Feature Based Fuzzy Matching of 2D Gel Electrophoresis Images

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

Institute of Chemistry, Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland, and Unilever Research and Development Vlaardingen, Olivier van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands

Received March 25, 2002

Automatic alignment (matching) of two-dimensional gel electrophoresis images is of primary interest in the evolving field of proteomics. In the present study, feature-based matching techniques, in their classical and robust versions, are described, and an automatic method of fuzzy alignment (FA) is introduced. This method allows automatic matching of two gel images with different numbers of features with unknown correspondence. Performance of FA is tested on simulated and real data sets.

1. INTRODUCTION

In 2002, we witness the shift of research interests from genomics to proteomics. The term proteome embraces all the proteins expressed by a genome, and thus proteomics involves identification of proteins and determination of their physiological and pathophysiological importance.

Transfer of interests from genomics to proteomics is caused by the fact that most diseases manifest themselves at the level of protein activity, whereas genetic data alone is insufficient to predict protein profiles of healthy or diseased tissues. From each gene, multiple proteins can be obtained due to the post-translation modification and mRNA splicing. The genome remains to a large extent unchanged, but proteins in any particular cell can change dramatically, as genes are turned on and off in response to their environment.

Study of proteins is more difficult than study of nucleic acids, mainly due to the secondary and tertiary structure of proteins, their inability of amplification like DNA, the possibility of their denaturation, or poor solubility of some proteins.

The first step in a typical proteomics workflow is proteins separation. There are numerous techniques used to separate proteins derived from a tissue, but the current state of art is the two-dimensional (2D) electrophoresis.

Two-dimensional gel electrophoresis, introduced in 1975 by O'Farrel[1] combines two modes of electrophoretic separation. In the first direction, the molecules are separated according to their charges and in the second direction, according to their molecular weights.[2] The first direction is isoelectric focusing. The net charge of a molecule depends on the pH value of the surrounding media and at a specific pH value it equals zero. This pH value is called the isoelectric point of a molecule (pI). At the pH = pI, the molecule stops migrating because of its nil net charge. In the second direction, separation according to the molecular weight is obtained, usually by applying the strong anionic detergent SDS (sodium dodecyl sulfate) to form the electrically charged micelles and polyacrylamide gel (PAG) as electrophoretic bed. Hence this electrophoretic mode is denoted as SDS-PAGE. SDS makes proteins unfold and masks their charge, allowing separation according to molecular weights. After separation, proteins are visualized by staining with different dyes, reaction with silver or radioactive labeling.

Advances in 2D gel electrophoresis revolutionized proteomics by providing the means to separate different proteins for further analysis. According to Celis and Gromov[3] currently there is no substitute to the 2D gel technology, but, of course, there are a number of its areas to be improved. Major issues concerning 2D gels are discussed in detail in ref 4.

The future of proteomics is determined by the progress of automation (e.g. ref 5) and by the advances in information sciences. Informatics can enter the proteomics workflow at many stages (e.g. ref 4), one of them being spot detection and gel-to-gel matching. There are many software packages supporting this step of analysis,[6−11] but the majority of them require user guidance and allow only two gels to be compared. However, a proteomics-based study of a disease ought to rely on large sample numbers, to statistically validate the differences in protein abundance between the control and the diseased groups. A possibility of a simultaneous comparison of many gel images is also of great importance for the chemists for optimization of separation.

A far-reaching goal of our study is to develop an automatic approach to gels matching. In this part of our study the feature based approach of image matching is introduced and a possibility of automatic matching of feature sets with unknown correspondence is studied thoroughly on real and simulated data sets.

2. THEORY

2.1. Matching of 2D Gel Plates. There are many possible approaches to matching of the gel images. Their taxonomy can be based on the following characteristics:[12] (1) primitives selected for matching, (2) models (transforms) used for mapping between primitives, (3) similarity measure between primitives, and eventually (4) strategy selected to control matching algorithm.

* Corresponding author phone/fax: (+48-32)-25-99-978; e-mail: beata@tc3.ich.us.edu.pl.

[†] Silesian University.

[‡] Unilever Research and Development Vlaardingen.

10.1021/ci020266k CCC: $22.00 © 2002 American Chemical Society
Published on Web 10/05/2002

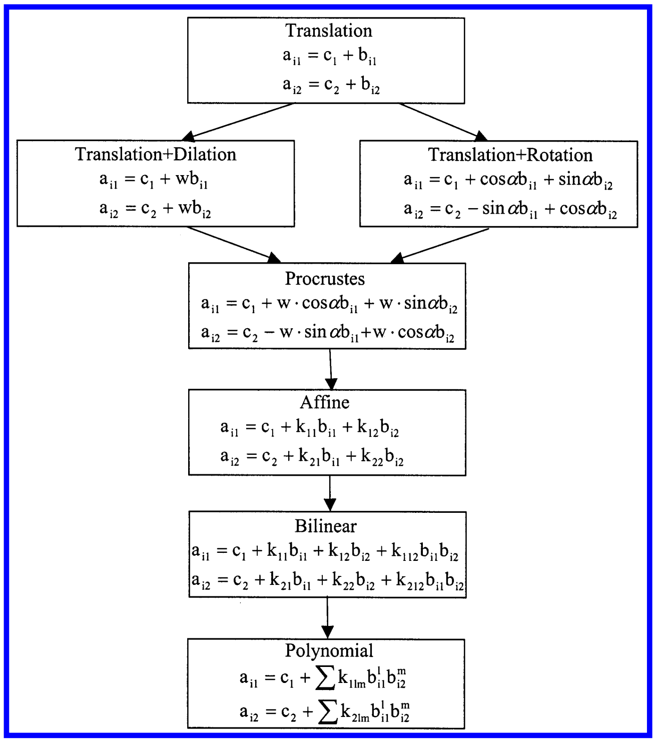
Translation
$a_{i1} = c_1 + b_{i1}$
$a_{i2} = c_2 + b_{i2}$
Translation+Dilation
$a_{i1} = c_1 + wb_{i1}$
$a_{i2} = c_2 + wb_{i2}$
Translation+Rotation
$a_{i1} = c_1 + \cos\alpha b_{i1} + \sin\alpha b_{i2}$
$a_{i2} = c_2 - \sin\alpha b_{i1} + \cos\alpha b_{i2}$
Procrustes
$a_{i1} = c_1 + w \cdot \cos\alpha b_{i1} + w \cdot \sin\alpha b_{i2}$
$a_{i2} = c_2 - w \cdot \sin\alpha b_{i1} + w \cdot \cos\alpha b_{i2}$
Affine
$a_{i1} = c_1 + k_{11}b_{i1} + k_{12}b_{i2}$
$a_{i2} = c_2 + k_{21}b_{i1} + k_{22}b_{i2}$
Bilinear
$a_{i1} = c_1 + k_{11}b_{i1} + k_{12}b_{i2} + k_{112}b_{i1}b_{i2}$
$a_{i2} = c_2 + k_{21}b_{i1} + k_{22}b_{i2} + k_{212}b_{i1}b_{i2}$
Polynomial
$a_{i1} = c_1 + \sum k_{1lm} b_{i1}^{l} b_{i2}^{m}$
$a_{i2} = c_2 + \sum k_{2lm} b_{i1}^{l} b_{i2}^{m}$

Figure 1. Hierarchy of parametric transformations. The arrows indicate transformations that generalize others;[14] a and b denote coordinates of images A and B, respectively, *i* denotes the consecutive objects, superscripts 1 and 2 refer to axes x- and y of the images; w, k, and c are regression coefficients, and α is the angle of rotation.

Primitives Selected for Matching. There are two types of primitives used for image matching: (1) windows composed of gray values or (2) features extracted in each image individually.

Depending on the type of primitives used in the matching procedure, the resulting algorithms are referred to as the area based matching (ABM) and the feature based matching (FBM). Each feature can be characterized by a set of attributes and its image coordinates.

In our study, the feature based matching will be considered. In FBM, an image A is converted into a matrix **A** with rows referring to the detected spots and containing their attributes such as coordinates, intensity, etc. In the present study, only coordinates of spots centers are taken into consideration, so matrices representing individual images have dimensionality (n × 2), where n denotes the number of detected spots. Further in the text, the term “feature” is used interchangeably with the term “spot” or “object”.

Models Used for Mapping between Features. Depending on the problem at hand, we can be interested in the so-called *registration* or *shape* parameters.[13] Registration parameters are associated with the object’s location, size, and rotation, while shape is defined as the entity of geometric information that remains when the location, scale, and rotational effect are filtered out from the object.

