A Genetic Approach to Spot Detection in Two-Dimensional Gel Electrophoresis Images

Dimitris K. Iakovidis, *Member, IEEE*, Dimitris Maroulis, *Member, IEEE*, Eleni Zacharia, and Sofia Kossida

Abstract—**Two-Dimensional Polyacrylamide Gel Electrophoresis (2D PAGE) is a proteomic technique that allows the analysis of large collections and complex mixtures of proteins. The 2D-PAGE gel images depict protein signals as spots of various intensities and sizes. In this paper, we present a novel approach to unsupervised protein spot detection in 2D-PAGE images based on a genetic algorithm. This approach involves three main steps: a) wavelet-based noise reduction, b) segmentation of the input images into regions around the local maxima of the image intensities, c) detection and model-based quantification of the spots within each region using a genetic algorithm. This algorithm searches within a multidimensional parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots. The detection and quantification of the spots is achieved by superposition of diffusion functions modeling adjacent spots. Experiments with 16-bit 2D-PAGE images show that the proposed method is effective and results in low spurious spot detection rate.**

I. Introduction

PROTEOMIC research deals with the systematic analysis of protein profiles expressed in a given cell, tissue or biological system at a given time. In this field, two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE) analysis [1], is a well-established and widely used technique for the analysis of large collections and complex mixtures of proteins. Images produced by digitization of 2D-PAGE gels contain spots, of various intensities and sizes, that correspond to proteins. Detection and quantification of the protein spots may reveal alterations in protein expression within a given biological system. However, this is not a straightforward process. It can be rather complicated due to the presence of noise, the inhomogeneneous background, and the overlap between the spots.

A variety of software packages have been developed for protein spot detection [2]. Many of these packages implement image segmentation methods based on edge detection algorithms such as Laplacian filtering, in conjunction with smoothing or morphological operators [2]-[4]. However, if a 2D-PAGE image contains artifacts it is likely that the boundaries of the artifacts have similar characteristics with the boundaries of the actual spots, leading to spurious spot detection. Moreover, the segmentation produced by the edge detection methods is particularly dependent on the preparation of the 2D-PAGE gels. The watershed algorithm has also been a popular choice for 2D-PAGE image segmentation [4]. It usually performs better than the methods based on edge detection algorithms, however it tends to oversegment the images. To alleviate possible oversegmentation effects, post-processing techniques, such as region merging, are usually applied. State of the art approaches to spot detection by image segmentation include geometric algorithms [5], and the pixel value collection method [6].

2D-PAGE image segmentation is usually followed by characterization and representation of the protein spots with a list of parameters over which further analysis can be carried out. Spot characterization algorithms span two categories: parametric and nonparametric. Nonparametric methods [6]-[7] involve heuristic post-processing of the segmentation boundaries for the delineation of spots, which are then represented by a set of measurements calculated over the detected spot regions. These methods do not impose any explicit constraint on the shape of the boundaries or the appearance of the spots. However, they exhibit poor performance with complex images.

Parametric methods utilize model functions to parameterize protein spots. Models represent prior knowledge used to impose constraints on the analysis procedure. This in turn improves the robustness of the solution. Early approaches to modeling protein spots in 2D-PAGE images include the use of 2D-Gaussian functions [8][9]. This model provides a good representation of some spots, but has proved inadequate as a general model. More precisely, in [10] it is noted that when the local concentration of protein is high, saturation effects occur and the spot can not be accurately modeled by a Gaussian function. Instead, a simplified diffusion model is suggested as more appropriate. Optimization of a model's parameters usually involves supervised techniques. For example, Melanie, a popular software package for 2D-PAGE analysis, uses the Polak-Ribiere variant of the conjugate gradient method to optimize the parameters of a Gaussian model [10]. Latest advances in spot modeling include the construction of spot models based

Manuscript received June 30, 2006. This work was supported by the Greek General Secretariat of Research and Technology and the European Social Fund, through the PENED 2003 program (grant no, 03ED332).

