MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1282

A Block-matching based technique for the analysis of 2D gel images

Ana Freire[a,d], José A. Seoane[a,d], Álvaro Rodríguez[a,d], Cristina Ruiz-Romero[b,d], Guillermo López-Campos[c], Julián Dorado[a,d]

[a] *Information and Communications Technologies Department, Faculty of Computer Science, University of A Coruña, Spain*
[b] *Complejo Hospitalario Universitario de A Coruña (CHUAC), Spain*
[c] *Bioinformatic and Public Health Unit, Institute of Health Carlos III. Spain*
[d] *Biomedical Research Institute of A Coruña (INIBIC), Spain*

Abstract

Research at protein level is a useful practice in personalized medicine. More specifically, 2D gel images obtained after electrophoresis process can lead to an accurate diagnosis. Several computational approaches try to help the clinicians to establish the correspondence between pairs of proteins of multiple 2D gel images. Most of them perform the alignment of a patient image referred to a reference image. In this work, an approach based on block-matching techniques is developed. Its main characteristic is that it does not need to perform the whole alignment between two images considering each protein separately. A comparison with other published methods is presented. It can be concluded that this method works over broad range of proteomic images, although they have a high level of difficulty.

Keywords:

2-D gel electrophoresis, Block-matching, Proteomics, Computer-assisted image processing

Introduction

The images of 2D gels [1] resulting from electrophoresis are a powerful biomedical diagnosis mechanism. This process is based on the separation and analysis of proteins extracted from tissues, blood, cell, etc. This type of image is obtained by staining the proteins that have been separated in a polyacrylamide gel after applying electrical potential difference to it. The separation follows a bidimensional pattern according to their molecular weight and isoelectrical point.

When analyzing images from 2D gels, there is a reference image that represents the distribution of a sample of proteins in reference conditions (normal or healthy status). In such case, molecules are labelled and their spatial location is known. Test images are then presented. In this case, the spatial location of the proteins is unknown. Usually, the comparison between a test image and the reference image is performed in order to establish the correspondence between proteins. Subsequently, both images are compared in order to establish a diagnosis based on the differences in the pattern of the identified proteins.

Images, such as 2D gels, are increasingly used in the biomedical field. The analysis of such images is difficult due to the variability of the different electrophoresis processes. Consequently, images with an apparently complex correspondence may be obtained. Location, shape, size and intensity of a given protein may vary from one image to another, it may not even appear, so the correspondence can be difficult or it might not be established. Due to this reasons, computational techniques are essential for image analysis.

There are several current software packages which try to solve this problem (i.e.: Nonlinear Dynamics Samespots [2], Decodon Delta2D [3] and Genebio Melanie[4]). These approaches tend to perform the whole alignment between two images. This is done by using several image transformations. The user must correct manually possible misalignments in order to obtain reliable results in later stages of the analysis. The system marks each protein with the same edge over all the gels of the same experiment. In this way, the user can select the spots of interest and compare them over the different images and obtain several conclusions based on their differences.

Most of those software packages do not indicate the identifiers or the names of the proteins in the test images, so the user must identify visually the proteins from a labelled reference image.

Apart from the software packages, there are different image analysis methods to align 2D gel images. They can be divided in two groups: those which use landmarks and those which use intensity information.

Two key concepts should initially be defined. A landmark refers to a characteristic of the image from which information is intended to be obtained. In proteomic images, landmarks are identified as proteins (dark spots on a clear background) with an unknown spatial location which will then be obtained. Intensity relates to the function that describes the grey level values corresponding to each pixel in the image.

The approach based on landmarks is a research line applied by several previous works. In [5], the detection of coordinates for the protein centres is firstly performed using an approach based on gradient and Watershed transformation (which performs the image segmentation according to its grey levels). Typical protein characteristics are also considered for molecule detection, like circular or elliptic rim. A local matching process between the reference image and the test image based on the Delaunay triangulation is then performed. This technique consists of a network of triangles where the circumferences circumscribed to each triangle do not contain any vertex of other triangle. A system based on matching the spot centres is presented in [6]. In this case, the rim detection has been carried out, as in [5], by considering different characteristics of the proteins in order to separate protein overlapping. The matching based on the coordinates of the protein centres is also carried out in [7] by following the approach known as fuzzy matching. This approach calculates the nearest protein to every reference protein within a certain range.