The registration parameters are of primary importance, if we are interested in the matching of different images, but they can be ignored if the main goal is estimation of an average shape and its variability in the objects population. If the primary goal is localization and identification of an object in the image, then both, registration and shape parameters, are important.

Registration parameters are—differently speaking—the parameters of the transformation t, which allows for transforming **B** to **A**

$$\mathbf{A} = \mathrm{t}(\mathbf{B}) + \mathbf{E} \tag{1}$$

where **E** denotes the residual matrix.

The estimates of transformation parameters can be calculated by minimizing an objective function, for instance, the sum of squared residuals:

$$\text{minimize } (||\mathbf{E}||^2) = \text{minimize}(||\mathbf{A} - \mathrm{t}(\mathbf{B})||^2) \tag{2}$$

Then the adequacy of the match can be expressed as

$$\mathrm{d}^2(\mathbf{A},\mathbf{B}) = \sum_i \sum_j (\mathbf{E}^2) \tag{3}$$

Transformations. Among the popular transformations there are, e.g., translation, isotropic scaling, and orthogonal rotation (and also reflection). These transformations do not effect the object’s shape. For objects localization and identification, affine linear and nonlinear transformations can be applied as well.

These transformations can be considered in some hierarchy, where the consecutive transformations are generalizations of the previous ones (see Figure 1).[14] The simplest transformation is translation

$$\mathbf{A} = \mathbf{1}_n\mathbf{c}^T + \mathbf{B} + \mathbf{E} \tag{4}$$

where $\mathbf{1}_n$ is a (n × 1) unit vector, **c** is translation vector (2 × 1), and **E** denotes the (n × 2) residual matrix.

Translation can be accompanied by dilation

$$\mathbf{A} = \mathbf{1}_n\mathbf{c}^T + \mathrm{w}\mathbf{B} + \mathbf{E} \tag{5}$$

where w is the scale or by rotation

$$\mathbf{A} = \mathbf{1}_n\mathbf{c}^T + \mathbf{BT} + \mathbf{E} \tag{6}$$

where **T** is an orthogonal rotation matrix (2 × 2).

Transformation including translation, isotropic scaling, and orthogonal rotation simultaneously

$$\mathbf{A} = \mathbf{1}_n\mathbf{c}^T + \mathrm{w}\mathbf{BT} + \mathbf{E} \tag{7}$$

is called the Procrustes analysis.

Generalization of the Procrustes analysis is affine transformation, defined as

$$\mathbf{A} = [\mathbf{1}_n\mathbf{B}]\mathbf{T} + \mathbf{E} \tag{8}$$

where **T** is (3 × 2) affine transformation matrix. Its elements are calculated via familiar least squares regression of **A** on $[\mathbf{1}_n\mathbf{B}]$:

$$\mathbf{T} = ([\mathbf{1}_n\mathbf{B}]^T [\mathbf{1}_n\mathbf{B}])^{-1} [\mathbf{1}_n\mathbf{B}]^T \mathbf{A} \tag{9}$$

More general approaches include quadratic transformation or bilinear transformation, etc.[14] The most general transformation preserving object shape is Procrustes analysis.[13,15] In this type of analysis, data sets **A** and **B** are centered and scaled to unit variance ($\hat{\mathbf{A}}$,$\hat{\mathbf{B}}$) and then rotated. Parameters of rotation matrix, **T**, cannot be estimated as linear regression

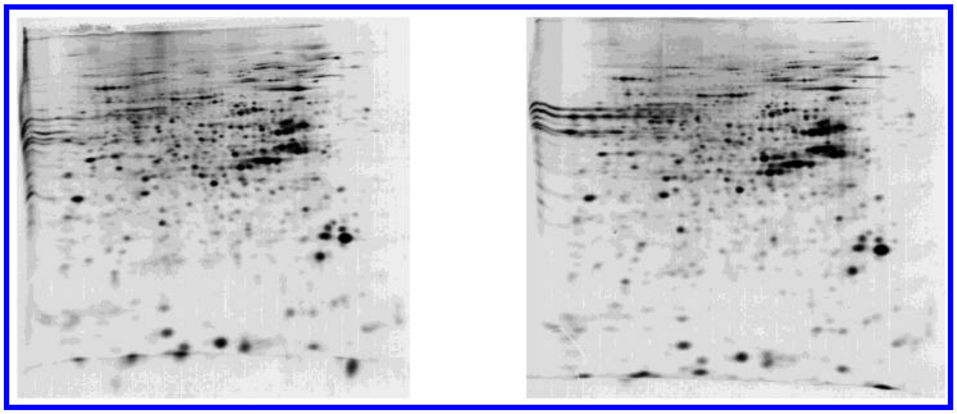

Figure 2. 2D electrophoresis gels for human skin, silver stained.[16]

parameters due to the orthogonality constraint, i.e., $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, but they can be determined using Singular Value Decomposition (SVD) of covariance matrix $\hat{\mathbf{A}}^T\hat{\mathbf{B}}$, and namely if svd($\hat{\mathbf{A}}^T\hat{\mathbf{B}}$) = $\mathbf{USV}^T$ then rotation matrix $\mathbf{T} = \mathbf{UV}^T$.

For two-dimensional data sets, Procrustes analysis can be simplified (and limited to linear regression), due to the possibility of working on complex numbers. Complex numbers are an algebraic way of coding points on an ordinary Euclidean plane so that translation (shift of position) corresponds to the addition of complex numbers and both rescaling (enlargement or shrinking) and rotation corresponds to multiplication of complex numbers. In this system of notation, the *x*-axis is identified with “real numbers” (ordinary decimal numbers) and the *y*-axis is identified with “imaginary numbers” (the square roots of negative numbers).

If data sets **A** and **B** are represented as complex numbers, **a** and **b**, then the regression equation has the following form

$$\mathbf{a} = \mathrm{w}\mathbf{b}e^{i\alpha} + \mathbf{1}_n c + \mathbf{e} \qquad (10)$$

where c is location (complex number), w is the scale (real number), $\alpha \in [0\ 2\pi)$ is the angle of rotation, and **e** is the complex n-vector of residuals.

For two-dimensional data sets, angle of rotation can also be estimated, based on the following equation

$$\alpha = \arctan\left(\frac{\sum_{i=1}^{n}(a^c_{i2}b^c_{i1} - a^c_{i1}b^c_{i2})}{\sum_{i=1}^{n}(a^c_{i1}b^c_{i1} - a^c_{i2}b^c_{i2})}\right) \qquad (11)$$

where

$$\mathbf{a}_i{}^c = \frac{\mathbf{a}_i - \bar{\mathbf{a}}}{\mathbf{s}_\mathbf{A}} \text{ and } \mathbf{b}_i{}^c = \frac{\mathbf{b}_i - \bar{\mathbf{b}}}{\mathbf{s}_\mathbf{B}} \qquad (12)$$

$$\bar{\mathbf{a}} = \frac{\sum_{i=1}^{n}\mathbf{a}_i}{n_1} \text{ and } \bar{\mathbf{b}} = \frac{\sum_{i=1}^{n}\mathbf{b}_i}{n_2} \qquad (13)$$

$$\mathbf{s}^2_\mathbf{A} = \sum_{i=1}^{n_1}||\mathbf{a}_i - \bar{\mathbf{a}}||^2 \text{ and } \mathbf{s}^2_\mathbf{B} = \sum_{i=1}^{n_2}||\mathbf{b}_i - \bar{\mathbf{b}}||^2 \qquad (14)$$

Similarity Measures. Similarity measures play an important role in each matching algorithm. In the case of FBM, similarity must be based on the attributes of the features used. In most cases the differences in the geometric and radiometric attribute values (i.e. in coordinates and intensities) are combined to construct a cost function.

In the case of Procrustes analysis, two types of distances are introduced:[13] the full Procrustes distance

$$d_1{}^2 = ||\hat{\mathbf{A}} - (\mathbf{1}_n\mathbf{c} + \mathrm{w}\hat{\mathbf{B}}\mathbf{T})||^2 = 1 - \mathrm{trace}(\mathbf{\Lambda})^2 \qquad (15)$$

and the partial Procrustes distance

$$d^2 = ||\hat{\mathbf{A}} - (\mathbf{1}_n\mathbf{c} + \hat{\mathbf{B}}\mathbf{T})||^2 = 2(1 - \mathrm{trace}(\mathbf{\Lambda})) \qquad (16)$$

where $\mathbf{\Lambda}$ is a 2 × 2 matrix with the elements given by the singular values of $\hat{\mathbf{A}}^T\hat{\mathbf{B}}$ where $\hat{\mathbf{A}}$,$\hat{\mathbf{B}}$ are preshapes of **A** and **B**, i.e., matrices **A** and **B** after centering and (isotropic or overall) scaling; trace(.) is the trace of a matrix, i.e., the sum of all diagonal elements.