D. K. Iakovidis, D. Maroulis, and E. Zacharia, are with the Department of Informatics and Telecommunications, University of Athens, GR 15784 Panepistimiopolis, Ilisia, Greece (e-mail: rtsimage@di.uoa.gr).

S. Kossida, is with the Division of Biotechnology, Center of Basic Research, Foundation of Biomedical Research of the Academy of Athens, Greece (e-mail: skossida@bioacademy.gr).

on a population of characterized spots [11]. However, this is also a supervised approach.

In this paper, we present a novel approach to unsupervised protein spot detection in 2D-PAGE images based on a genetic algorithm. This algorithm searches within a multidimensional parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots. The detection and quantification of the spots is achieved by superposition of diffusion functions modeling adjacent spots. To the best of our knowledge genetic algorithms have not been previously applied to protein spot detection.

The rest of this paper is structured in four sections. Section II describes the diffusion modeling of protein spots used in this study. The proposed approach to spot detection is described in Section III. The results from its application on real 2D-PAGE images are shown in Section IV, whereas the conclusions of this study are summarized in Section V.

II. Diffusion Modeling of Protein Spots

Protein spots share with each other some common characteristics, such as having an approximately elliptical shape and a limited range of intensities that peak centrally and diminish towards their perimeter. These simple characteristics can be captured by tuning the parameters of a mathematical model so that it fits an image region containing a spot.

The diffusion model proposed in [10] suggests that the protein spots are modeled by a mathematical function representing the actual diffusion process of a protein into a 2D-PAGE gel. The assumptions about the process include: a) the medium of the diffusion is two dimensional and anisotropic, i.e. there are two main directions of diffusion (x and y) with different diffusion constants D_x and D_y, b) the diffusing substance is initially distributed uniformly on a disc with radius a.

The equation of the diffusion model represents the concentration of the spot's substance as a function of (x, y):

$$C(x,y) = B + \frac{C_0}{2}\left[\operatorname{erf}\left(\frac{a'+r'}{2}\right) + \operatorname{erf}\left(\frac{a'-r'}{2}\right)\right] + \frac{C_0}{r'\sqrt{p}}\left[\exp\left(-\left(\frac{a'+r'}{2}\right)^2\right) + \exp\left(-\left(\frac{a'-r'}{2}\right)^2\right)\right] \quad (1)$$

where

$$r' = \sqrt{\frac{(\mathrm{x} - x_0)^2}{D'_x} + \frac{(\mathrm{y} - y_0)^2}{D'_y}} \quad (2)$$

and

$$\operatorname{erf}(z) = \frac{1}{\sqrt{p}}\int_0^z \exp(-t^2)\,dt \quad (3)$$

is the error function encountered in integrating the normal distribution, B is the background intensity, C_0 is the initial concentration of the substance, D'_x and D'_y are related to the diffusion constants in the two main directions of diffusion ($D'_x = D_x t$ and $D'_y = D_y t$, where t is the time), (x_0, y_0) are the coordinates of the substance on the plane, and $a' = a\sqrt{D/t}$ is the area of the disc containing the substance. For $a' \to 0$ Eq. (1) becomes the 2D-Gaussian function.

III. A Genetic Approach to Spot Detection

The proposed approach to protein spot detection consists of 3 main steps: a) noise reduction, b) segmentation of the input images into regions around the local maxima of the image intensities, c) detection and model-based quantification of the spots within each region using a genetic algorithm.

A. Noise Reduction

The scanning devices used for the production of the 2D-PAGE images often pick up dust particles corrupting the images with impulse noise. This type of noise can be effectively reduced by wavelet-based filtering [6], which is employed as a pre-processing step to improve the quality of the input images prior to segmentation.

B. Image Segmentation

The pre-processed 2D-PAGE images are segmented into regions around the local maxima of image intensities by using a variant of the pixel value collection algorithm [6].

