INFORMATICA, 2010, Vol. 21, No. 3, 409–424

Treatment of Over-Saturated Protein Spots in Two-Dimensional Electrophoresis Gel Images

Artūras SERACKIS, Dalius NAVAKAUSKAS
Department of Electronics Systems, VGTU
Naugarduko 41-413, 422, LT-03227 Vilnius, Lithuania
e-mail: arturas.serackis@el.vgtu.lt; dalius.navakauskas@el.vgtu.lt

Received: November 2009; accepted: April 2010

Abstract. The paper addresses the over-saturated protein spot detection and extraction problem in two-dimensional electrophoresis gel images. The effective technique for detection and reconstruction of over-saturated protein spots is proposed. The paper presents: an algorithm of the median filter mask adaptation for initial filtering of gel image; the models of over-saturation used for gel image analysis; several models of protein spots used for reconstruction; technique of the automatic over-saturated protein spot search and reconstruction. Experimental investigation confirms that proposed search technique lets to find up to 96% of over-saturated protein spots. Moreover the proposed flexible protein spot shape models for reconstruction are faster and more accurate in comparison to the flexible diffusion model.

Keywords: two-dimensional electrophoresis, protein spot search, image processing, image reconstruction, modelling.

1. Introduction

The health state of living organisms has a strong mutual dependency with proteins and their change in time. In proteomics, process of two-dimensional electrophoresis (2DE) is used for protein separation in the gel according to their isoelectric point (pI) and molecular mass (MM). This enables the experts (specially trained physicians and bioengineers) to analyse health state of the patient (Iwadate, 2008; Lopez and Bermudez-Crespo, 2007). The detected changes of the proteins in the samples, used for 2DE, make it possible to determine the influence of the applied medical treatment to the patient after some period of time. An automated protein spot detection and parametrisation is able to help the expert to make comparisons of two different 2DE gels (2DEG) easier and quickly identify the type and amount of changed proteins. Number of 2DEG analysis systems (Goldfarb, 2007; Srinark and Kambhamettu, 2008) uses image processing techniques for protein spot detection and parametrisation in the scanned gel images (see Fig. 1). Though the fully automatic protein spot detection and parametrisation is still not possible.

The 2DEG analysis systems are not able to deal with specific protein spot distortions found in the gel images. The over-saturated protein spots prevent proper segmentation of the 2DEG images. Automatic detection and reconstruction of the over-saturated protein

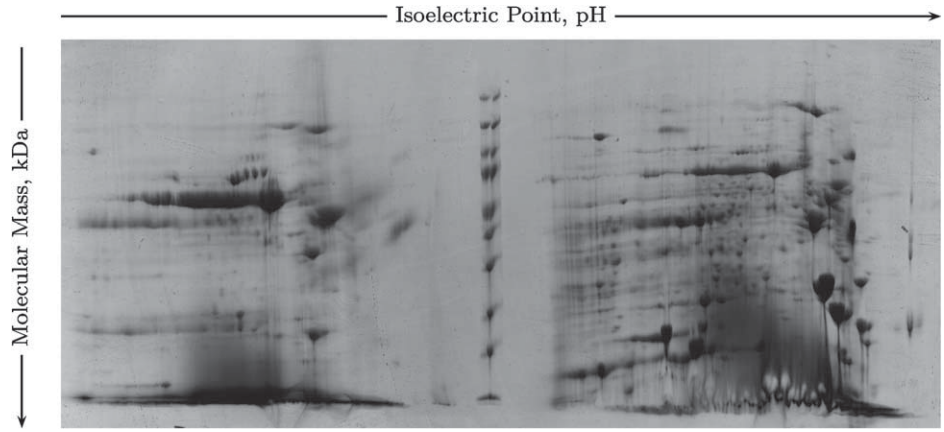


Fig. 1. An example of two-dimensional electrophoresis gel image.

spots is impossible using available protein analysis software. Usually the expensive 2DE procedure has to be repeated.

About 10,000 various proteins may appear in one 2DE gel of size 40×30 cm. Analysis and comparison of such gels is complicated and time consuming. Automatic 2DEG analysis systems can be made by combining computer performance and modern digital image processing techniques (Sutiene *et al.*, 2010). The result of 2DE process has to be scanned with high sensitivity scanner for the further image analysis. Various artifacts, donuts, scratches are often present in 2DEG images. In most cases for the same sample it is not possible to have several 2DEG due to expensivity of the ingredients used. The image processing techniques, applied for the scanned gel, has to be flexible and tunable to cope with the influence of the all distortions present in the image (Wu *et al.*, 2009).

Number of software tools for 2DEG image analysis is powerful and precise for detection and parametrisation of the undistorted protein spots. Because of the unavoidable distortions, present in the 2DEG images, those programs works in semiautomatic mode (Campostrini *et al.*, 2004; Pietrogrande *et al.*, 2003). To fully automate the 2DEG image analysis process, the new methods for distorted protein spot analysis has to be developed (Serackis, 2008).

Here we present the new technique for the search and reconstruction of over-saturated protein spots in 2DEG images. Firstly we discuss main preprocessing steps needed to be taken. Then we introduce two mathematical models for search and other models for reconstruction of over-saturated protein spots. Later we in details present developed technique for treatment of over-saturated protein spots. Finally, results of experimental study of newly technique are shown and discussed.

2. Preprocessing of the 2D Electrophoresis Gel Images

Protein spots in 2DEG images has to be detected and parametrised. For parametrisation of the protein spots the 2DEG images are segmented into regions having one possible

protein spot in each. Most widely used 2DEG image segmentation technique is based on regional minima search and watershed transformation (Matuzevičius and Navakauskas, 2005).

Each scanned 2DEG image has an individual intensity level of background, consisting of high intensity pixels surrounding proteins spots. The minor change of background intensities brings no useful information about protein spots and may complicate the gel image segmentation procedure. The simplest solution for the background extraction is the threshold filtering, succesfully used in various applications.

