ACTA CHROMATOGRAPHICA, NO. 15, 2005

BASELINE REDUCTION IN TWO DIMENSIONAL GEL ELECTROPHORESIS IMAGES

K. Kaczmarek[1,*], *B. Walczak*[1], *S. de Jong*[2], and *B. G. M. Vandeginste*[2]

[1]Institute of Chemistry, Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland
[2]Unilver R&D Vlaardingen, Olivier van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands

INTRODUCTION

The proteins are the controllers of all cell functions and, as such, are closely connected with many diseases and metabolic processes. Although proteins are coded by genes, it has been shown there is a poor correlation between protein and mRNA abundance [1]. For this reason a full knowledge of the genome is not sufficient for determination of the protein composition of cells [2]. This ineffectiveness of genomics caused a big shift of interest from genomics to proteomics. The main analytical tool used for protein separation in proteomics is two-dimensional gel electrophoresis (2DGE) [3,4]. This technique separates proteins according to their masses and charges. These independent attributes enable separations of thousands of proteins in a single analysis. After separation all proteins can be identified by mass spectroscopy. In studies performed to identify specific proteins related to a given metabolic process or disease it is, however, much more efficient to detect and identify only proteins differentiating groups of samples [5]. In comparative studies of biological material a large number of analyses must be conducted to suppress natural differences and to increase differences related to process being studied. This results in a huge amount of data to be analyzed and generates a need for a rapid, efficient and fully automated method for matching and comparing gel images. The images may differ significantly and also contain noise of different characteristics and a varying baseline (Fig. 1). This necessitates careful preprocessing, i.e. noise and background removal.

Signal noise hampers spot detection with methods based on signal derivatives, because they magnify the noise, causing identification of false peaks (spots) and incorrect determination of the borders of spots. In turn, varying signal background interferes with methods based on histogram segmentation and thresholding [6]. The background component is added to the

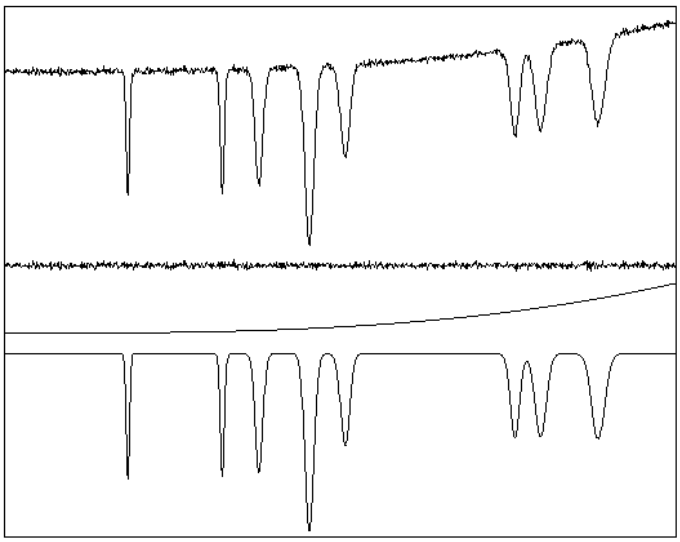


Fig. 1

Components of 2DGE image presented on its profile (from top to bottom): profile of 2DGE image, noise, baseline, and real signal.

real signal and overstates the intensities of spots. Another serious problem caused by varying background is difficulty with matching images. In area-based matching [7,8] images are matched to maximize the similarity measure between them [9]. Varying background present in warped images makes it difficult to properly calculate their similarity. Filtering in the wavelet domain has been found to be the method best suited for dealing with noise present in 2DGE images [10].

Many numerical methods have been developed for estimation of varying background present in one-dimensional signals [11–25]. Among these techniques are methods based on digital filters [11–15]. Such filters usually introduce artefacts and simultaneously distort the real signal. Other approaches rely on automated peak rejection [16]. These algorithms fit some functions to find regions of signal that consist only of the baseline without peaks of real signal. The functions being fitted may have different forms, e.g. polynomials [17,18], splines [19]. The main disadvantages of peak-rejection approaches are difficulties related to identification of peak-free regions. On the other hand, threshold-based rejection of peaks gives good results when the baseline is relatively smooth [20], and fails for signals with significantly varying baseline.

Because of difficulties caused by automatic peak rejection other approaches have been designed to fit a baseline without detecting the peaks. In Ref. [17] the baseline is fitted with a low-order polynomial that prevents it from fitting the real signal peaks. For signals with many positive peaks, however, e.g. electrophoretograms, the baseline estimated in this way has values which are too high. Subtraction of a background with values which are too high from a signal introduces significant distortions to the analyzed signal, i.e. the values for the spots (peaks) are too low. Other approaches rely on statistical methods, such as maximum entropy [21,22]. There are also approaches based on baseline removal in the wavelet domain [23-25]. In this paper we focus on the method, proposed by Eilers [26] for background elimination in two-dimensional signals based on asymmetric least squares splines regression and evaluate its potential as an automated approach.