Local maxima are the most probable candidates for spot centers and each maximum is assigned a unique label. However, highly overlapping spots may not be sufficiently discriminated by local maxima (Fig. 1, c-d.).

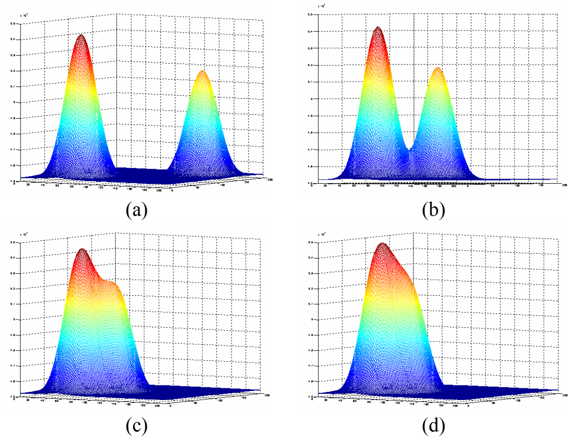
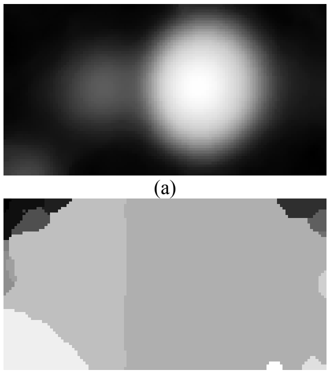

(a)
(b)
(c)
(d)

Fig. 1. Three-Dimensional representation of protein spots: (a) Non-overlapping spots, (b) Minor overlap: the segmentation algorithm can separate them (c) Substantial overlap: the segmentation algorithm cannot separate them, but they can be separated (d) Complete overlap: the spots can not be separated.

For each integer p from the maximum to the minimum value of the dynamic range of image intensities, the algorithm proceeds to labeling the pixels of intensity p. For each pixel a majority voting criterion is applied among the labels of its adjacent pixels. If all the adjacent pixels are unlabeled, the pixel is assigned a new label. The algorithm results in a mosaic of labeled regions, each of which is likely to contain either a single protein spot, exhibiting no or minor overlap with it's adjacent spots (Fig. 1, a-b), or multiple protein spots with substantial or complete overlap (Fig. 1, c-d). Each region contains exactly one local maximum. An example segmentation of a 2D-PAGE image is illustrated in Fig. 2.



(a)

(b)

Fig. 2. 2D-PAGE image segmentation: (a) input, (b) output.

This variant of the pixel value collection algorithm can be advantageous for spot detection over its original form. It labels all the pixels in the image by growing the regions containing the protein spots to the greatest possible extent. The original algorithm stops the growing of a region if a spot is found to be merged with another. In this case the detected spot does not reach its correct boundary.

C. Detection and Model-Based Quantification of Spots

This step aims to determine the optimal diffusion model for each protein spot, in the pre-processed 2D-PAGE image. Finding the optimal model parameters is not straightforward due to the overlap between the spots and the imperfect diffusion of the spot substance across the gel medium [14]. In order to automatically tune the parameters of the diffusion models so that they optimally fit the protein spots, we developed a novel method based on a genetic algorithm capable of dealing with the afore-mentioned situations.

Genetic algorithms are stochastic non-linear optimization algorithms based on the theory of natural selection and evolution [16]. Compared to traditional search and optimization procedures, genetic algorithms are parallel, robust optimizers, suitable for solving problems for which there is a little or no a priori knowledge about the underlying processes.