The noise is also always present in the 2DEG images. Background extraction techniques are not able to remove noise influence to the protein spots and speckles in the background. The noise in 2DEG images can be filtered using local or double threshold function (Matuzevičius and Navakauskas, 2005). The low and high frequency noise picked up by image scanner could be filtered using low-pass and high-pass two-dimensional filters. The protein spots with similar frequency characteristics could be filtered, too. The most popular for such noise cancellation in 2DEG images are linear averaging (Grigaitis *et al.*, 2007) or nonlinear median filter of size $(2k+1) \times (2l+1)$:

$$\mathbf{I}_{\mathrm{F}}(i_{\mathrm{I}}, j_{\mathrm{I}}; \mathbf{M}) = \mathrm{median}\left\{\mathbf{I}_{(i_{\mathrm{I}}-k, j_{\mathrm{I}}-l)}, \ldots, \mathbf{I}_{(i_{\mathrm{I}}, j_{\mathrm{I}})}, \ldots, \mathbf{I}_{(i_{\mathrm{I}}+k, j_{\mathrm{I}}+l)}\right\}, \tag{1}$$

here $\mathbf{M}$ – median filter mask; $(i_{\mathrm{I}}, j_{\mathrm{I}})$ – indices of the image matrix; k and l determine the size of filter mask and usually $k = l$.

The size of the filter mask has to be properly chosen. Too large size filter mask may lead to the overlapping of neighbouring protein spots in 2DEG image, e.g., for image with resolution 2165×1375 pixels, filter mask of size 24×24 pixels gives the overlap of neighbouring protein spots. Experimental investigation confirmed that the filter mask size has to be chosen twice larger than the noise element to be filtered.

In the proposed treatment technique the size of median filter mask for 2DEG image can be chosen using one of the two ways:

- by the analysis of mean-square error resulting from the application of median filter with different sizes of mask;
- by the image analysis that estimates the size of the biggest noise element in the image.

In the first case, the mean-square error is computed using cross-validation method (Reeves, 1995). This way of filter mask selection is suitable for comparison of several 2DEG images when the set of already detected protein spots is known.

During visual analysis of 2DEG images (cf. Fig. 1) one can notice the noise, which results in peaks of intensities with high gradient change. Those features are not typical for protein spots and have to be filtered as high frequency noise in the image.

The filter mask selection requires an algorithm for detection of high frequency noise elements in 2DEG image. The shape analysis of noise elements gives possibility to choose a filter mask of size twice larger than the highest noise element detected. The new algorithm, based on analysis of image gradient changes, for filter mask size estimation is proposed (Fig. 2).

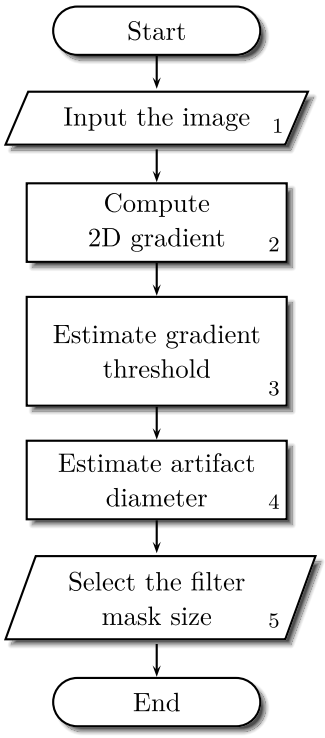


Fig. 2. The algorithm for the selection of median filter mask size.

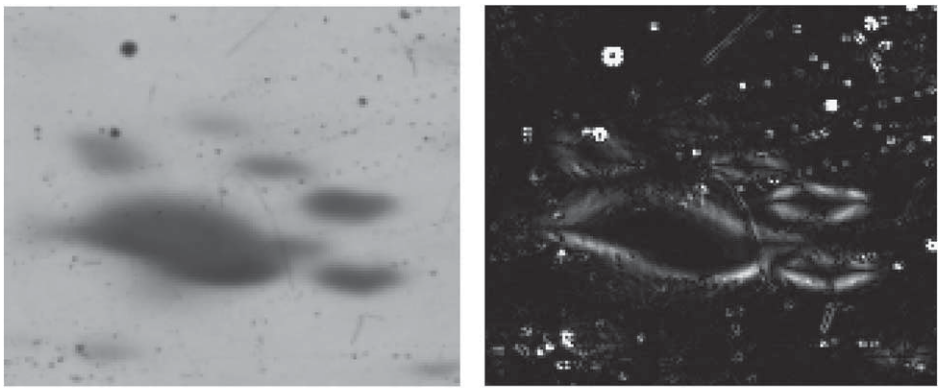
The analysis of the 2DEG image is performed only on intensities of the pixels, thus no colour information has to be detected and stored after input of an image.

The changes of the pixel intensities among axes of isoelectric point, x_{I} and molecular mass, y_{I} are analysed by the use of gradient function:

$$\nabla \mathbf{I} = \frac{\partial \mathbf{I}}{\partial x_{\mathrm{I}}} + \frac{\partial \mathbf{I}}{\partial y_{\mathrm{I}}}. \tag{2}$$

The aim of proposed algorithm is to detect sharp intensity changes – not typical for the protein spots. The application of gradient function for the image $\mathbf{I}$ returns matrix of the same size as the image. During the estimation of gradient matrix, the changes of pixel intensities result in positive and negative values, dependent on the direction of computation. For further analysis absolute values of gradient matrix are considered. The analysis of two-dimensional gradient matrix is performed. This matrix is estimated by multiplication of two gradient functions: $\nabla \mathbf{I}_{\mathrm{x}}$ and $\nabla \mathbf{I}_{\mathrm{y}}$. The sharp changes of intensities are expressed by gradient matrix elements with extremely high values.

The illustration of the contours of sharpest elements in gradient matrix is shown in Fig. 3. If the noise elements are in shape of ellipse, the contour of the artifact shape is closed. The area of the largest noise element is computed and chosen as the basis for the selection of filter mask size.



(a) Initial 2DEG image view (b) Gradient matrix view

Fig. 3. Illustration of filter mask selection procedure.