THEORY

The new baseline-correction procedure proposed by Eilers [26] may be regarded as a method similar to the 'peak rejection' approaches; there is, however, no need to detect peaks. This procedure is based on the Whittaker smoother [27,28], which minimizes the following cost function:

$$Q = \sum_i v_i \left(y_i - f_i\right)^2 + \lambda \left(\Delta^d f_i\right)^2 \qquad (1)$$

where y is the analyzed signal, f is a smooth approximation of y (baseline), i denotes the consecutive values of the signal, d is the order of differences Δ, and v are weights.

Weights v should have high values in parts of the signal where the signal analyzed is allowed to affect estimation of the baseline. In all other regions of the signal, values of v are zero. The positive parameter λ is the regularization parameter and controls the significance of the penalty term $(\lambda(\Delta^d f_i)^2)$, i.e. the higher the value of λ, the smoother the estimated baseline.

Because of the asymmetry problem in baseline estimation, the weights should be chosen in a way that will enable 'rejection' of the peaks. To achieve this, the weights are assigned as:

$$v_i = \begin{cases} p & \text{if } y_i > f_i \\ 1 - p & \text{if } y_i \le f_i \end{cases} \qquad (2)$$

where $0 < p < 1$.

The positive deviations from the estimated baseline (peaks) have low weights while the negative deviations (baseline) obtain high weights. There is, however, a problem of simultaneous determination of weights (v) and baseline (f). Without the weights it is impossible to calculate the baseline and without the baseline it is impossible to determine the weights. This problem is solved iteratively, i.e., in the first iteration all weights get the same value, i.e. unity. Using these weights, the first estimate of the baseline is calculated. Iterating between calculation of the baseline and setting weights, gives a good estimate of the baseline in a few iterations. The use of p close to zero and large λ enables baseline estimation to follow the baseline exactly. The shape of the estimated baseline is not too flat and simultaneously does not follow the peaks of the real signal.

The penalty term in eq. (1) can also be formulated differently. In Ref. [29] Eilers proposed the splines-based approach to smoothing instrumental signals. The multidimensional extension of the spline-based approach was presented by Eilers in Ref. [30].

The two dimensional signal is described by a data matrix **Y** containing $i \times j$ intensity values. To estimate background, let **B** ($j \times l$) be a B-spline basis along columns of **Y** matrix and $\bar{\mathbf{B}}$ ($i \times k$) be a B-spline basis along rows of **Y** matrix. The spline basis along columns of signal constructed from five basis functions is presented in Fig. 2.

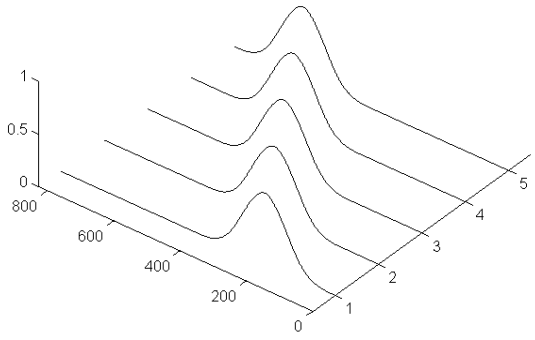


Fig. 2

Spline basis along columns of signal constructed from five basis functions

As a compromise between speed of calculation and memory requirements on the one hand and accuracy on the other, ten basis functions, i.e., $k = l = 10$ were used in our study. Then, estimation of a smooth surface **F** can be presented in the equation:

$$f_{i,j} = \sum_{k,l} \breve{b}_{ki} b_{lj} a_{kl} \tag{3}$$

where a_{kl} is the k,lth element of matrix **A** containing regression coefficients.

The matrix of regression coefficients can be calculated by minimizing the cost function:

$$Q = \sum_{i,j} v_{i,j} \left(y_{i,j} - f_{i,j} \right)^2 + P \tag{4}$$

where P is a penalty term defined as:

$$P = \lambda \sum_{l} \left(\Delta^d \mathbf{a}_{\bullet \mathbf{l}} \right)^2 + \breve{\lambda} \sum_{k} \left(\breve{\Delta}^d \mathbf{a}_{\mathbf{k}\bullet} \right) \tag{5}$$

The first part of eq. (5) is a difference Δ of order d calculated for each column of **A** ($\mathbf{a}_{\bullet \mathbf{l}}$) and the second part is the difference $\breve{\Delta}$ of order $\breve{d}$ calculated for each row of **A** ($\mathbf{a}_{\mathbf{k}\bullet}$). From eq. (5) it is apparent that the penalty may have different values for the vertical and horizontal directions, because there are two different regularization parameters ($\lambda, \breve{\lambda}$). As the backgrounds in 2DGE images do not have different spatial structure for the horizontal and vertical directions, however, one value for both regularization parameters will be used ($\lambda = \breve{\lambda}$).