The genetic approach to spot detection proposed in this paper assumes that adjacent spots may be overlapping. Fig. 3 illustrates a hypothetical 2D-PAGE subimage segmented into eight regions using the image segmentation algorithm described previously. The developed genetic algorithm performs a parallel search for the optimal parameters of:

- the diffusion model that correspond to the protein spot(s) contained in the central region labeled as A (Fig. 3) and

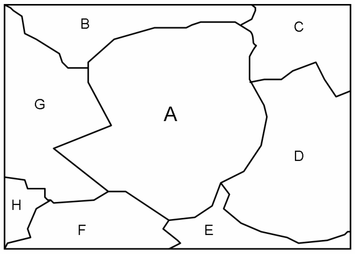
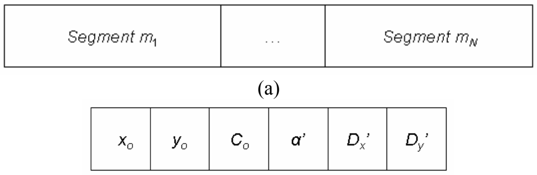


Fig. 3. Sketch of a hypothetical 2D-PAGE subimage segmented into eight labeled regions containing protein spots.

- the diffusion models that correspond to the protein spots of the adjacent regions, labeled as B, C, D, E, F and G (Fig. 3).

1) Chromosome: The parameters of the diffusion models that correspond to the protein spots contained in the central and the adjacent regions are encoded into a single chromosome m (Fig. 4a). The chromosome consists of N segments m_i, i=1,2,…N. The segment m_1 encodes the model parameters of the spot contained in the central region, whereas the segments m_i, i=2,3,…N encode the model parameters of the spots contained in the adjacent regions. In the case of the hypothetical 2D-PAGE subimage illustrated in Fig. 3, N=7, and the chromosome segments m_1 to m_7 correspond to the regions A to G.

Each chromosome segment (Fig. 4b) is a string of real values representing the parameters of the corresponding diffusion model. Such real-coded chromosomes exhibit various advantages over binary coded chromosomes as they can use large or unknown domains for the variables they code. On the other hand, assuming that the chromosome has a fixed length, binary implementations cannot increase the domain without sacrificing precision [17].



(a)

(b)

Fig. 4. The chromosome used in the genetic algorithm: (a) Real-coded segments comprising the chromosome, (b) The parameters encoded in a segment of the chromosome.

It should be noted that the background parameter $B(m_i)$ of the diffusion model encoded by the chromosome segment m_i has not been included in the chromosome segment because it can be computed by the following equation:

$$B(m_i) = I(x_0, y_0) - C(x_0, y_0 \mid m_i)\big|_{B=0} \qquad (4)$$

where (x_0, y_0) are the coordinates of the center of the spot modeled by C in the pre-processed 2D-PAGE image, and I is the pre-processed 2D-PAGE image.

2) Genetic Operations: Beginning with an initial population of randomly generated chromosomes, the genetic algorithm evolves the population by subsequent elitist reproduction [18], uniform crossover [19] and random mutation [16] operations.

The genetic algorithm is executed for each labeled region of the segmented image. The spot models are stored and their cross section with the image plane is depicted, so that the boundaries of the spots become visible.

3) Fitness Function: The fitness of a chromosome *m* as a solution to the particular optimization problem is defined by the following equation:

$$f(m) = \max\left[f_L(m_1), f_C(m)\right] \tag{5}$$

where the real valued functions $f_L(m_i)$ and $f_C(m)$, are named *local* of a chromosome segment m_i, i=1,2,…N, and *central* fitness of the chromosome, respectively.

The *local fitness* of a chromosome segment m_i is computed by the following equation:

$$f_L(m_i) = \sum_{\substack{[(x,y)\in i]\,\wedge \\ [(x,y):C(x,y|m_i)>B(m_i)]}} \frac{d_L(x,y\,|\,m_i)}{E(m_i)} \tag{6}$$

where

$$d_L(x,y\,|\,m_i) = \begin{cases} 1, & \text{if } \left|C(x,y\,|\,m_i) - I(x,y)\right| < a \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

and

$$E(m_i) = \sum_{(x,y)\in i} 1\,. \tag{8}$$

$E(m_i)$ is the number of pixels contained within the region i, (x, y) are pixel coordinates in the image region i, $C(x, y \mid m_i)$ is the value of the diffusion model encoded by the chromosome segment m_i, $I(x, y)$ is the intensity of the pre-processed image, $a = k \cdot I(i, j)$ and $0 < k \leq 1$ is constant.