As it is shown in Fig. 3(b), the contours of the sharp protein spots are also noticeable in the gradient matrix view. However the gradient values of such contours are lower comparing to the contours of noise elements. The threshold for elimination of sharp protein spots from further analysis has to be chosen.

Mostly the sharp noise elements appear in cylindrical shape and the contour of such shape is approximate circle. The size of the noise element is computed as the diameter of the circle:

$$d_{\mathrm{tr}} = 2\sqrt{S_{\mathrm{tr}}/\pi}, \tag{3}$$

here S_{tr} – the area of the largest noise element in the image.

After the application of the threshold function to the gradient matrix the centre of the noise element disappears, see Fig. 3(b). The additional morphological operations are applied to remove empty areas within the contour of noise elements.

The elimination of artifacts in more complex shapes is still not solved in the analysis of 2DEG images. Various image analysis techniques, proposed in the literature (Guzaitis and Verikas, 2008). There are many recommendations for proper 2DE gel drying and handling to eliminate possible dust marks and scratches in the gel. If such artifacts still appear in the 2DEG images, the expensive experiment has to be repeated.

3. Models of Over-Saturated Protein Spots

The over-saturation of the spots are also observed in micro-array images (Ekstrom *et al.*, 2004; Kim *et al.*, 2007; Ridgway and Godsill, 2006). Several shape models are proposed in literature for spot modelling in micro-array images: cylindrical, symmetrical Gaussian shape model, two Gaussian shapes difference model and hyperbolic polynomial model.

The shape model for over-saturated gene spot in micro-array image is expressed as a difference of two Gaussian shapes:

$$\mathbf{S}(l_{\mathrm{s}}) = \frac{1}{1-s}\left[\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(l_{\mathrm{s}}-c_{\mathrm{c}})^2}{2\sigma_1^2}\right) - \frac{s}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(l_{\mathrm{s}}-c_{\mathrm{c}})^2}{2\sigma_2^2}\right)\right], \tag{4}$$

here l_{s} – Euclidean distance to the centre of the spot; s – scale factor, $s \in [0,1]$; σ_1 and σ_2 – the standard deviation for the first and second shape in the equation respectively.

This over-saturated gene spot shape model is easily adopted for non-saturated spot modelling by setting the scale factor $s = 0$. Comparing it to the symmetrical Gaussian shape model, the over-saturated spot shape model adds two additional parameters: scale factor s and standard deviation σ_2. The over-saturation distortion is modelled by introducing an inverted Gaussian shape to the model. Such shape also can be used for the search of over-saturation in 2DEG image.

Over-saturated protein spot distortion model based on Gaussian function – *Gaussian protein over-saturation model*, used for cross-correlation as matrix with indices x and y, is expressed:

$$\mathbf{S}_{\mathrm{NG}}(x,y) = \frac{m}{\sqrt{2\pi(m\sigma)^2}}\exp\left(-\frac{(x-x_{\mathrm{c}})^2}{2(m\sigma)^2}\right)\exp\left(-\frac{(y-y_{\mathrm{c}})^2}{2(m\sigma)^2}\right), \tag{5}$$

here $c_{\mathrm{c}}(x_{\mathrm{c}}, y_{\mathrm{c}})$ – centre of the protein spot; σ^2 – variance; m – scale factor, $m \in [0,1]$.

Over-saturated protein spot distortion model based on Mexican Hat shape – *Mexican Hat protein over-saturation model*, used for cross-correlation as matrix with indices x and y, is expressed:

$$\mathbf{S}_{\mathrm{MS}}(x,y) = m\frac{\sin(\sqrt{x^2+y^2})}{\sqrt{x^2+y^2}}, \tag{6}$$

here m – scale factor, $m \in [0,1]$.

In order to reconstruct the over-saturated protein spot shape, flexible mathematical models has to be used. The shape models for saturated protein spots have to be: (1) flexible to match undistorted and saturated protein spots; (2) able to acquire protein spots of varying size; (3) able to deal with asymmetric protein spots; (4) resistant to the non-protein spots, i.e., by giving bad approximation results. According to the stated requirements four protein spot shape models for reconstruction were proposed and in the following presented.

Anisotropic Bell spot shape model, used for parametrisation of protein spot in 2DEG image segment with indices of the image pixels set by x and y, is expressed:

$$\mathbf{S}_{\mathrm{2VF}}(x,y) = B + \frac{I_{\mathrm{s}}}{1+\left|\frac{x-x_{\mathrm{c}}}{l_{\mathrm{x}}}\right|^{2q_{\mathrm{x}}}} \cdot \frac{1}{1+\left|\frac{y-y_{\mathrm{c}}}{l_{\mathrm{y}}}\right|^{2q_{\mathrm{y}}}}, \tag{7}$$

here B – background intensity of the 2DEG image; I_{s} – intensity of the spot; $l_{\mathrm{s}}(l_{\mathrm{x}}, l_{\mathrm{y}})$ – distance to the centre of the spot; q_{x} and q_{y} – inclination of the spot shape slope; $c_{\mathrm{c}}(x_{\mathrm{c}}, y_{\mathrm{c}})$ – centre of the protein spot.

Four splines spot shape model, used for parametrisation of protein spot in 2DEG image segment with indices of the image pixels set by x and y, is expressed:

$$S_{4\mathrm{S}}(x,y) = B + I_{\mathrm{s}} \cdot S_{\mathrm{E}} \cdot S_{\mathrm{W}} \cdot S_{\mathrm{N}} \cdot S_{\mathrm{S}}, \tag{8a}$$

$$S_{\mathrm{E}} = \begin{cases} 0, & \text{for } x \leqslant a; \\ \dfrac{2(x-a_{\mathrm{x}})^2}{(x_{\mathrm{c}}-b_{\mathrm{x}}-a_{\mathrm{x}})^2}, & \text{for } a_{\mathrm{x}} < x \leqslant (a_{\mathrm{x}}+x_{\mathrm{c}}-b_{\mathrm{x}})/2; \\ 1-\dfrac{2(x_{\mathrm{c}}-b_{\mathrm{x}}-x)^2}{(x_{\mathrm{c}}-b_{\mathrm{x}}-a_{\mathrm{x}})^2}, & \text{for } (a_{\mathrm{x}}+x_{\mathrm{c}}-b_{\mathrm{x}}/2) < x \leqslant x_{\mathrm{c}}-b_{\mathrm{x}}; \\ 1, & \text{for } x > x_{\mathrm{c}}-b_{\mathrm{x}}; \end{cases} \tag{8b}$$