DATA

Real 2DGE images have been used for visual inspection of baseline estimation accuracy. These images contain results from separations conducted for human and animal (e.g. mouse) tissues and have been taken from public databases [31,32]. For evaluation of Monte Carlo performance simulated gel images have been used. Images have been simulated as squares of different side length (512 to 1024 pixels) and different numbers of spots (500 to 3000 spots) placed in random coordinates. The spots have been simulated as two-dimensional Gaussian functions. Random white noise has been added to each simulated image, resulting in images with signal-to-noise ratios varying from 30 to 50. Also, the varying background has been added. The background has been simulated as a smooth surface in the following way:

1. five or nine points were selected, four lying in the corners of the image, one in the middle of the image, and, for nine points, four halfway between the centre of the image and its edges;

2. random intensity values (not exceeding the 25% of the highest intensity in the image) were assigned to these points;
3. a smoothing function was used to interpolate the background values between the points.

The procedure described resulted in smooth backgrounds similar to those present in real 2DGE images (Fig. 3).

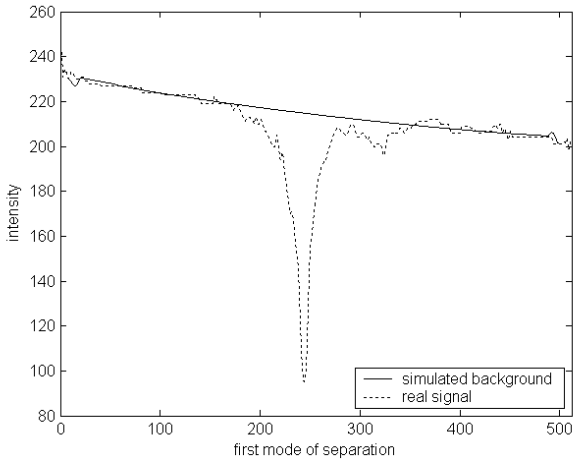


Fig. 3

Comparison of simulated background and typical real 2DGE image

RESULTS AND DISCUSSION

Typical results from baseline estimation using the 2D extension of the spline-based approach are presented in Fig. 4. Baseline estimation was performed on a typical real 2DGE image [31], which is presented in Fig. 4a. The image with removed background is presented in Fig. 4b.

Details of the background estimation process are better visible on a single profile of a 2DGE image. For this reason Fig. 5 shows the baseline estimation process for a single profile of a simulated 2DGE image. For this profile, the process of baseline estimation converged in twelve iterations; a few of these iterations (1st, 2nd, 3rd, and final) are presented. This profile has negative peaks, so the different weights assigned to data points must be used (eq. 2). In this case the points lying above the estimated baseline must obtain weights with high values, so p must be close to unity. In the example above following parameters were used: $\lambda = 100$, $p = 0.999$, and $d = 3$.

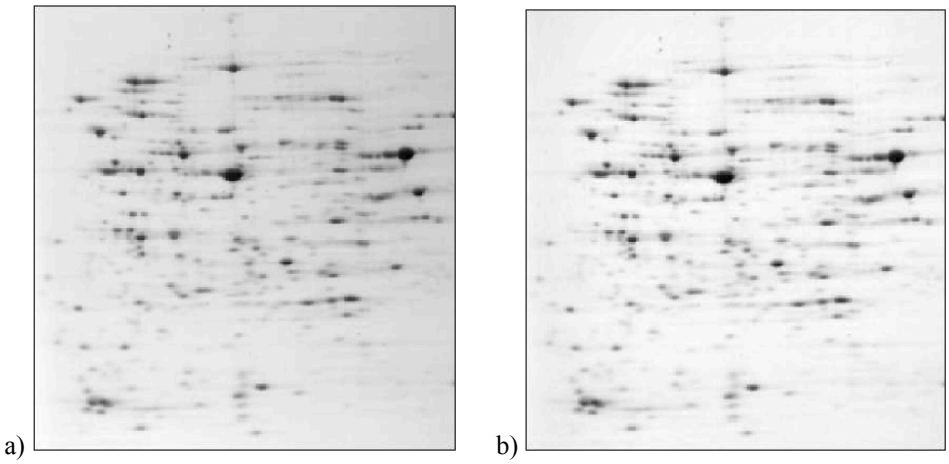


Fig. 4

Results from baseline removal: (a) real 2DGE image [31] and (b) its counterpart with removed background