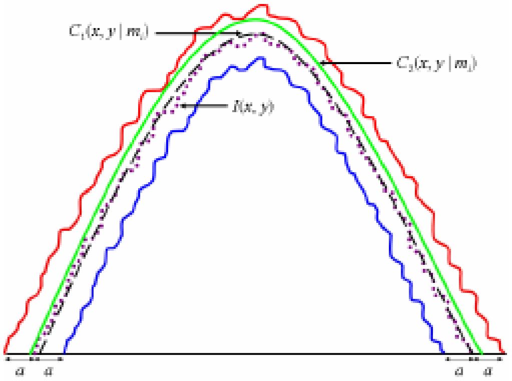


Fig. 5. The dotted curve represents the intensity of a real protein spot in a 2D-PAGE image. The dashed curve represents a diffusion model $C_1(x, y \mid m_i)$ that optimally fits the real protein spot. Chromosome segments m_i, i=1,2,…N encoding diffusion models such as $C_2(x, y \mid m_i)$, with values that fall within the margin defined by the outer curves have unitary *local fitness*.

The *local fitness* expresses the percentage of pixels of an image region i for which $C(x, y \mid m_i)$ differs from $I(x, y)$ less than a (Fig. 5). If $\left|C(x,y\,|\,m_i) - I(x,y)\right| < a$ and $C(x,y\,|\,m_i) > B(m_i)$ then $f_L(m_i) = 1$. The parameter a controls the tolerance of the *local fitness* to include as fittest solutions, models that approximate spots with irregularities or asymmetry, with arbitrary precision. Such spots may appear due to the imperfect diffusion of their substance across the gel medium.

The *central fitness* of the chromosome *m* is computed by the following equation:

$$f_C(m) = \sum_{[(x,y):C(x,y|m_1)>B(m_1)]} \frac{d_C(x,y\,|\,m)}{E(m_1)} \tag{9}$$

where

$$d_C(x,y\,|\,m) = \begin{cases} 1, & \text{if } \left|\sum_{j=1}^{N} S(x,y\,|\,m_j) - I(x,y)\right| < a \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

and

$$S(x,y\,|\,m_j) = \begin{cases} C(x,y\,|\,m_j), & \text{if } j = 1 \\ C(x,y\,|\,m_j) - B(m_j), & \text{if } \begin{cases} f_L(x,y\,|\,m_j) > T\,\cdot \\ \qquad\vee \\ j \neq 1 \end{cases} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

The *central fitness* expresses the percentage of pixels of the image region i for which $C(x,y\,|\,m_1) > B(m_1)$ and the superposition $S(x, y \mid m_i)$ of the adjacent regions differs from $I(x, y)$ less than a. In Eq.(11), T is a threshold for the *local fitness* of the adjacent spot models beyond which they can be included in $S(x, y \mid m_i)$.