The term “full” is used, if minimization is over a full set of similarity transformations (translation, scaling, and rotation), whereas the term “partial” is used if minimization is only over translation and rotation.

2.2. Visual Alignment of Corresponding Spots. Efficient as it is, Procrustes Analysis (or any other transform) of course requires that the corresponding objects (those to be matched) are known. In practice, this condition is the most difficult one to fulfill. Even a comparison of two 2D electrophoresis images of the same biological material separated under the similar experimental conditions can lead to some difficulties, as an appearance of any particular spots can change and their automatic assignment is difficult. In Figure 2, two 2D electrophoresis images of human skin are presented.[16] As one can notice, resolution of these images differs to some extent and the spot alignment is required.

The situation becomes more complex, when the 2D electrophoresis images represent separation of related but different biological material. In this case, only a part of spots is expected to be the same. The analogous situation is faced, if a particular image has to be compared with the images from the available databases. In a majority of the available approaches, spot alignment is performed by a user, i.e., the available software is highly interactive.[16,17] The user is

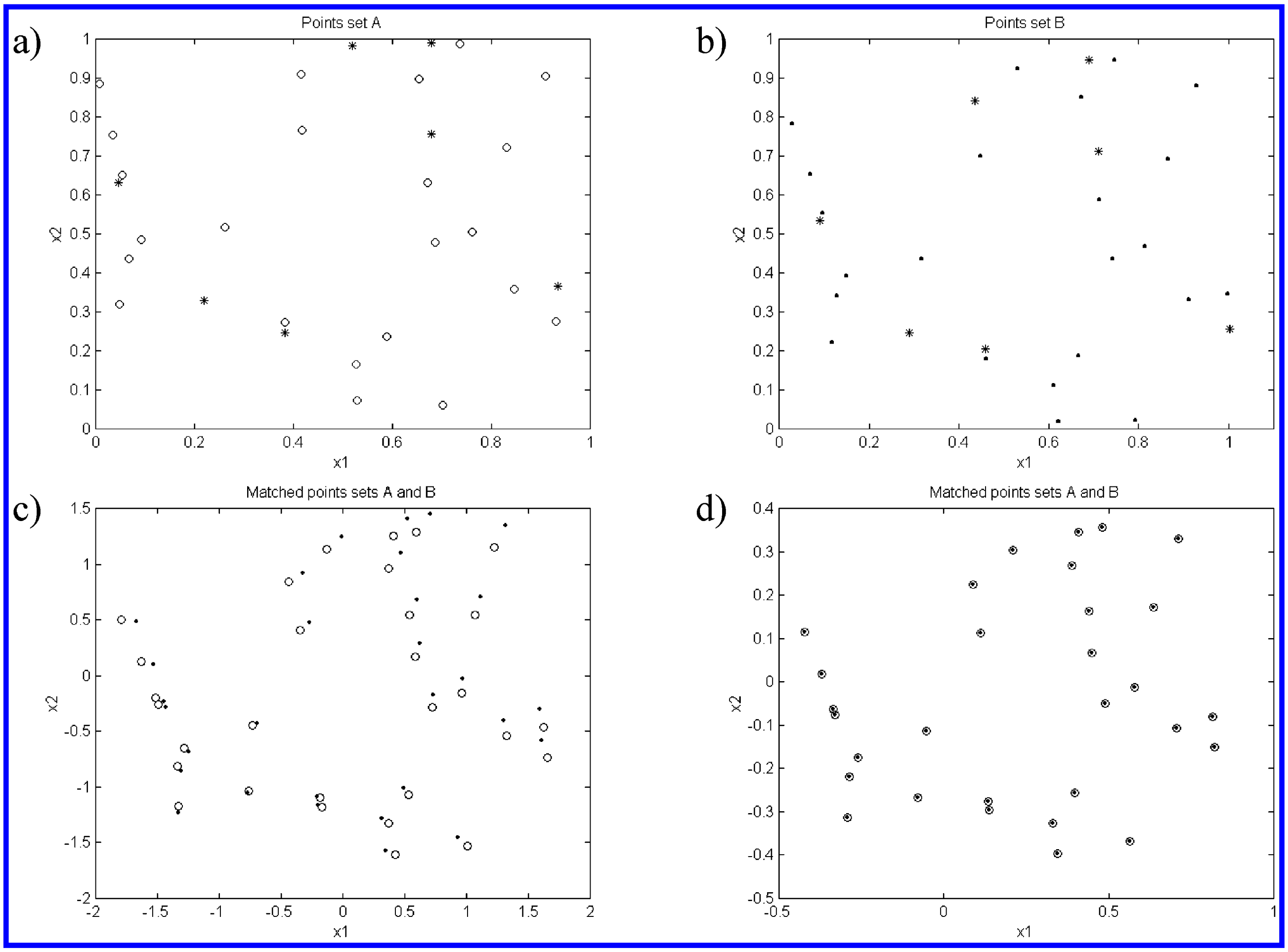

Figure 3. Spot patterns to be matched: (a) set **A** and (b) set **B** with the selected landmarks; (c) results of the Procrustes analysis; and (d) results of the robust Procrustes analysis. The selected spots are denoted as an asterisk (*).

obliged to mark at least three common (i.e. corresponding) spots and then the transform is performed on them. Estimation of transformation parameters is, of course, more successful, if more than three corresponding spots are used. Visual selection of a higher number of the corresponding spots can lead to mistakes and any least squares transform method is very sensitive to outliers. For instance, the alignment of spots presented in Figure 3a,b leads to the results visualized in Figure 3c. Majority of the corresponding spots was detected well, but there are wrong alignments too. Procrustes transformation determined by this set of spots leads to a wrong warping of images.

To overcome a pinpointed problem, robust versions of transform ought to be applied.

Robust Matching. To eliminate the problem with wrongly selected landmarks, it is possible to use a robust version of a transformation of interest. The type of the applied transform determines the minimal number of landmarks, necessary to calculate transform parameters. In the case of the robust version, the number of necessary landmarks has to be increased by a number depending on the assumed fraction of the noncorresponding landmarks (wrongly selected spots or outliers).

Without any loss of generality, an idea of robust matching will be further presented for Procrustes Analysis. As the minimal number of landmarks for Procrustes Analysis is three and assuming that half of the landmarks can be selected wrongly, at least six landmarks are required for the robust version. Robust Procrustes Analysis is performed by selecting the best parameters of transform, calculated for all possible combinations of the three out of six landmarks.

For each combination of the three landmarks, scaling, translation, and rotation are calculated (just like in the plain Procrustes Analysis) and transform of all features of the set **B** is performed, leading to the set t(**B**). Matching efficiency is then determined by robust measure of similarity between the two point sets, **A** and t(**B**), and namely by the partial Hausdorff distance.[18]

Calculation of the partial Hausdorff distance starts with construction of the distance matrix **D** ($n_1 \times n_2$), where n_1 and n_2 denote numbers of objects in the sets **A** and t(**B)**, respectively. The ij-th element of this matrix, i.e., $d_{ij,}$ represents the Euclidean distance between the i-th point from data set **A** and the j-th point from data set t(**B**). Thus each row of **D**, i.e., $\mathbf{d}_i$, consists of the distances between the i-th point from set **A** and all points from data set t(**B**). Minimal element of vector $\mathbf{d}_i$ ($1 \times n_2$) represents the distance of the i-th object from set **A** to its nearest neighbor from set **B**. The set of nearest neighbor distances, (i.e. vector **nnd** ($n_1 \times 1$)) obtained for n_1 objects from set **A** is then sorted, and the k-th ranked value of the nearest neighbor distances is used as a robust measure of transform performance.

Application of robust Procrustes Analysis to the data presented in Figure 3a,b leads to the results presented in Figure 3d.