$$S_{\mathrm{W}} = \begin{cases} 1, & \text{for } x_{\mathrm{c}}+c_{\mathrm{x}} < x \leqslant d_{\mathrm{x}}; \\ 1-\dfrac{2(x_{\mathrm{c}}-c_{\mathrm{x}}-x)^2}{(x_{\mathrm{c}}+c_{\mathrm{x}}-d_{\mathrm{x}})^2}, & \text{for } d_{\mathrm{x}} < x \leqslant (d_{\mathrm{x}}+x_{\mathrm{c}}+c_{\mathrm{x}})/2; \\ \dfrac{2(x-d_{\mathrm{x}})^2}{(x_{\mathrm{c}}+c_{\mathrm{x}}-d_{\mathrm{x}})^2}, & \text{for } (d_{\mathrm{x}}+x_{\mathrm{c}}+c_{\mathrm{x}})/2 < x \leqslant x_{\mathrm{c}}+c_{\mathrm{x}}; \\ 1, & \text{for } x > x_{\mathrm{c}}+c_{\mathrm{x}}; \end{cases} \tag{8c}$$

$$S_{\mathrm{N}} = \begin{cases} 0, & \text{for } y \leqslant a_{\mathrm{y}}; \\ \dfrac{2(y-a_{\mathrm{y}})^2}{(y_{\mathrm{c}}-b_{\mathrm{y}}-a_{\mathrm{y}})^2}, & \text{for } a_{\mathrm{y}} < y \leqslant (a_{\mathrm{y}}+y_{\mathrm{c}}-b_{\mathrm{y}})/2; \\ 1-\dfrac{2(y_{\mathrm{c}}-b_{\mathrm{y}}-y)^2}{(y_{\mathrm{c}}-b_{\mathrm{y}}-a_{\mathrm{y}})^2}, & \text{for } (a_{\mathrm{y}}+y_{\mathrm{c}}-b_{\mathrm{y}})/2 < y \leqslant y_{\mathrm{c}}-b_{\mathrm{y}}; \\ 1, & \text{for } y > y_{\mathrm{c}}-b_{\mathrm{y}}; \end{cases} \tag{8d}$$

$$S_{\mathrm{S}} = \begin{cases} 1, & \text{for } y \leqslant d_{\mathrm{y}}; \\ 1-\dfrac{2(y_{\mathrm{c}}+c_{\mathrm{y}}-y)^2}{(y_{\mathrm{c}}+c_{\mathrm{y}}-d_{\mathrm{y}})^2}, & \text{for } d_{\mathrm{y}} < y \leqslant (d_{\mathrm{y}}+y_{\mathrm{c}}+c_{\mathrm{y}})/2; \\ \dfrac{2(y-d_{\mathrm{y}})^2}{(y_{\mathrm{c}}+c_{\mathrm{y}}-d_{\mathrm{y}})^2}, & \text{for } (d_{\mathrm{y}}+y_{\mathrm{c}}+c_{\mathrm{y}})/2 < y \leqslant y_{\mathrm{c}}+c_{\mathrm{y}}; \\ 0, & \text{for } y > y_{\mathrm{c}}+c_{\mathrm{y}}, \end{cases} \tag{8e}$$

here B – background intensity of the 2DEG image; I_{s} – intensity of the spot; $a_{\mathrm{x}}, a_{\mathrm{y}}, d_{\mathrm{x}}$ and d_{y} – parameters forming the bottom part of the shape; $b_{\mathrm{x}}, b_{\mathrm{y}}, c_{\mathrm{x}}$ and c_{y} – parameters forming the top of the shape.

Four Gaussian functions spot shape model, used for parametrisation of protein spot in 2DEG image segment with indices of the image pixels set by x and y, is expressed:

$$\begin{aligned} \mathbf{S}_{4\mathrm{G}}(x,y) \;=\; & B + I_{\mathrm{s}}\left[\delta_{\mathrm{x}} \cdot \exp\left(\frac{(x-x_{\mathrm{c1}})^2}{-2\sigma_{\mathrm{x1}}^2}\right) + \overline{\delta_{\mathrm{x}}} \cdot \exp\left(\frac{(x-x_{\mathrm{c2}})^2}{-2\sigma_{\mathrm{x2}}^2}\right)\right] \\ & \times \left[\delta_{\mathrm{y}} \cdot \exp\left(\frac{(y-y_{\mathrm{c1}})^2}{-2\sigma_{\mathrm{y1}}^2}\right) + \overline{\delta_{\mathrm{y}}} \cdot \exp\left(\frac{(y-y_{\mathrm{c2}})^2}{-2\sigma_{\mathrm{y2}}^2}\right)\right], \end{aligned} \tag{9}$$

here B – background intensity of the 2DEG image; y_{c1}, y_{c2}, x_{c1}, x_{c2} – centre points for Gaussian functions; I_s – intensity of the spot; $\sigma^2_{x1}, \sigma^2_{x2}, \sigma^2_{y1}$ and σ^2_{y2} – variance of the Gaussian functions; the threshold functions to form the shape: $\delta_x = 1$, when $x \leqslant x_c$; $\delta_y = 1$, when $y \leqslant y_c$.

Three Gaussian functions model is a partial case of the four Gaussian functions model and is obtained by the use of one variance and single centre point for the Gaussian function along isoelectric point axis (we skip its presentation in order to save space).