Optimization of Input Parameters

Because each of the input parameters used for calculation of the cost function in eq. (3) (p, λ, d) affects the results obtained differently, they must be chosen carefully to ensure proper estimation of the baseline. For this reason the automated method for baseline correction should enable automated determination of input parameters or should be insensitive to changes of input parameters over a wide range of their values. Examples of results obtained for different values of λ and d are presented in Fig. 6

It is apparent that the values of both parameters determine the accuracy of baseline estimation. The higher the value of λ (smoothing parameter) the flatter the estimated baseline becomes. In turn, d (order of the differences) decides how well baselines of different shape, i.e. polynomials of different degree, will be estimated. For 2DGE images second-order differences seem to be a reasonable choice, because they yield a good estimate of the baseline. The value of p (weights) also affects the estimates obtained; its closeness to unity, however, e.g. 0.999, ensures correct results for signals containing negative peaks.

Methods are available for automated estimation of regularization parameters (λ) for signals without an asymmetry problem. The methods commonly used for determination of λ include generalized cross-validation [33], the L-curve criterion [34], the quasi-optimality criterion [35], and the

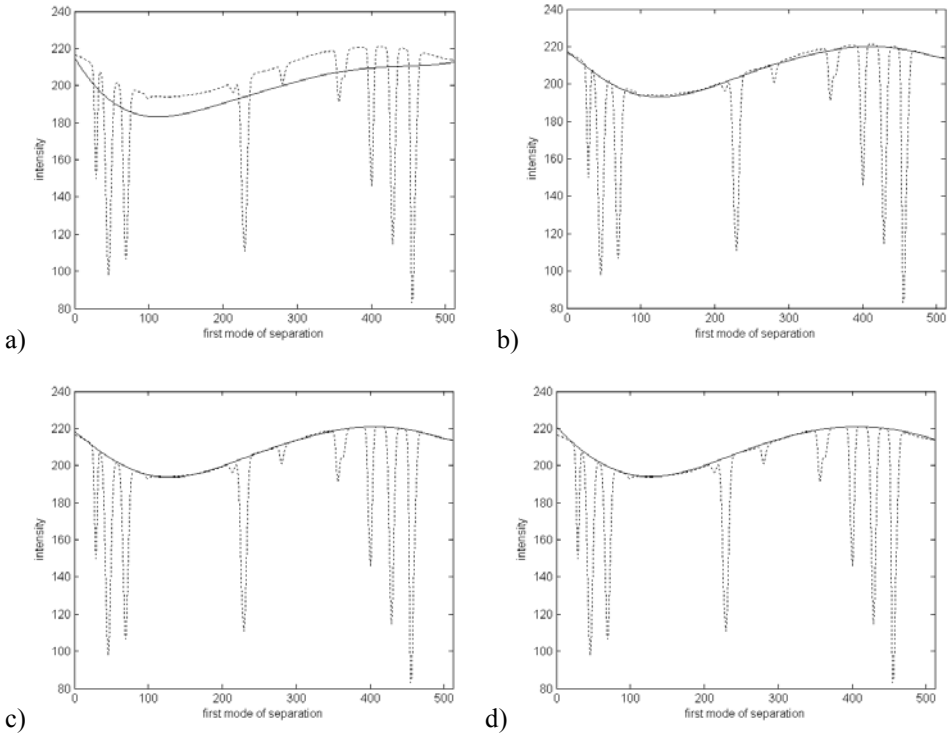


Fig. 5

Illustration of the baseline estimation process in consecutive iterations: (a) first, (b) second, (c) third, and (d) final iteration (λ = 500, p = 0.999 d = 3). The estimated baseline is depicted as the solid line.