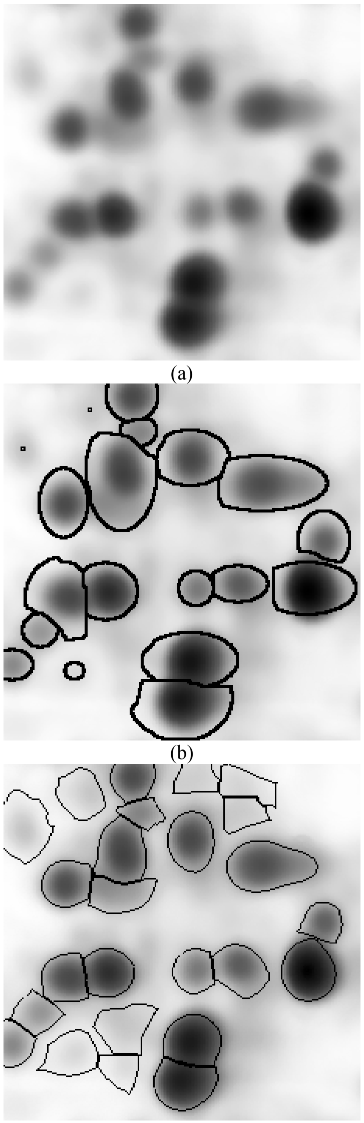
IV. Results

Experiments were performed to evaluate the performance of the proposed algorithm on a set of real 2D-PAGE images digitized at 2250×3000-pixels at 16-bit grey level depth. Each gel contains approx. 1500 spots. A subimage of a 2D-PAGE image used in the experiments is illustrated in Fig. 6a.

A population of 100 chromosomes was used, as uniform crossover has been observed to operate better when the population size is small [21]. In each generation of the genetic algorithm, 10% of the best chromosomes were maintained in the population, whereas the rest were reproduced by crossover and mutation operations.

In accordance with [22] a high crossover probability of 0.8 was chosen. Best results were achieved using an also high mutation probability of 0.8. In [23] it is suggested that the real-coded genetic algorithm may take advantage of high mutation rates. The reason is that the real-coded genetic algorithm does not provide enough diversity through the crossover operation alone. Mutation on the other hand can select a new real value within the allowable range of each designed gene of the chromosome. A threshold value T=0.4 (Eq.11) was found to be adequate for the adjacent spots to contribute in the computation of the *central fitness*.

The results of the experiments are summarized in Table I. It presents the spot detection performance of the proposed method in comparison with the spot detection performance of the Melanie 5 software package.


(a)

(b)

(c)

Fig.6. Indicative protein spot detection results: (a) input image, (b) output of the proposed approach, (c) output of the Melanie 5 software package.

TABLE I

SUMMARY OF RESULTS

Spots	Proposed method (%)	Melanie 5 (%)
True positive	95.3	95.3
False positive	**3.0**	8.6
Not detected	4.7	4.7

It can be observed that the percentage of the real spots (true positive) detected with the proposed method is comparable with that detected with Melanie. Both methods failed to detect 4.7% of the real spots. However, the percentage of spurious spots detected with the proposed method was clearly lower.

Example output images containing indicative spot detection results are illustrated in Fig. 6. This figure shows that the proposed method did not find any spurious spot whereas Melanie found 6 spurious spots. Both methods detected all the 19 real spots contained in the image. It should be noted that the two points appearing at the upper left corner of Fig. 6b indicate that two spots have been detected, but their boundaries have not been developed enough to capture the whole region of the spot.

V. CONCLUSION

We presented a novel method to detect and quantify protein spots in 2D-PAGE images based on a genetic algorithm. The genetic algorithm searches within a multidimensional parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots.

The proposed method has the following advantages: a) it does not require a training phase; b) it is capable of detecting overlapping spots; c) it is capable of detecting spots distorted by imperfect diffusion of the spot substance across the gel medium; d) compared with the state of the art commercial Melanie 5 software package results in clearly lower spurious spot detection rate.

Future work includes further experimentation, optimization and parallelization of the proposed method, and its integration in a complete user-friendly software application. Also variation of the proposed method will be applied to other biomedical data, such as microarrays.

REFERENCES

[1] R. Westermeier, *Electrophoresis in practice: A guide to theory and practice*, VCH, Weinheim 1993.

[2] B. Raman, A. Cheung, and M. R. Marten, “Quantitative comparison and evaluation of two commercially available, two dimensional electrophoresis image analysis software packages, Z3 and Melanie,” *Electrophoresis*, vol.23 no.14, 2002, pp. 194-200.

[3] P. F. Lemkin, and L. E. Lipkin, “Electrophoresis gel data base analysis: aspects of data structures and search strategies in GELLAB,” *Electrophoresis*, vol. 4, 1983, pp. 71–81.

[4] K. Conradsen, and J. Pedersen, “Analysis of two-dimensional electrophoresis gels,” *Biometrics*, vol. 48, 1992, pp.1273-1287.