4. Technique for Detection and Reconstruction of Over-Saturated Protein Spots

The proposed technique for detection and reconstruction of the over-saturated protein spots in 2DEG images consists of five stages:

1. Preprocessing of the 2DEG image.
2. Creation of the over-saturated protein spot models.
3. Detection of the over-saturated protein spots in 2DEG image.
4. Extraction of the image segments with protein spot distortions.
5. Reconstruction of the over-saturated protein spots.

Detection of the over-saturated protein spots is performed in five steps: (1) creation of distortion examples; (2) computation of the cross-correlation between the distortion example and 2DEG image; (3) analysis of the correlation matrix; (4) marking of the distorted segments; (5) revision of the results.

The examples of protein spot distortions are created using over-saturation model and scale multiplier m, in order to create several examples of distortions varying in size. The detection of the over-saturation is performed applying the cross-correlation of the 2DEG image and the over-saturation models. Image regions similar to the modelled over-saturation of the protein spot gives regional maxima in cross-correlation matrix. The regions with the cross-correlation values above defined threshold are marked as over-saturated protein spots and the reconstruction of the protein spots is performed. Two types of models were suggested for over-saturation search in the 2DEG images; Eqs. (5) and (6).

After detection of the over-saturated protein spot, the reconstruction – shape restoration – of the protein spot has to be performed in order to improve proper image segmentation and protein spot detection in 2DEG image. The position of the protein spot in 2DEG image makes possible to identify the protein according to the values of isoelectric point and molecular mass. Therefore the proper reconstruction of the over-saturated protein spot is essential. The new protein spot shape model, combined from three Gaussian shapes is chosen for reconstruction of the over-saturated protein spot.

The new algorithm for over-saturated protein spots reconstruction is proposed and presented in Fig. 4. The algorithm is adapted to the selected three Gaussian protein spot shape model. The reconstruction of the over-saturated protein spots is performed during

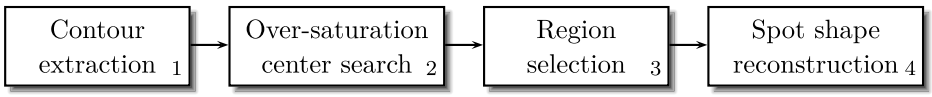


Fig. 4. Over-saturated protein spot reconstruction operation flow.

four main steps: (1) contour detection of the over-saturated protein spot; (2) computation of the spot centre; (3) selection of the analysis regions for reconstruction model parameters estimation; (4) reconstruction of the protein spot. To adjust reconstruction algorithm for other protein spot shape models, proposed in this paper, 3rd and 4th steps has to be modified.

The input data for reconstruction algorithm is filtered segment of 2DEG image $\mathbf{I}_{\mathrm{F}}$ with one over-saturated protein spot.

At the first stage of the over-saturated protein spot reconstruction algorithm the contour of over-saturation is estimated. The disjunctive boundary k_{s} between the distorted part and the rest part of the spot shape is created. The contour k_{s} in the filtered image region $\mathbf{I}_{\mathrm{F}}$ is estimated using watershed transformation based algorithm; for the example see Fig. 5(a).

All the pixels in the segmented region are marked by identification number. The region pixels values are changed to 1 and all other pixels are set to the value of 0. The received binary image is additionally processed using morphological erosion and dilation to leave the contour pixels surrounding segmented region of the over-saturated spot; see Fig. 5(b).

Each concave in the image has its own local minima $I_{\min,n}$; cf. Fig. 5(c). We assume that the centre of the over-saturated protein spot is the highest intensity regional minima found: $I_{\min} = \max\{I_{\min,1},\ I_{\min,2}, \ldots,\ I_{\min,n}\}$; cf. Fig. 5(d). Frequently there are several regional minima with intensity value of $I_{\min}$ in one image segment. To select a point as a centre of the region the additional morphological operations has to be performed. The centre of mass for set of regional minima is estimated and selected as the regional minima $I_{\min}$ of the image segment.

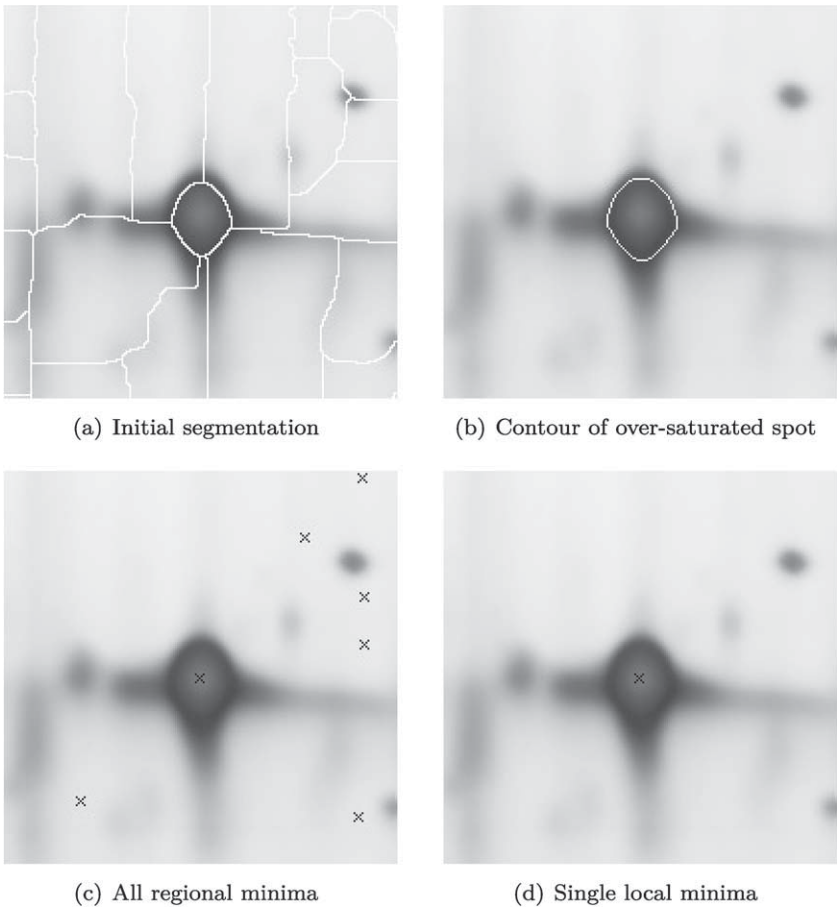
The centre of the spot also can be found by fitting an ellipse to the contour of the over-saturated spot $p(x, y)$. Ellipse fitting procedures are frequently used in image analysis tasks (Treigys *et al.*, 2008). The ellipse fitting error is computed by the use of least squares algorithm by searching the ellipse parameter vector $\mathbf{a}$. The distance between the ellipse points and the contour vector is minimised:

$$D(\mathbf{a}) = \sum_{i=1}^{N} F(\mathbf{p}_i, \mathbf{a})^2, \tag{10}$$

here $F(\mathbf{p}_i, \mathbf{a})$ – distance function between pixel array $\mathbf{p} = (x^2\ xy\ y^2\ x\ y\ 1)^{\mathrm{T}}$ and ellipse parameters vector $\mathbf{a} = (a\ b\ c\ d\ e\ f)^{\mathrm{T}}$.

The parameters of ellipse are related to the pixel coordinates by equation:

$$ax^2 + by^2 + cxy + dx + ey + f = 0. \tag{11}$$



(a) Initial segmentation (b) Contour of over-saturated spot

(c) All regional minima (d) Single local minima

Fig. 5. Illustration of the over-saturated protein spot reconstruction superimposing original image with intermediate results.

Reconstruction of the protein spot is done by the use of the protein spot shape model. The three Gaussian shape based protein spot model is used. Varying variance of Gaussian function forms a shape similar to the protein spot. Different variance values are used to form an asymmetric protein shape among the axis indicating the molecular mass value. The symmetrical Gaussian shape is used to form the shape of protein spot among the perpendicular axis.

The application of the protein spot shape model requires the shape slope gradient changes to be similar or equal to the gradient changes of protein spot slope. The parameters of the model are computed according to the four regions of the protein spot slope. The model is fitted by minimising the differences between slopes of the protein spot and spot shape model (see Fig. 6), used for reconstruction. Additionally the initial values for protein spot model parameters: initial centre $c(x_{\mathrm{c}}, y_{\mathrm{c}})$, σ_1, σ_2 and σ_3) has to be set.

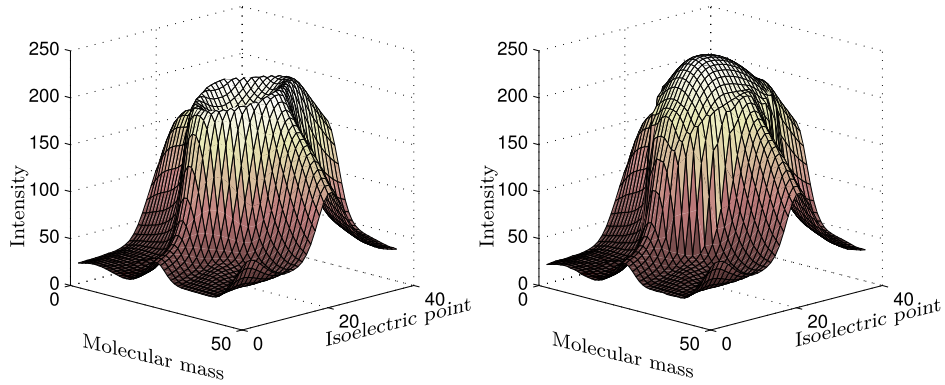


Fig. 6. Three-dimensional view of the over-saturated protein spot before and after the reconstruction.

5. Results of Experimental Study

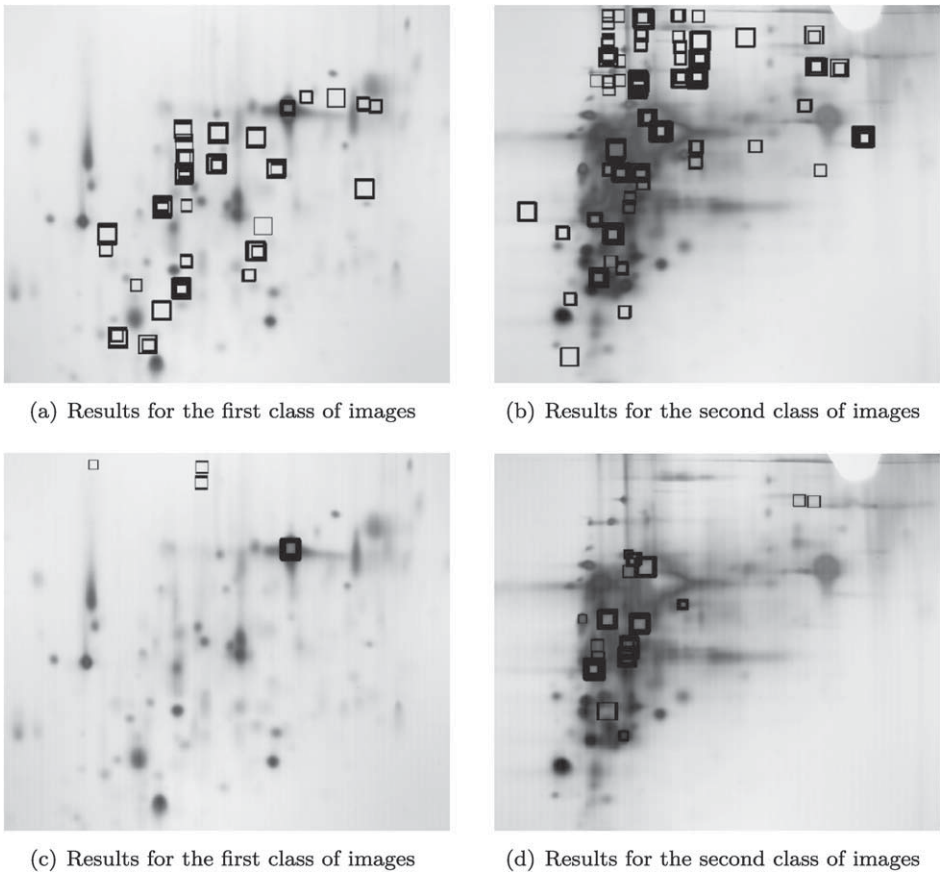
The experimental investigation of the proposed technique is performed in two stages. Gaussian and Mexican Hat based shape models are compared for over-saturated protein spot search at first stage. The models are used to create a set of protein spot over-saturation examples. The examples are used as a mask for two-dimensional cross-correlation performed on 2DEG images. The results of over-saturated protein spot search using two different shape models are compared. The over-saturation model with fewer incorrect results is selected. At the second stage of the experimental investigation, the reconstruction of over-saturated protein spots is performed. Protein spot reconstruction algorithm is tested on real scanned gel images and synthetic over-saturated protein spots.