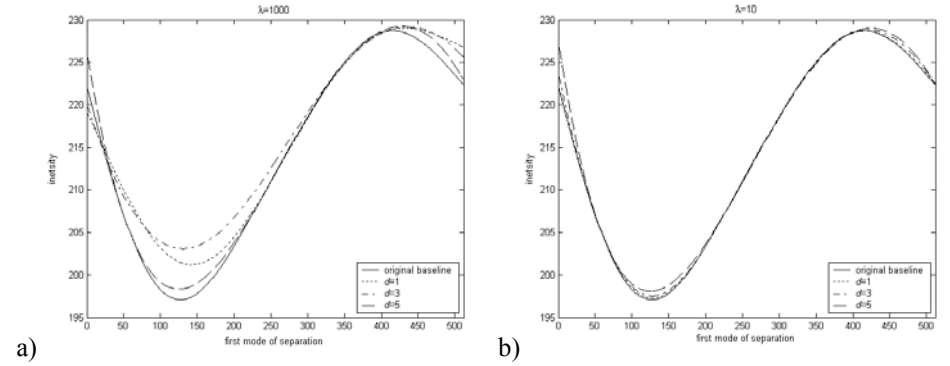


Fig. 6

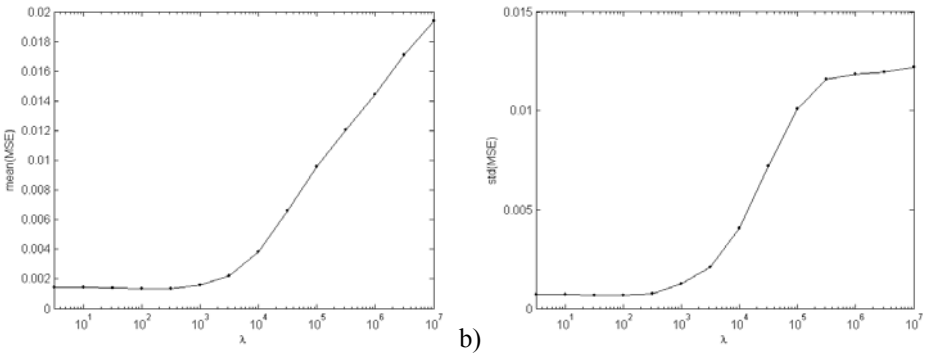
Baseline estimates obtained for different values of λ and d: (a) λ=1000, (b) λ=10

discrepancy principle [36]. None of these methods can be used for automatic determination of the λ value for 2DGE images, because they have significant asymmetry.

If there is a lack of methods for automatic determination of the regularization parameter it is possible to conduct a Monte Carlo study on simulated data. Such a study enables determination of a value of λ which yields proper results for simulated data and also for real data with similar characteristics. For simulated images it is possible to determine the quality of baseline estimation by comparing the real (known) background with that estimated. As a measure of quality the MSE (mean square error) may be used:

$$\mathrm{MSE} = \frac{\sum_{n,m} \left(\left| b_{n,m} - \hat{b}_{n,m} \right| \right)}{n \cdot m} \tag{5}$$

where n and m denote the vertical and horizontal size of the image and $b_{n,m}$ and $\hat{b}_{n,m}$ are values of a single pixel from the real background and the estimated background, respectively. Results from real and estimated background comparison are presented in Fig. 7.



a) b)

Fig. 7

(a) Mean values of MSE for 100 different images, obtained for different values of the regularization parameter, and (b) their standard deviation

The results were obtained for 100 images. Fifty were simulated using five points and fifty were simulated using nine points for background generation. Comparison of the real and estimated baselines indicate that optimum values of the regularization parameter λ are the range 10^1 to 10^3. This study enabled determination of the λ value for simulated signals. This

value will also yield correct results for real images with background characteristics similar to those present in simulated data. It was also shown that the described method is very robust and resistant to changes in signals.

Baseline Correction for Noisy Signals

For real, noisy signals even points belonging to the baseline may lie under the estimated baseline in a given iteration (Fig. 8a). As such, they will not be taken into calculation of baseline in next iteration, causing overestimation of baseline, i.e., background will obtain too high values. Because of that, there is a need for estimation of noise level present in analyzed signal. We propose two methods for estimation of the baseline for noisy 2DGE images.

1. The standard deviation of the noise can be easily estimated on the basis of wavelet coefficients [37]. After estimation of the noise it is possible to calculate the background using not only points lying above the estimated baseline – points lying under the baseline also are used for estimation if their distance to baseline is smaller than the estimated noise. The baseline presented in Fig. 8a is estimated using only points lying above the baseline whereas the baseline presented in Fig. 8b is estimated using also points lying under the baseline within the range of estimated noise.
2. The second way of dealing with noise present in analyzed signals is heavy smoothing without attempting to preserve narrow peaks. Heavy smoothing may be achieved by use of median filtering with a broad window, e.g. equal to 5% of signal length. Such smoothing does not change the background but very effectively removes the noise, and of course, most of the narrow peaks. This enables proper estimation of the baseline even for noisy signals (Fig. 8c).

Baselines estimated using these two different approaches are compared in Fig. 9. It is apparent the baseline estimated without any consideration of the noise present in the signal is significantly higher than the real background. Both approaches ensure proper estimation of the baseline without the influence of noise present in the image. Median smoothing seems to be less complicated and faster than the wavelet approach. Usually, however, image processing also involves noise reduction and the wavelet approach enables simultaneous, efficient noise removal [10] with estimation of the noise present in the image.