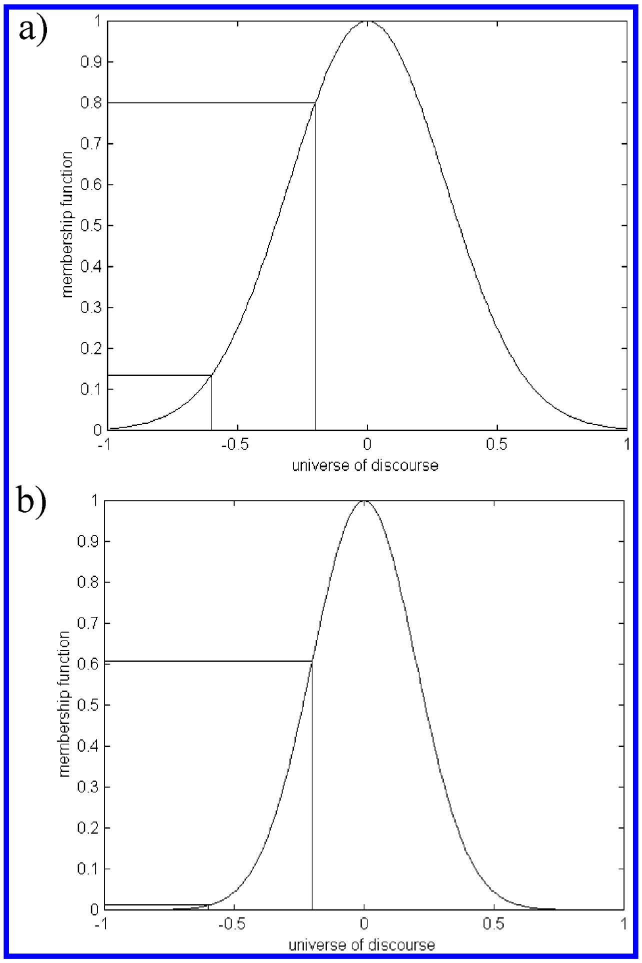

Figure 4. Illustration of an idea of fuzzy matching, using the Gaussian function with different values of σ; (a) $\sigma = 0.3$ and (b) $\sigma = 0.2$; Gaussian function is centered on one feature from set **A** (0 on abscissa). For the features from set **B** (-0.6 and -0.3 on abscissa), values of this function depend on the degree of fuzziness, i.e., on the σ value of the Gaussian function.

Table 1. Simulated Sets A and B

set **A**		set **B**	
x	y	x	y
0.9501	0.4057	1.0050	0.4944
0.2311	0.9355	0.1916	0.8631
0.6068	0.9169	0.5629	0.9230
0.4860	0.4103	0.5501	0.4023
0.8913	0.8936	0.8460	0.9594
0.7621	0.0579	0.8934	0.1151
0.4565	0.3529	0.5331	0.3401
0.0185	0.8132	0.0090	0.6992
0.8214	0.0099	0.9614	0.0804
0.4447	0.1389	0.5661	0.1283
0.6154	0.2028	0.7198	0.2263
0.7919	0.1987	0.8933	0.2590
0.9218	0.6038	0.9361	0.6823
0.7382	0.2722	0.8255	0.3197
0.1763	0.1988	0.2911	0.1311
		0.9981	0.4028
		0.1796	0.6095
		0.8413	0.3289
		0.5395	0.2046
		0.9918	0.0897

2.3. From Visual to Automated Spots Alignment and Matching. When features are derived from the landmarks, there is no problem of a correspondence between the spots, but if features are derived for the two images automatically, then an additional problem of correspondence between the unlabeled features appears. Things become even more complicated, if the numbers of the extracted features to a certain extent differ.

Now let us assume that gel A is the target gel (standard), with which gel B is to be compared. It means that gel B ought to be transformed to match gel A. Correspondence between the features of **A** and **B** can be represented in the form of a binary matrix **M** ($n_1 \times n_2$), where n_1 and n_2 denote the number of features in the two considered images:

$$m_{ij} = \begin{matrix} 1 \text{ if spot } \mathbf{a}_i \text{ corresponds with spot } \mathbf{b}_j \\ 0 \text{ otherwise} \end{matrix} \tag{17}$$

Matrix **M** ought to fulfill certain requirements, and namely, each row and each column ought to contain only one “1” element, to ensure the one-to-one correspondence between the features from **A** and **B**. In the other words, one spot from image A ought to correspond with only one spot from image B and vice-versa. This matrix can be constructed in an iterative way, using an algorithm that is a combination of correspondence search and spatial mapping.

The idea of this approach will be presented for the Procrustes mapping.[19] To perform the Procrustes analysis, we should center and standardize the features of **A** and **B**. As the number of features can differ and a correspondence is expected for a part of spots only, the weighted mean and variance ought to be used, defined as

$$\bar{\mathbf{a}} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}\mathbf{a}_i}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}} \tag{18}$$

$$\bar{\mathbf{b}} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}\mathbf{b}_i}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}} \tag{19}$$

$$s_{\mathbf{A}}^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}||\mathbf{a}_i - \bar{\mathbf{a}}||^2 \tag{20}$$

$$\bar{s}_{\mathbf{B}}^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}||\mathbf{b}_i - \bar{\mathbf{b}}||^2 \tag{21}$$

The parameters of correspondence are also required to calculate a rotation angle:

$$\alpha = \arctan\left(\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}(a_{i2}^c b_{i1}^c - a_{i1}^c b_{i2}^c)}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} m_{ij}(a_{i1}^c b_{i1}^c - a_{i2}^c b_{i2}^c)}\right) \tag{22}$$

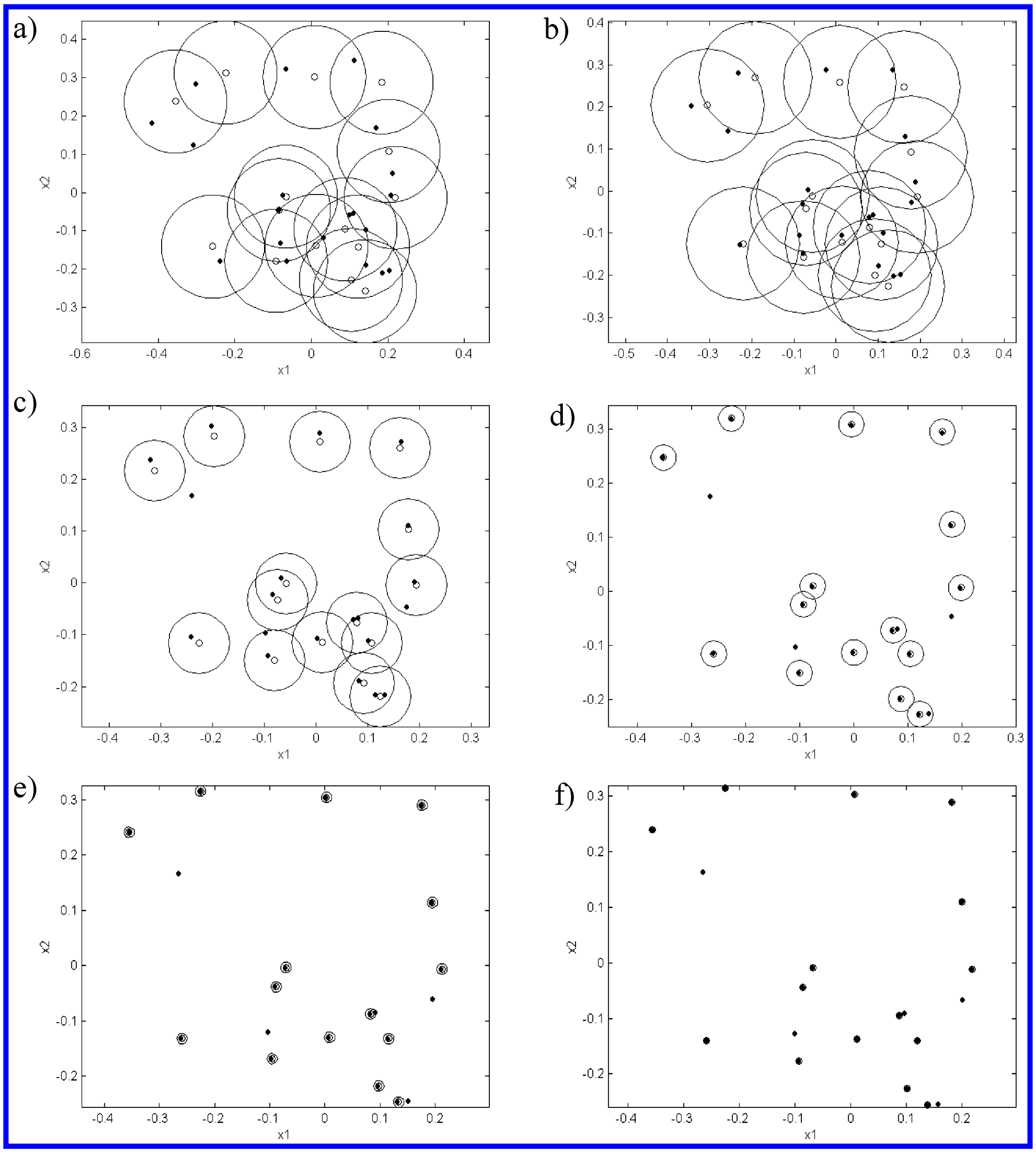

Figure 5. Degrees of fuzziness, determined by the σ of the Gaussian function, and represented by the different radii of the circles for (a) the initial sets, and those after: (b) 1 iteration, (c) 10 iterations, (d) 20 iterations, (e) 30 iterations, and (f) 40 iterations.