The 2DEG images used for experimental investigation are taken from various proteomic databases with open access. All collected images are classified into two classes: 2DEG images with more than three and 2DEG images with one, two or three over-saturated spots.

The size of the protein spot over-saturation example depends on the resolution of the selected image. The resolution of the analysed 2DEG images is 606×463 pixels. The size of the biggest example used for two-dimensional cross-correlation is 60×60 pixels. Four examples for each model were created by the use of scale factor with values $m = 1$, 0.7, 0.5 and 0.3 (the resulting matrices were of size 42×42, 30×30 and 18×18).

During over-saturated protein spot search the cross-correlation matrix $\mathbf{R}_{m,n}$ for each model is received. For current set of 2DEG images threshold value of $d_{\mathrm{R}} = 0.7$ is selected to extract the regional maxima in the cross-correlation matrix. The extracted maxima are defining possible locations of the over-saturated protein spots – the 2DEG image segments similar to the modelled by the Gaussian or Mexican Hat model.

The results of the over-saturated protein spot search are shown by drawing the rectangular in the image regions with possible over-saturations (see Fig. 7). The most expected regions are marked with several rectangular due to higher values in each cross-correlation matrix received.



(a) Results for the first class of images (b) Results for the second class of images

(c) Results for the first class of images (d) Results for the second class of images

Fig. 7. Over-saturated protein spot search results using different models: (a, b) – Gaussian; (c, d) – Mexican Hat.

The experimental investigation of the over-saturation search in 2DEG images, performed using two types of artifact models, showed that using Mexican Hat model, from 15% to 25% of identified image regions were not marked as a over-saturated protein spot by the expert in the first class of 2DEG images (see Fig. 7(c)). For the second class of images, the Mexican Hat model gives from 18% to 28% incorrect results (see Fig. 7(d)). This model gave 3–5 times less wrong over-saturated spots found in 2DEG images comparing to Gaussian shape over-saturation model: from 78% to 96% incorrect results for the first class (see Fig. 7(a)) and from 75% to 85% for the second set (see Fig. 7(b)).

The experimental investigation of over-saturated protein spot reconstruction was performed on 883 spot set. 240 spots were selected from original scanned 2DEG images and other – synthetic spots (used to compute the quality of spot reconstruction). Two different protein spot shape models and fitting algorithms were used for comparison: Anisotropic Gaussian and Three Gaussian functions.

Table 1
Reconstruction results of the over-saturated spots

Algorithm	Intensity		Volume	
	Mean	Deviation	Mean	Deviation
2 Gaussian	2.22	+18.68	17.12	+2.31
		−7.96		−3.95
3 Gaussian	2.23	+18.69	16.55	+2.33
		−7.96		−4.04

Table 2
Results of approximation residual error

Shape model	Type of the protein spot					
	Saturated		Big-size		Small-size	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Diffusion	2.300	1.850	1.170	1.270	1.160	1.020
Bell	**2.030**	1.740	1.050	1.190	0.992	0.829
4 Gaussian	2.270	1.960	**0.811**	0.839	**0.824**	0.753
4 Splines	2.730	1.970	1.440	1.020	1.530	0.922

Table 3
Results of relative processing time

Shape model	Type of the protein spot					
	Saturated		Big-size		Small-size	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Diffusion	1.000	0.000	1.000	0.000	1.000	0.000
Bell	0.169	0.103	0.118	0.069	0.126	0.149
4 Gaussian	**0.111**	0.073	**0.068**	0.038	**0.057**	0.024
4 Splines	0.351	0.280	0.170	0.130	0.175	0.111

The results of the experiments are shown in Table 1. The use of three Gaussian functions protein spot shape model increases precision of the reconstruction, but not significantly (0.6%). The reconstruction of the over-saturated protein spots adds 2.23% of intensity and 16.55% of spot shape volume at an average.

Other protein spot shape models are also available for reconstruction. The additional experimental investigation was performed to compare the average model fitting time and precision of proposed protein shape models for the saturated protein spots.

The experimental investigation of the protein spot models was made comparing to the Diffusion model (Bettens *et al.*, 1997). 84 natural protein spots were selected, dividing

those into three groups:

- over-saturated protein spots ($I_{\mathrm{s}} > 160$);
- big-size protein spots ($I_{\mathrm{s}} > 160$);
- small-size protein spots ($I_{\mathrm{s}} < 100$).

The results of model fitting are shown in Table 2. In the Table 3 results of the protein spot model relative fitting time are shown.

The use of four Gaussian functions model gives the shortest protein spot processing time and fits best for undistorted protein spots.

6. Conclusions

In the paper the new technique for treatment of over-saturated protein spots in 2DEG images are presented. From 75% to 96% of over-saturated protein spots can be found by the use of proposed two distortion models. The erroneous over-saturations found during experimental investigation showed the imperfection of the Gaussian distortion model. The proposed Mexican Hat model reduces the amount of erroneous results by 3–5 times.

Anisotropic Gaussian and three Gaussian functions spot shape models were used in proposed spot reconstruction technique. The accuracy of reconstruction was verified by experimental investigation performed on synthetic protein spots. Proposed technique restores 2.23% of the spot intensity and 16.55% in average of the spot volume during reconstruction.