For typical gel images, the method of background estimation presented yields reasonable results for a wide range of parameter values. The

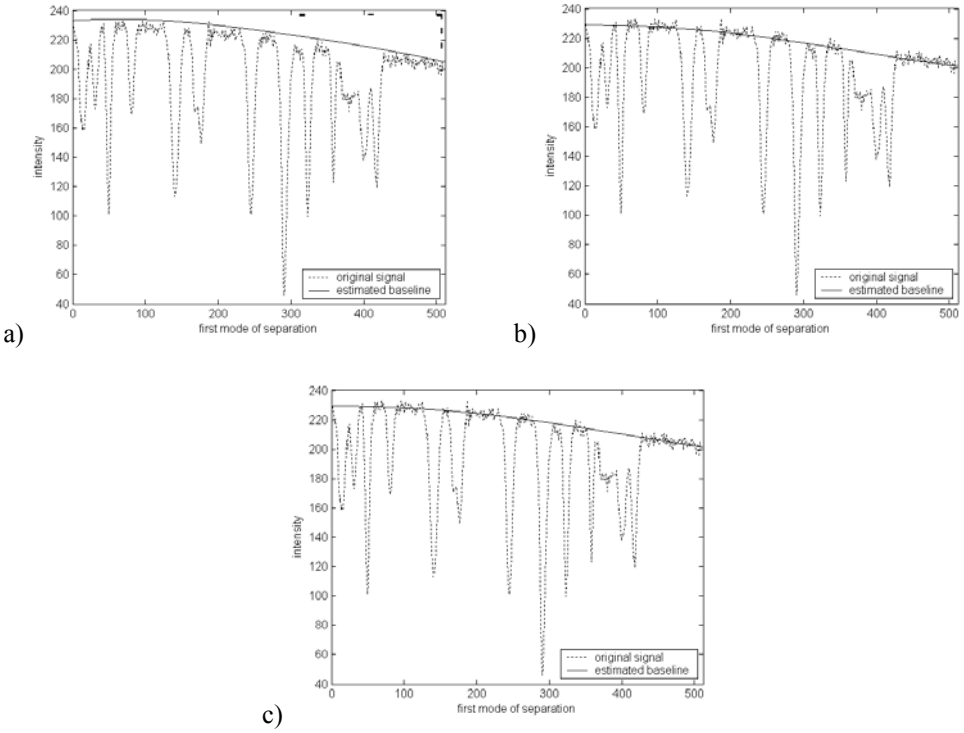


Fig. 8

Examples of baselines estimated for (a) a noisy signal, and for the same signal with (b) wavelet estimation of noise ($\lambda = 10^4$, $p = 0.999$, $d = 2$) and (c) using a heavily de-noised signal ($\lambda = 10^4$, $p = 0.999$, $d = 2$)

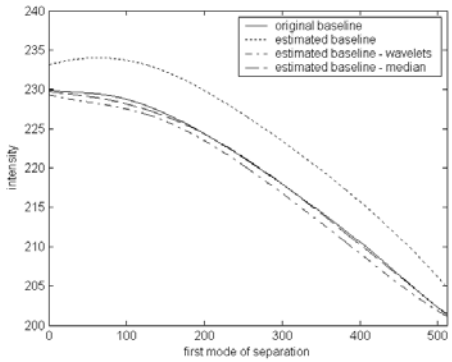
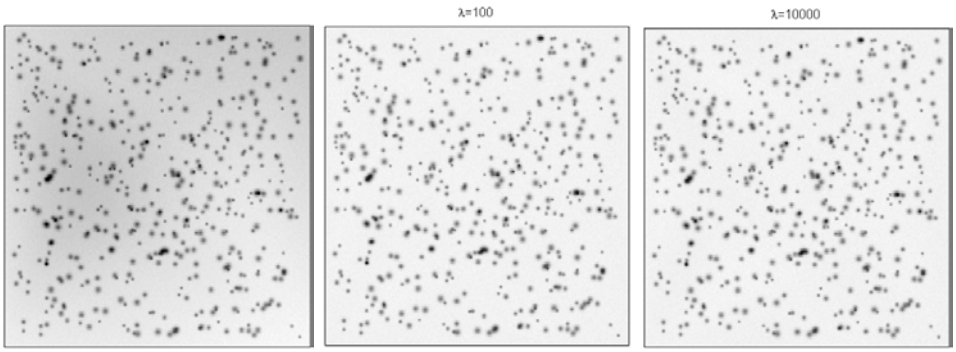


Fig. 9

Comparison of background estimates obtained by use of different methods for dealing with signal noise