To properly calculate the correspondence between features, we have to transform image B to image A. The solution to this problem can be an iterative algorithm alternating between correspondence and mapping:

1. For a given correspondence matrix calculate parameters of transform.

2. For transformed images update correspondence matrix and go to 1, till the convergence is attained.

The main problem with iterative algorithms consists of their parameters' initialization. The elements of matrix **M** ought to somehow represent similarity between the spots **A** and **B**.

To calculate the elements of matrix **M**, we propose to use fuzzy matching between all pairs of features with varying degrees of fuzziness, i.e., starting with a high degree of fuzziness and decreasing it in the consecutive steps of the algorithm. In each iteration, the results of fuzzy matching are collected in matrix **G**. The *i*-th row of matrix **G** is an output of the Gaussian function centered at $\mathbf{a}_i$, where $i = 1,2,...,n_1$, calculated for all n_2 features of **B**

$$g_{ij} = \exp\left(\frac{-\sum_{k=1}^{2}(\mathbf{a}_i - \mathbf{b}_j)^2}{\sigma^2}\right) \tag{23}$$

where $j = 1$: n_2, and σ denotes the "radius" (or width) of the Gaussian function, chosen as a membership function.

The shorter the distance between feature $\mathbf{b}_j$ and the function center, i.e., $\mathbf{a}_i$, the higher is the membership value. The idea of fuzzy matching for one-dimensional data is illustrated in Figure 4.

A high value of σ causes a high degree of fuzziness, i.e., for many spots belonging to B, the elements of matrix **G** attain high values and each feature $\mathbf{b}_j$ can considerably belong to different centers $\mathbf{a}_i$. Decreasing σ leads to a more precise, i.e., less fuzzy assignment of features $\mathbf{b}_j$.

The elements of matrix **G** cannot be used, however, as the weight factors in the above equations, because it can

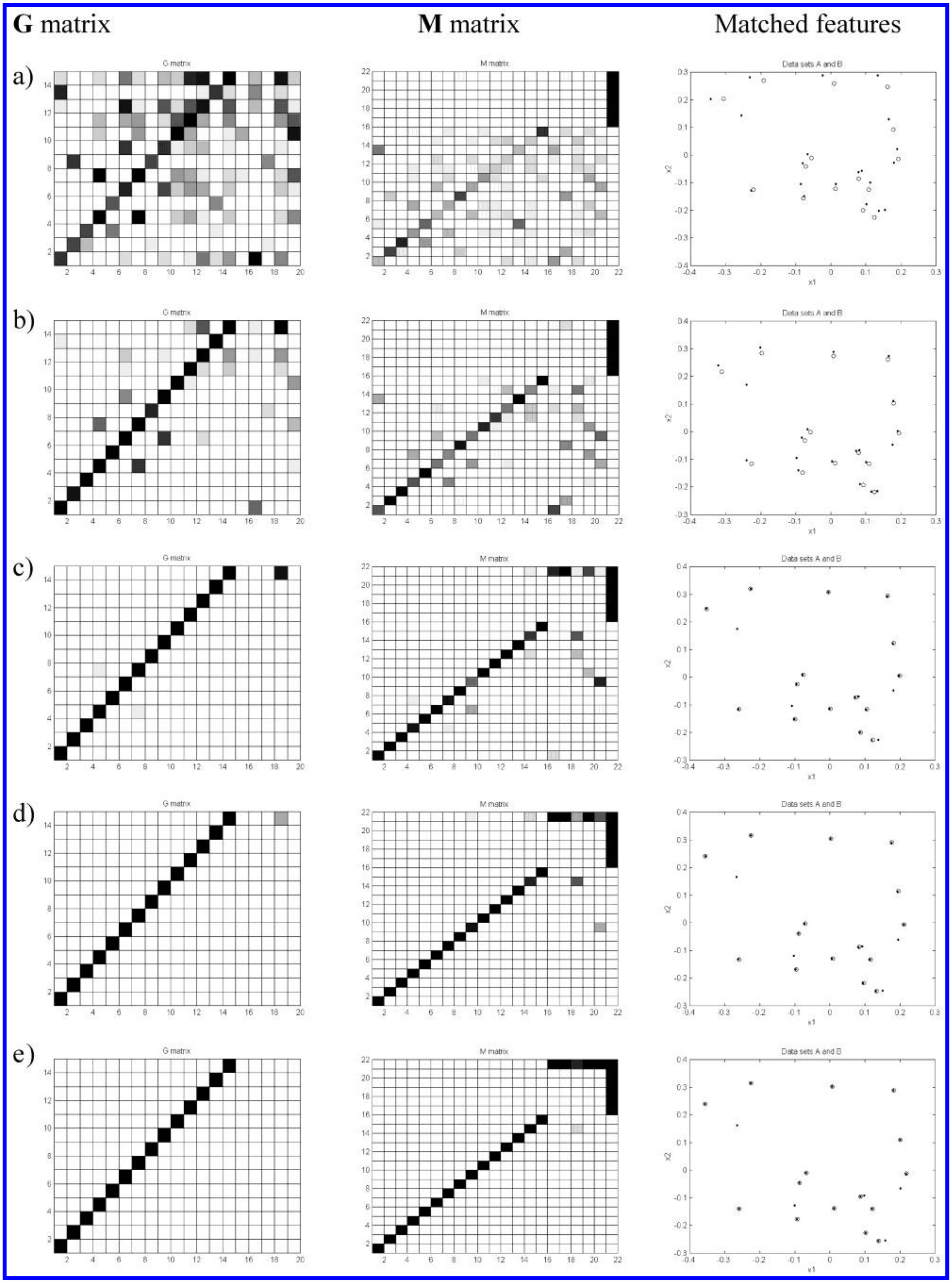

Figure 6. Changes in matrices **G** and **M** and the results of the matching of images during the analysis, (a) the initial values, (b) those after 10 iterations, (c) 20 iterations, (d) 30 iterations, and (e) 40 iterations (final values); each element of matrices **G** and **M** is represented by a corresponding pixel, with a color varying from white (0) to black (1).

happen that two or more spots of **B** are assigned in equal proportions to the same spot of **A**. To avoid this problem, a process of alternating row and column normalization can be applied to matrix **G** in order to obtain matrix **M**. According