Three additional flexible protein spot shape models were proposed. The comparison to the flexible Diffusion model showed such advantages of the proposed models: shorter spot shape fitting time and reduced approximation residual error.

Acknowledgements

The research in part was supported by project No. T-08126 (contract No. T-111/08) from Lithuanian State Science and Studies Foundation.

References

Bettens, E., Scheunders, P., Dyck, D.V., Moens, L., Osta, P.V. (1997). Computer analysis of two-dimensional electrophoresis gels: a new segmentation and modelling algorithm. *Electrophoresis*, 18(5), 792–798.

Campostrini, N., Areces, L.B., Rappsilber, J., Pietrogrande, M.C., Dondi, F., Pastorino, F., Ponzoni, M., Righetti, P.G. (2004). Spot overlapping in two-dimensional maps: a serious problem ignored for much too long. *Proteomics*, 5(9), 2385–2395.

Ekstrom, C.T., Bak, S., Kristensen, C., Rudemo, M. (2004). Spot shape modelling and data transformations for microarrays. *Bioinformatics*, 20(14), 2270–2278.

Goldfarb, M. (2007). Computer analysis of two-dimensional gels. *Journal of Biomolecular Techniques*, 18(3), 143–146.

Grigaitis, D., Bartkutė, V., Sakalauskas, L. (2007). An optimization of system for automatic recognition of ischemic stroke areas in computed tomography images. *Informatica*, 18(4), 603–614.

Guzaitis, J., Verikas, A. (2008). An efficient technique to detect visual defects in particleboards. *Informatica*, 19(3), 363–376.

Iwadate, Y. (2008). Clinical proteomics in cancer research-promises and limitations of current two-dimensional gel electrophoresis. *Current Medicinal Chemistry*, 15(23), 2393–2400.

Kim, H.Y., Lee, S.E., Kim, M.J., Han, J.I., Kim, B.K., Lee, Y.S., Lee, Y.S., Kim, J.H. (2007). Characterization and simulation of cdna microarray spots using a novel mathematical model. *BMC Bioinformatics*, 8(1).

Lopez, J., Bermudezl-Crespo, J. (2007). Possibilities of two-dimensional gel electrophoresis in the understanding of human disease. *Current Proteomics*, 4(4), 187–197.

Matuzevičius, D., Navakauskas, D. (2005). Investigation of segmentation methods for proteomics. *Electronics and Electrical Engineering*, 7(63), 66–70.

Pietrogrande, M.C., Marchetti, N., Dondi, F., Righetti, P. G. (2003). Spot overlapping in two-dimensional polyacrylamide gel electrophoresis maps: relevance to proteomics. *Electrophoresis*, 24, 217–224.

Reeves, S. J. (1995). On the selection of median structure for image filtering. *IEEE Transactions on Circuits and Systems: Analog and Digital Signal Processing*, 42(8), 556–558.

Ridgway, G., Godsill, S. (2006). Bayesian modelling of microarray images. In: *GENSIPS '06. IEEE International Workshop on Genomic Signal Processing and Statistics*, pp. 41–42.

Serackis, A. (2008). Image reconstruction technologies for protein spot parametrisation. *Doctoral Dissertation*, Vilnius Gediminas Technical University, Vilnius.

Srinark, T., Kambhamettu, C. (2008). An image analysis suite for spot detection and spot matching in two-dimensional electrophoresis gels. *Electrophoresis*, 29(3), 706–715.

Sutiene, K., Makackas, D., Pranevicius, H. (2010). Multistage K-means clustering for scenario tree construction. *Informatica*, 21(1), 123–138.

Treigys, P., Šaltenis, V., Dzemyda, G., Barzdziukas, V., Paunksnis, A. (2008). Automated optic nerve disc parameterization. *Informatica*, 19(3), 403–420.

Wu, J., Ruan, Q., An, G. (2009) A joint-diffused inpainting model for underexposure image preserving the linear geometric structure. *Informatica*, 20(1), 151–163.

A. Serackis born 1980 in Vilnius, Lithuania. Received engineering diploma (2002), MSc degree (2004) and PhD degree (2008). Presently working as associate professor at Electronics Systems Department of Vilnius Gediminas Technical University. Research interests are image analysis, signal processing and intelligent systems.

D. Navakauskas is a professor at Electronics Systems Department of Vilnius Gediminas Technical University. He received honour diploma of radio-electronics engineer in 1992, MSc in electronics degree in 1994, doctor of electrical and electronic engineering degree in 1999, passed habilitation procedure in informatics engineering in 2005, received professor title in 2008 all at Vilnius Gediminas Technical University. He is a senior member of IEEE and currently serves as IEEE Joint SP/CIS/COM chapter chair in Lithuania. His main research interests include computational intelligence, signal and image processing, bioinformatics.

Persisotinusių baltymų pėdsakų apdorojimas dvimatės elektroforezės gelių vaizduose

Artūras SERACKIS, Dalius NAVAKAUSKAS

Straipsnyje nagrinėjama baltymų pėdsakų atpažinimo ir parametrizavimo problema dvimatės elektroforezės gelių vaizduose. Pristatomas naujas būdas baltymų persisotinimams dvimatės elektroforezės gelių vaizduose aptikti ir rekonstruoti. Siūlomą paieškos ir rekonstravimo būdą sudaro keli etapai: gelio vaizdo paruošimas taikant naują medianos filtro kaukės dydžio parinkimo algoritmą, baltymų persisotinimų paieška taikant autorių siūlomus iškraipymų modelius, automatinis persisotinusio baltymų pėdsako išskyrimas ir rekonstravimas. Straipsnyje pateikti eksperimentinio tyrimo rezultatai įrodo, kad siūlomas būdas leidžia atpažinti iki 96% baltymų persisotinimų gelių vaizduose taikant vieną iš dviejų autorių siūlomų iškraipymų modelių. Baltymų persisotinimų rekonstravimui straipsnyje siūlomi nauji baltymų pėdsakų modeliai, gebantys tiksliau ir sparčiau atstatyti baltymo pėdsako formą nei alternatyvusis difuzinis modelis.