[5] A. Efrat, F. Hoffmann, K. Kriegel, C. Schultz, and C. Wenk, “Geometric algorithms for the analysis of 2D-electrophoresis gels,” *Journal of Computational Biology*, vol 9, 2002, pp. 299-315.

[6] P. Cutler, G. Heald, I.R. White, and J. Ruan, “A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection,” *Proteomics*, vol. 3, 2003, pp. 392–401.

[7] R. D. Appel, J. Vargas, P. M. Palagi, D. Walther, and D. F. Hochstrasser, “MelanieII: a third generation software package for analysis of two-dimensional electrophoresis images-II Algorithms,” *Electrophoresis*, vol. 18, 1997, pp.2735–2748.

[8] K. P. Pleissner, F. Hoffmann, K. Kriegel, C.Wenk, C. Wegner, A Sahlstrohm, H Oswald, H.Alt, and E. Fleck, “New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gels,” *Electrophoresis*, vol. 20, 1999, pp.755–765.

[9] J. J. Tyson, and R.H. Haralick, “Computer analysis of two-dimensional gels by a general image processing system,” *Electrophoresis*, vol.7, 1986, pp.107-113.

[10] Y. Wu, P. Lemkin, and K. Upton, “A fast spot segmentation algorithm for two-dimensional gel electrophoresis analysis,” *Electrophoresis*, vol. 14, 1993, pp. 1351-1356.

[11] P.F. Lemkin, J. E. Myrick,and K. M.Upton, “Splitting merged spots

in two-dimensional polyacrylamide gel electrophoresis gel images," *Appl Theor Electrophor*, vol.3, 1993, pp. 163-72.

[12] J. I. Garrels, "The QUEST system for quantitative analysis of two-dimensional gels," *Journal of Biological Chemistry*, vol. 264, 1989, 5269-5282.

[13] E. Bettens, P. Scheunders, D. van Dyck, L. Moens,amd P. van Osta, "Computer analysis of two-dimensional electrophoresis gels: a new segmentation and modeling algorithm," *Electrophoresis*, vol. 18, 1997, pp. 792-798.

[14] M. Rogers, J. Graham, and R. P. Tonge, "Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images," *Proteomics*, vol. 3, 2003, pp. 887–896.

[15] K. Kaczmarek, B. Walczak, S. de Jong, and B.G.M. Vandeginste, "Preprocessing of two-dimensional gel electrophoresis images," *Proteomics*, vol. 4, 2004, pp. 2377–2389.

[16] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, Reading, MA, 1989.

[17] F. Herrera, M. Lozano, and J.L. Verdegay, "Tackling Real Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis," Artificial Intelligence Review, vol.12, 1998, pp. 265–319.

[18] G. Rudolph, "Convergence Analysis of Canonical Genetic Algorithms," *IEEE Trans. on Neural Networks*, vol. 5, pp. 96-101, 1994.

[19] G. Syswerda, "Uniform crossover in genetic algorithms," in Proc. *ICGA-89*, J. D. Schaffer, Ed. Morgan Kaufmann, pp. 2—9, 1989.

[20] S. Y. Wang, and K. Tai, "Graph Representation for Structural Topology Optimization Using Genetic Algorithms," *Computer & Structures*, vol. 82, 2004, pp. 1609-1622.

[21] A. Bhutyan, V. Ampornramveth, S. Muto, H. Ueno, "Face detection and facial localization for human-machine interface," *NII journal*, vol. 3, no .5, 2003, pp. 25-39

[22] M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, "Feature selection for computer-aided polyp detection using genetic algorithms," *Proceedings of SPIE*, vol. 5031, 2003, pp. 102-110.

[23] C. Z. Janikow, Z. Michalewicz, "An experimental comparison of binary and floating point representations in genetic algorithms," in *Proc. 4th International Conf. on Genetic Algorithms*, San Diego, 1991, pp. 31–6.