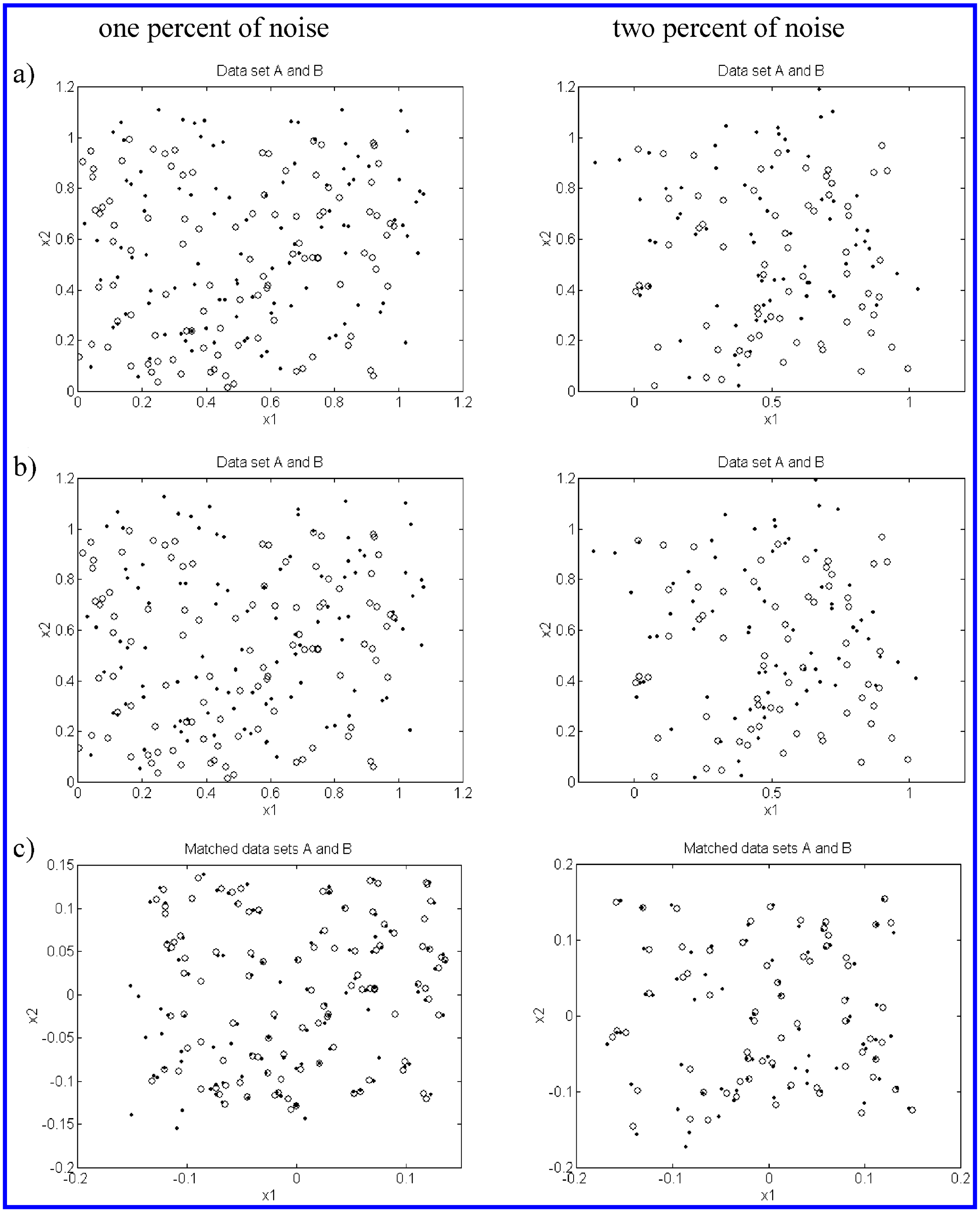

Figure 7. Examples of simulated data sets without (a) and with added noise (b). In subplots (c) the same data sets after the alignment.

to the theorem of Sinkhorn,[20] applying this procedure to any square matrix with positive elements leads to a so-called double stochastic matrix, i.e., to a matrix with the rows and columns summing to 1.

The main steps of an iterative automated matching are as follows:

0. Initialize degree of fuzziness, i.e., value of σ

1. Construct matrix **G** ($n_1 \times n_2$), the elements thereof representing outputs of the Gaussian functions, centered at the features of **A**, for all the features of **B**

2. Use Sinkhorn's theorem to obtain "a double stochastic matrix" **M** ($\max(n_1,n_2)+1$, $\max(n_1,n_2)+1$)

3. For a given correspondence (**M**) calculate parameters of Procrustes analysis, i.e., perform matching

4. If convergence is not achieved, change the degree of fuzziness and return to step 1

To describe step 2 of the above algorithm, some additional details are required. To accommodate the fact that in images A and B there are noncorresponding features, normalization procedures are performed for matrix **M** with an additional

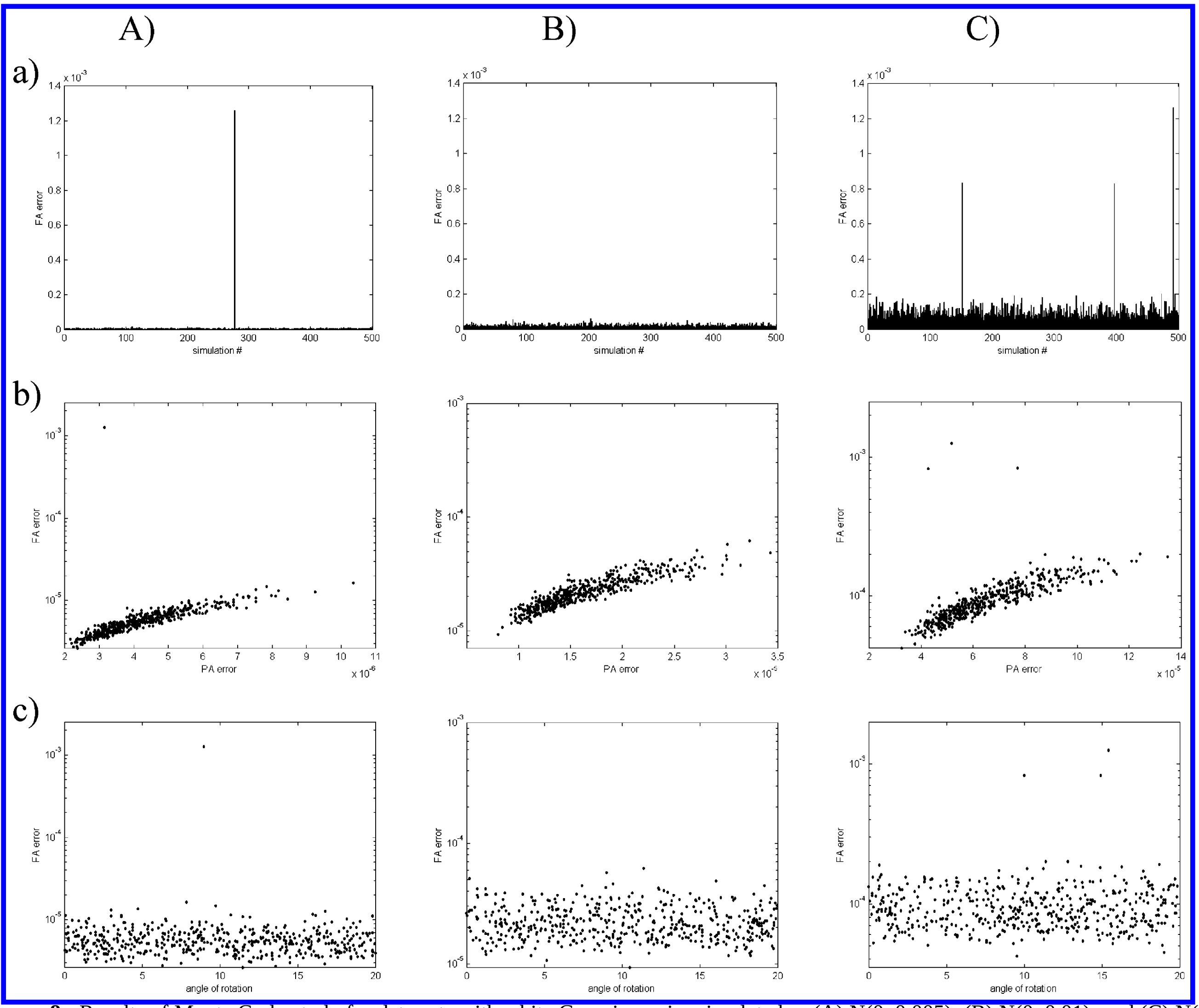

Figure 8. Results of Monte Carlo study for data sets with white Gaussian noise simulated as (A) N(0, 0.005); (B) N(0, 0.01); and (C) N(0, 0.02); (a) error of Fuzzy Alignment; (b) error of Procrustes Analysis (PA) versus error of Fuzzy Alignment (FA); and (c) plot of FA error versus the angle of rotation.

row and column. Features of **B**, which lack correspondence with any features of **A** attain a highest value of correspondence in an additional row, whereas the features of **A**, which do not correspond with any feature of **B**, will attain the value of 1 in an additional column of **M**.

3. DATA

Performance of the automated fuzzy alignment (FA) of the spots was studied for real and simulated data sets. The real data sets were loaded from databases available via the net,[21] whereas the Monte Carlo study was conducted for the simulated pairs of features sets, containing different fractions of corresponding spots and representing different degrees of deformation.

An individual pair of feature sets, **A** and **B**, was simulated as follows:

1. randomly select the number of corresponding spots (features), n (n $\in$ [50−100])

2. randomly generate coordinates from the range [0−1] for the n features (set **A**)

3. randomly select the parameters of transformation: angle of rotation ($\alpha \in$ [5−20°]), translation([0−0.2]), and scaling factor ([0.9−1.1])

4. transform **A** to obtain **B**

5. to set **B** add white Gaussian noise with predefined standard deviation, σ, simulated as N(0,σ)

6. randomly select two integer numbers r_1 and r_2, both from the range [1, n/2]

7. to the sets **A** and **B,** respectively, randomly add the generated coordinates r_1 and r_2 (from the range [0,1])

8. randomly order all (i.e., $n_2 = n + r_2$) features of **B.**

4. RESULTS AND DISCUSSION

Example of Automated Fuzzy Alignment (FA) of Simulated Features. Let us trace the convergence of the discussed FA for the simulated sets **A** and **B**, containing 15 and 20 features, respectively (which are presented in Table 1). All spots of the set **A** have their counterparts in **B**, but in **B** there are also five additional spots. For an illustrative purpose, the corresponding spots in **B** are placed in the first 15 positions.