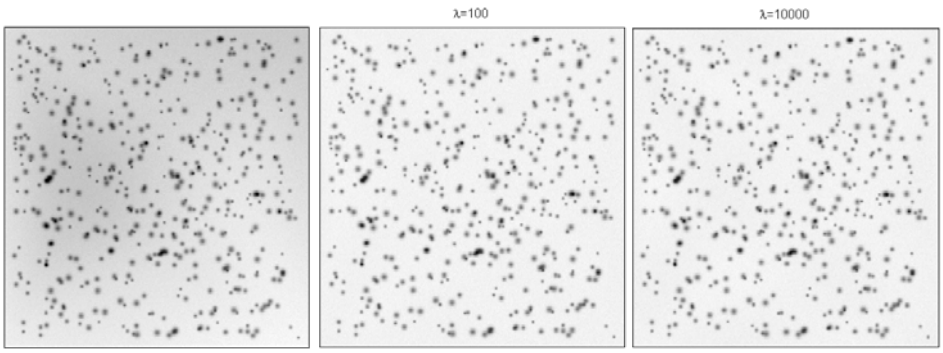
results obtained for simulated 2DGE images are presented in Figs. 10 and 11. The accuracy of method is immune even to large changes in λ and d.



a) b) c)

Fig. 10

Results from baseline removal: (a) simulated 2DGE image and (b, c) its counterparts with removed background using different values of the smoothing parameter (λ = 100 and λ = 10000) and d = 2



a) b) c)

Fig. 11

Results from baseline removal: (a) simulated 2DGE image and (b, c) its counterparts with removed background using different values of smoothing parameter (λ = 100 and λ = 10000) and d = 4

For the reasons already mentioned, the real images may be used solely for visual inspection of the result obtained. The results obtained for typical real 2DGE images are presented in Figs. 4, 12, and 13.

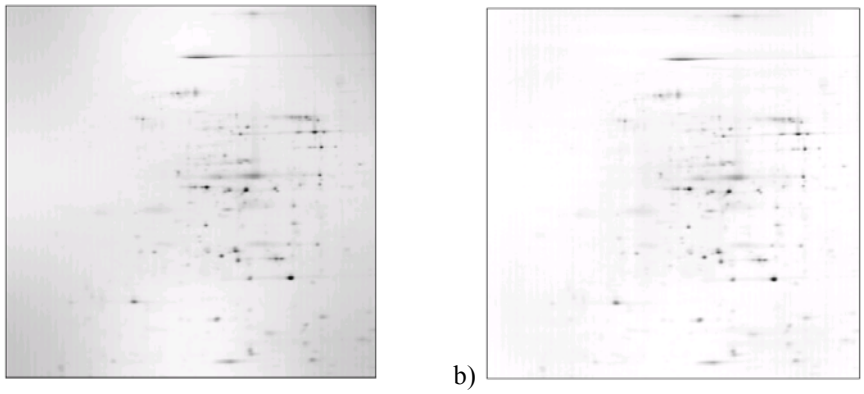


Fig. 12

2DGE image [31] with varying background (a) and the same image with removed background (b)

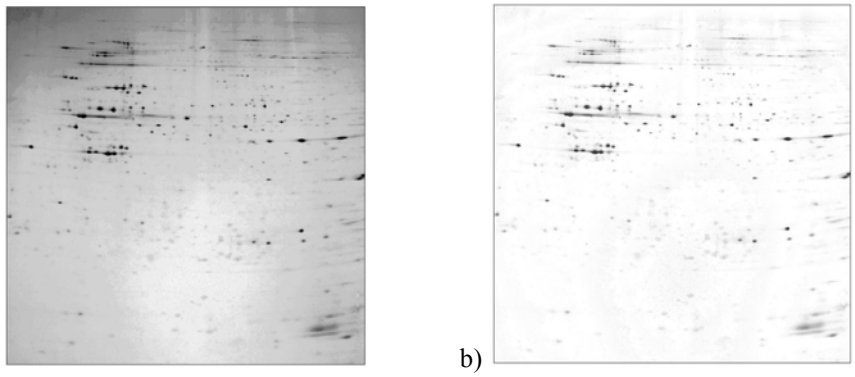


Fig. 13

2DGE image [32] with varying background (a) and the same image with removed background (b)

CONCLUSIONS

In the work discussed in this paper the suitability of baseline reduction methods based on splines regression were investigated. The tests were conducted on synthetic images in a Monte Carlo study and, because the backgrounds generated in the synthetic images have characteristics similar to those of backgrounds present in real images, the results should also

be valid for real 2DGE images. The evaluation demonstrated that the method presented deals very effectively with the background present in typical 2DGE images. The method is very robust and gives good results for wide range of values of the regularization parameter for typical images. This means that it is possible to deal very effectively and automatically with the background present in typical gel electrophoresis images. Occasionally, however, the need for fine tuning of the input parameters may arise. For typical images the error of estimation is very low, irrespective of values of the smoothing parameter and the size of the images. The proposed techniques of heavy smoothing and wavelet estimation of the noise enable baseline estimation without bias introduced by the noise always present in instrumental signals.

