all models to multispot regions are visually poor (Fig. 7(b)). Here, the decrease in average residual achieved by the statistical model is 16.9% and 1.7% compared to Gaussian and diffusion models respectively. Clearly, the statistical model improves the fit for single spots significantly more than for multiple spot regions. The careful training of the model gives a representation that is specific to single spots, and therefore cannot 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. For the statistical model, four of the five single spot fits have lower residuals than all of the multipot regions. For the Gaussian model, only three single spots have a residual that is lower than all the Gaussian multispot fits. For the diffusion only two single spot fits are lower than all multipot regions. This example shows that in general it is not possible to set a single threshold on model fit residual that will identify all fitting regions containing multiple spots. However, a model that is specific to the observed appearance of gel spots, such as our statistical shape model, will fit more closely to genuine single spots than a more general heuristic model, such as a simple Gaussian function. A specific model also fits poorly to invalid data, increasing the likelihood of detecting invalid model fits.

These results suggest that the statistical model can improve the likelihood of detecting erroneously fitted models. However the figure shows only a small set of example spots. It is necessary to view the full set of residuals to determine whether this trend is genuine. Figure 8 shows reconstructions of the silver image (Fig. 5(a)) using the Gaussian, diffusion and statistical models. The images have been constructed using model values within the fitting regions (as defined by the watersheds in Fig. 5) and have been filled by interpolation in the other areas of the image. The images are displayed together with a map of the model fit residual for each fitting region. These error maps have been constructed by setting each pixel with each fitting region with the value of its associated residual after model fitting. The error maps give a visual impres-
sion of the spatial distribution of residual values. High values (light pixels) in these images indicate poor model fits. The appearance of the statistical reconstruction (Fig. 8(c)) away from very badly fitted regions (multiple spot regions) is rather more convincing than the reconstructions of the Gaussian and diffusion models (Figs. 8(a) and (b)). The pattern of high residuals in the error maps for each of the models is similar. If a threshold value were chosen for each model to discriminate regions where the model had been incorrectly fitted to multiple spots, approximately the same regions would be identified, regardless of the model used. However, the statistical model can better represent irregular single spots leading to significantly lower fitting residuals in these regions. A Gaussian model would result in high fit residuals for these single spots and thresholding may erroneously identify the regions as containing multiple spots. An area of each image in Fig. 8 has been highlighted with an ellipse. The highlighted area contains many spots which have been adequately separated into fitting regions by the original watershed process (see Fig. 5(a)). The difference between the Gaussian and statistical reconstruction images is visually apparent as a general ‘sharpening’ of the highlighted area, together with some differences in the shapes of some spots. The Gaussian reconstruction of the highlighted area contains several regions with relatively high residual and many regions with moderate residual values.

Setting a sensible threshold on fit residual for the image would probably identify several multiple spot regions in this area. The same highlighted area in the diffusion reconstruction has generally improved fits for all spots, with some substantial reductions in residual. A threshold similar to that chosen for the Gaussian model would identify fewer multi-spot regions in the area. The same highlighted area in the statistical reconstruction contains only one region of high residual, which is present in both other reconstructions, indicating that only this spot should be classified as a multiple spot region. In this case, the high residual region corresponds to a relatively low volume spot group. Fit residuals in other regions have been substantially reduced. The statistical model fits more accurately to genuine single spots, reducing their residuals significantly and therefore reducing the likelihood of them being identified as multiple spot regions. This in turn increases model specificity and improves the ability to discriminate between single and multiple spot regions.

For statistical models, specificity is entirely determined by the training data and model generation process. The results of this evaluation show that our automatic model building scheme (Section 2.2) retains specificity that is at least as high as and in general better than that of Gaussian and diffusion models whilst increasing quantification accuracy.
4 Concluding remarks

In this paper, we have introduced a new statistical model of spot appearance. This model is both flexible and specific enough to represent the true range of protein spot appearance found in complex 2-DE gel images. The model provides more accurate descriptions of irregularly-shaped single spots without losing the specificity to distinguish multiple spot groups. Furthermore, the need to develop a sophisticated theoretical model of the physical processes driving irregular spot formation has been circumvented by learning the resulting shape variation in a statistical manner.

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

Received September 3, 2002

5 References


Figure 8. Reconstructed synthetic images and error maps corresponding to Fig. 5(a). (a) Gaussian protein spot model, (b) diffusion protein spot model and (c) statistical protein spot model. Values outside fitting regions have been filled by interpolation. Error maps are generated by filling fitting regions with corresponding residual value. An area of difference between the three images has been highlighted with a circle in each image.