An initial σ value of the Gaussian function can be estimated, based on the following equation

$$\sigma_1 = \frac{1}{4}\sqrt{\mathrm{range}(\mathbf{f}_1)\cdot\mathrm{range}(\mathbf{f}_2)/\pi} \qquad (24)$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ denotes vectors of spots' coordinates for more numerous set (**F**), i.e., in the example **F** = **B**.

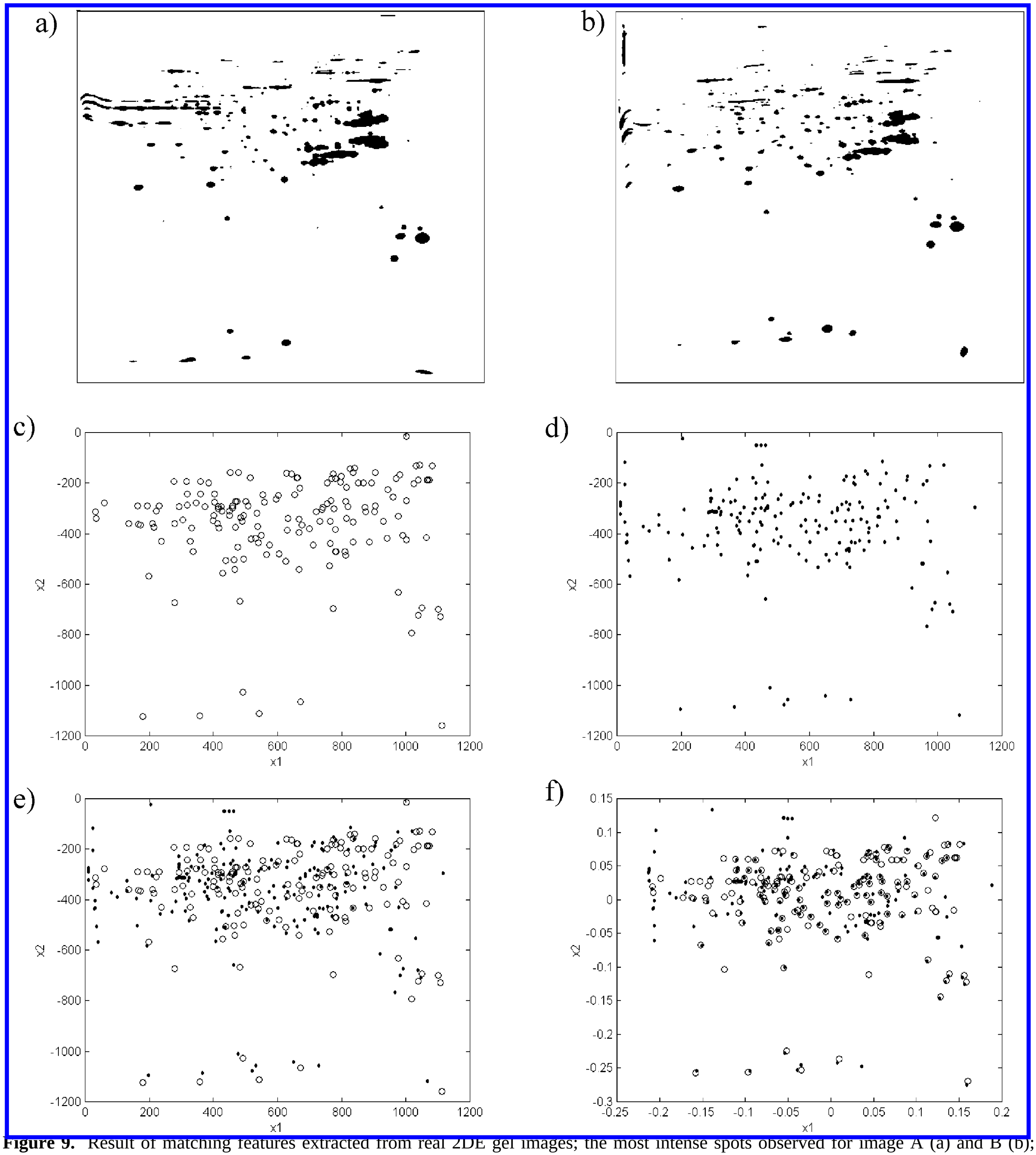

Figure 9. Result of matching features extracted from real 2DE gel images; the most intense spots observed for image A (a) and B (b); features corresponding to the most intensive spots for image A (c) and B (d); and superimposed features before (e) and after alignment (f).

For the example data set, σ equals 0.1084 and in each iteration it decreases, as given below:

$$\sigma_{\text{iter}+1} = \sigma_{\text{iter}} \cdot 0.833 \tag{25}$$

For the two-dimensional data sets, changes in the degree of fuzziness can be presented in the form of contour plots of Gaussian functions centered on features of **A**. In Figure 5, contour plots of 15 Gaussian functions, centered on 15 features of **A**, drawn for σ_0, σ_1, σ_{10}, σ_{20}, σ_{30}, and σ_{40} are presented.

Matrix **G** for the sets **A** and **B** has the dimensionality of 15 × 20 and represents 20 outputs (corresponding to the 20 objects from the set **B**) of the Gaussian functions centered at 15 objects from the set **A**. Matrix **M** of the dimensionality 21 × 21 represents the results of the Sinkhorn procedure. The main difference between matrices **G** and **M** consists of the fact that all rows and all columns of **M** sum up to 1, which is not the case with matrix **G**.

The initial matrices **G** and **M** and the matrices after 10, 20, 30, and 40 iterations are presented in form of the gray valued images in Figure 6. Due to an initial high degree of fuzziness there are many elements of matrix **G** with high values, but this matrix quickly converges to a binary matrix with the ones for the 15 corresponding spots and the zeros