ACKNOWLEDGEMENT

The authors are very grateful to Professor Eilers for the Matlab code of the 2D splines approach. K. Kaczmarek thanks Unilever R&D (Vlaardingen, Holland) for financial support of his PhD study.

REFERENCES

[1] S.P. Gygi, Y. Rochon, B.R. Franza, and R. Aebersold, Mol. Cell. Biol., **19**, 1720 (1999)
[2] H.F. Hebestreit, Curr. Opin. Pharmacol., **1**, 513 (2001)
[3] P.H. O'Farrell, J. Biol. Chem., **250**, 4007 (1975)
[4] S.J. Fey and P.M. Larsen, Curr. Opin. Chem. Biol., **5**, 26 (2001)
[5] W.P. Blackstock and M.P. Weir, Trends Biotechnol., **17**, 121 (1999)
[6] Y. Peng-Yeng and C. Ling-Hwei, Signal Process., **60**, 305 (1997)
[7] S. Veeser, M.J. Dunn, and G.Z. Yang, Proteomics, **1**, 856 (2001)
[8] Z. Smilansky, Electrophoresis, **22**, 1616 (2001)
[9] C. Heipke, Overview of Image Matching Techniques. OEEPE Workshop "Applications of Digital Photogrammetric Workstations", Proceedings, Lausanne, Switzerland, 1996, pp. 173–191
[10] K. Kaczmarek, B. Walczak, S. de Jong, and B.G.M. Vandeginste, Proteomics, **4**, 2377 (2004)
[11] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, and J.A. Dodd, J. Quant. Spectrosc. Radiat. Transfer, **68**, 179 (2001)

[12] M.A. Kneen and H.J. Annegarn, Nucl. Instrum. Methods Phys. Res. B, **109/110**, 209 (1996)
[13] J.A. Maxwell, J.L. Campbell, and W.J. Teesdale, Nucl. Instrum. Methods Phys. Res. B, **43**, 218 (1989)
[14] Y. Sun, K.L. Chan, and S.M. Krishnan, Comput. Biol. Med., **32**, 465 (2002)
[15] E.H. van Veen and M.T.C. de Loos-Vollebregt, Spectrochim. Acta B, **53**, 639 (1998)
[16] A. Rouh, M.A. Delsuc, G. Bertrand, and J.Y. Lallemand, J. Magn. Reson. Ser. A, **102**, 357 (1993)
[17] H. Abbink Spaink, T.T. Lub, R.P. Otjes, and H.C. Smith, Anal. Chim. Acta, **183**, 141 (1986)
[18] G. Dellalunga, R. Pogni, and R. Basosi, J. Magn. Reson. Ser. A, **108**, 65 (1994)
[19] G. Della Lunga and R. Basosi, J. Magn. Reson. Ser. A, **112**, 102 (1995)
[20] W. Dietrich, C.H. Rüdel, and M. Neumann, J. Magn. Reson., **91**, 1 (1991)
[21] J. Padayachee, V. Prozesky, W. von der Linden, M.S. Nkwinika, and V. Dose, Nucl. Instrum. Methods Phys. Res. B, **150**, 129 (1999)
[22] A.J. Phillips and P.A. Hamilton, Anal. Chem., **68**, 4020 (1996)
[23] X.-G. Ma and Z.-X. Zhang, Anal. Chim. Acta, **485**, 233 (2003)
[24] H-W. Tan and S.D. Brown, J. Chemom., **16**, 228 (2002)
[25] C. Perrin, B. Walczak, and D.L. Massart, Anal. Chem., **73**, 4903 (2001)
[26] P.H.C. Eilers, Anal. Chem., **76**, 404 (2004)
[27] P.H.C. Eilers, Anal. Chem., **75**, 3631 (2003)
[28] R.J. Verrall, Ins.: Mathematics Econ., **13**, 7 (1993)
[29] P.H.C. Eilers and B.D. Marx, Statist. Sci., **11**, 89 (1996)
[30] P.H.C Eilers, I.D. Currie, and M. Durbán, Comput. Stat. Data Anal., in press
[31] GelBank at http://gelabank.anl.gov
[32] World-2DPAGE - 2-D PAGE databases and services at http://www.expasy.ch/ch2d/2d-index.html
[33] G.H. Golub, M. Heath, and G. Wahba, Technometrics, **21**, 215 (1979)
[34] P.C. Hansen, SIAM Rev., **34**, 561 (1992)
[35] S. Morigi and F. Sgallari, Appl. Math. Comput., **121**, 55 (2001)
[36] H.W. Engl and H. Gfrerer, Appl. Numer. Math., **4**, 395 (1988)
[37] D.L. Donoho, De-noising via Soft Thresholding, IEEE Trans. Inform. Theory, **41**, 613 (1995)