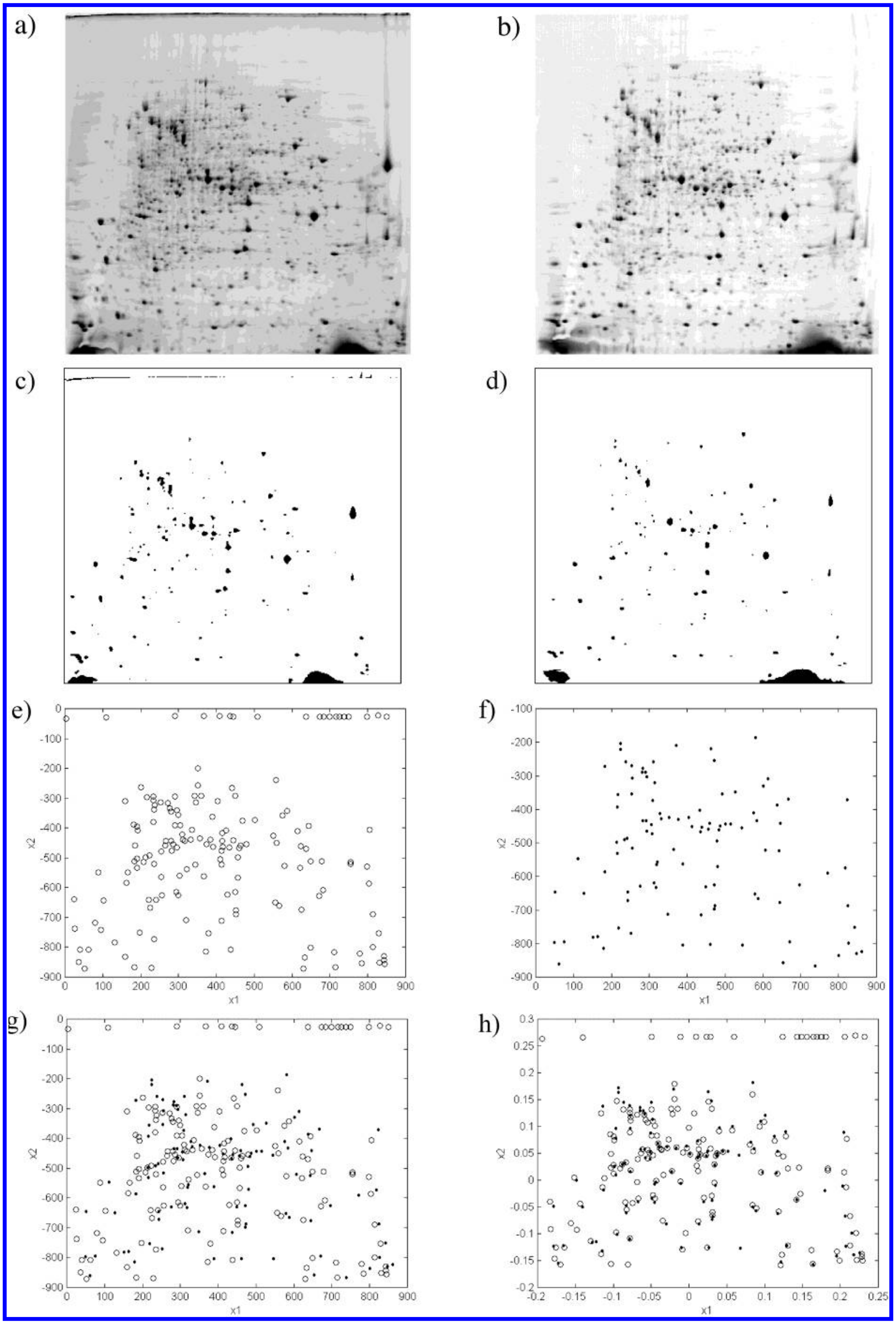

Figure 10. Result of matching features extracted from real 2DE gel images (a, b); the most intense spots observed for image A (c) and B (d); features corresponding to the most intensive spots for image A (e) and B (f); superimposed features before (g) and after alignment (h).

Table 2. Results of Monte Carlo Study[a]

parameters, mean values	noise level 0.005	noise level 0.01	noise level 0.02
n	74.84	75.71	75.66
n_1	18.93	19.47	18.81
n_2	18.95	19.45	19.98
angle	9.79	10.20	10.18
error of PA	$0.40 \bullet 10^{-5}$	$1.60 \bullet 10^{-5}$	$6.40 \bullet 10^{-5}$
error of FA	$0.58 \bullet 10^{-5}$	$2.21 \bullet 10^{-5}$	$9.60 \bullet 10^{-5}$
no. of iterations	53.21	46.07	39.63

[a] n, n_1, and n_2 denote number of corresponding spots in matched patterns, number of spots in **A**, and number of spots in **B**, respectively.

elsewhere. Five spots in **B**, having no corresponding spots in **A**, reach the approximate values of unity and group in an additional row of matrix M, indicating their outliness.

In the consecutive iterations, features **B** are transformed to match features **A**, with the weights described by the elements of matrix **M**.

Results of Monte Carlo Study. The proposed automated FA method was intensely tested for simulated pairs of data sets (**A**, **B**). The examples of simulated sets with different levels of the white Gaussian noise, before and after the alignment are presented in Figure 7.

Performance of matching is expressed in form of the mean squared distance between pairs of corresponding spots of images A and B, i.e.

$$\text{error of alignment} = \frac{\sum_{k=1:500} d_k}{500} \tag{26}$$

where

$$d_k = \frac{1}{n} \sum_{i=1:n} \sum_{j=1:2} (a_{ij} - t(b_{ij}))^2 \tag{27}$$

Once error of alignment is calculated for Procrustes Analysis (PA) performed for known corresponding features, the second time for Fuzzy Alignment (FA) with unknown features' correspondence.

Final results of Monte Carlo study for 500 pairs of data sets with varying degree of white, Gaussian noise are summarized in Table 2 and presented in Figure 8.

Mean values of the PA and FA errors observed for different σ of noise (see Table 2) give evidence of good performance of the discussed approach. The PA error, calculated for the known corresponding features is slightly lower than the FA error, calculated for the features with unknown correspondence. As visualized in Figure 8a, alignment of features with unknown correspondence failed for 4 out of 1500 cases only. In one case it happened for the data set with the lowest level of noise, and three times for the data sets with the highest levels of noise studied. All failures are probably caused by a too high number of the noncorresponding features. A decreasing number of iterations observed for the increasing σ is caused by the convergence criterion applied. Namely, the FA algorithm is stopped when any of the 50% highest outputs of the Gaussian functions starts decreasing. As shown in Figure 8c, the FA error is independent of the rotation angle.

An analogous study was performed for real data sets. Real data sets usually contain much higher numbers of features than the presented simulated data. To speed up the FA algorithm, one can calculate the parameters of transform using a limited number of features only, namely features corresponding to the most intense spots. For the sake of example two real patterns before and after fuzzy alignment are presented in Figures 9 and 10.

5. CONCLUSIONS

The proposed fuzzy alignment of features allows efficient automated matching of 2D gel images. Varying degree of fuzziness and alternating between column and rows standardization of correspondence matrix, it is possible to establish one-to-one correspondence between gels features and properly calculate transform parameters. The only limitation of the proposed approach is associated with its global nature.

ACKNOWLEDGMENT

K. Kaczmarek thanks Unilever R&D Vlaardingen (The Netherlands) for financial support of his Ph.D. study.

REFERENCES AND NOTES

(1) O'Farrell, P. O. High-resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **1975**, *250*, 4007−4021.

(2) Westermeier, R. *Electrophoresis in practice: a guide to theory and practice;* VCH: Weisenheim, 1993.

(3) Celis, J. E.; Gromov, P. 2D protein electrophoresis: can it be perfected? *Curr. Opin. Biotechnol.* **1999**, *10*, 16−21.

(4) Harry, J. L.; Wilkins, M. R.; Herbert, B. R.; Packer, N. H.; Gooley, A. A.; Williams, K. L. Proteomics: Capacity versus utility. *Electrophoresis* **2000**, *21*, 1071−1081.

(5) Quadroni, M.; James, P. Proteomics and Automation. *Electrophoresis* **1999**, *20*, 664−677.

(6) Olson, A. D.; Miller, M. J. Elsie 4: Quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* **1988**, *169*, 49−70.

(7) Pleissner, K. P. The CAROL system available over the Internet, http://gelmatching.inf.fu-berlin.de/.

(8) Lemkin, P. F.; Lipkin, L. E.; Lester, E. P. Some extensions to the GELLAB 2D electrophoresis gel analysis system. *Clin. Chem.* **1982**, *28*, 840−849.

(9) Lemkin, P. F. Comparing two-dimensional electrophoretic gel images across the Internet. *Electrophoresis* **1997**, *18*, 461−70.

(10) Lemkin, P. F.; Thornwall, G. Flicker image comparison of 2-D gel images for putative protein identyfication using the 2DWG Meta-Database. *Molecular Biotechnol.* **1999**, *12*, 159−172.

(11) Monardo, P. J.; Boutell, T.; Garrels, J. I.; Latter, G. I. A distributed system for two-dimensional gel analysis. *Comput. Applications Biosci.* **1994**, *10*, 137−143.

(12) Heipke, C. Overview of Image Matching Techniques. *Proceedings of the OEEPE Workshop on "Application of digital photogrammetric workstations"*; available from http://dgrwww.epfl.ch/PHOT/workshop/wks96/art_3_1.html.

(13) Dryden, I. General shape and registration analysis. In *Stochastic Geometry: Likelihood and Computation*; Barndorff-Nielsen, O. E., Kendall, W. S., van Lieshout, M. N. M., Eds.; Monographs on Statistics and Applied Probability, Chapman and Hall/CRC Press: Boca Raton, 1999; Vol. 80, pp 333−364.

(14) Mardia, K. V.; A review of image-warping methods. *J. Appl. Statistics* **1998**, *25*, 155−171.

(15) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: part B*; Elsevier: Amsterdam, 1998.

(16) Smilansky, Z. Automatic registration for images of two-dimensional protein gels. *Electrophoresis* **2001**, *22*, 1616−1626.

(17) Veeser, S.; Dunn, M. J.; Yang, G. Z. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics* **2001**, *1*, 856−870.

(18) Huttenlocher, D. P.; Klanderman, G. A.; Rucklidge, W. J. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Analysis Machine Intelligence* **1993**, *15*, 850−863.

(19) Rangarajan, A.; Chiu, H.; Bookstein, F. L. The softassign Procrustes matching algorithm. Proceedings of the 15th International Conference Information Processing in Medical Imaging. In *Lecture Notes in Computer Science*; Springer: Heidelberg, 1997; Vol. 1230, pp 29−42.

(20) Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals Mathematical Statistics* **1996**, *35*, 876−879.

(21) Loaded from WORLD-2DPAGE - 2-D PAGE databases and services from http://www.expasy.ch/ch2d/2d-index.html; now available over the Internet http://www.cto.us.edu.pl/∼kkaczm6/obrazki.

CI020266K