Other approaches are focused on the image intensity distribution in order to perform the matching between two gels. With the present approach, the cost of the landmarks extraction process is avoided. The principal objective of [8] is to find the alignment using regional matching (rectangles containing several molecules) instead of spots matching. In [9] the matching is performed by calculating the crossed correlation between the intensity distributions of the test image and the reference image. The strategy followed for image registration is an iterative solution based on the gradual selection of images' resolution, obtaining at each step a more precise transformation. There are two more recent works [10, 11], which are based on Robust Automated Image Normalization (RAIN).

There are new techniques that merge the two previous groups, that is, hybrid approaches. Particular cases of the Iterative Closest Point algorithm [12] (used for solving the alignment between the spots of a test image and the reference image using the Euclidean distance) were presented in [13] and [14]. Both works propose new distance metrics that combine the Euclidean distance with information related to the shape and the intensity of the spots. The aim of [15] is to find a function that makes image alignment possible using a nonlinear deformation model (B-splines). The optimization is based on Levenberg-Marquardt (LM) [16]. This method performs an iterative parametric fitting taking into account a predetermined mathematical model. In [17], some pairs of corresponding spots of both images are selected. The information about the correspondence of the landmarks is introduced as part of the energy function, which is minimized to perform the transformation. The centres are detected by modelling the proteins as 2D inverted Gaussian functions using LM fitting (as in the previous case). A new version of this work [18] was presented in 2008. This work uses the Navier equation, which represents the regularization of the deformation field. It is used with the aim of considering the crossed effects of some gel deformations.

As a consequence of this analysis, the main objective of this work was to find a new method which allows the clinicians to identify automatically the spots, where the alignment of the images is not necessary.

Methods

Establishing a correspondence between molecules of two different gel images might be a difficult task. As has been previously mentioned, this difficulty is due to potential displacements or appearance changes between proteins in two different images.

Several techniques have been studied with the aim of finding a solution for this type of scenario. These techniques, used for movement estimation, consist of mathematical procedures that analyse the intensity changes of a sequence of images

Block-Matching

Among the optical flow estimation techniques, regional fitting (Block-Matching) has been chosen because it is suitable for measuring displacements due to nonlinear movements with a high degree of deformation. Due to their characteristics, these techniques are especially suitable for fluid analysis. A good performance in this context is expected as the proteomic images are obtained from viscous fluids (polyacrylamide gels).

The Block-Matching approach considers a reduced space of the correspondence problem in order to achieve a better approach rather than the global case. These techniques usually work as follows [19]: the image is divided into some regions; different criteria might be followed to perform this division. The simplest approach implies selecting non-overlapping regions of predetermined size, known as blocks. The aim of the next step is to calculate the displacement for each region between two images. This is done assuming that the local distortions caused by the displacement are almost negligible. Thus, considering a region small enough and a time lapse short enough, the characteristics of each region will not be affected by the movement. This is the only assumption made in relation to the movement which will be calculated. Every block of an image is then compared with several possible blocks of the following one, maximizing a similarity measure or minimizing a distance measure.

Modified Block-Matching

As the previous process is general, it does not fit closely to the 2D gel scenario. This is due to the assumption made by the generic Block-matching algorithms: despite the displacement, within an area small enough, the visual texture of a region remains unaffected. In the field of 2D gels this hypothesis is not fulfilled because the images are not part of a temporal sequence affected by movement. In those sequences there are few changes from an image to the next one. In proteomics, the images are independent. Besides, the variability is increased due to the previously mentioned changes experienced by the protein samples in different images. Due to this, the use of certain similarity measures presents some restrictions, because

the difference between a region and its corresponding region in the other image is very high in 2D gels. Although the result of the exploration might be correct, it might not meet the minimum difference criterion. Because of that, the system will assume that the displacement produced is not detected. A similarity measure based on the statistical distribution of the intensity levels is then required. This measure needs to be robust despite those factors that vary among samples.

Due to the reasons explained previously, a new Block-Matching-based method has been developed. This method proposes different specific strategies to adapt Block-Matching to proteomic images of 2D gels. The process, show in figure 1, is carried out as follows:

- Firstly, the method needs as input the list of coordinates of the spots in the reference image. This part can be done in different ways: getting the output of a spot picking robot, such as GelPix [20], or using several image analysis techniques. Present work does not focus on this phase, but on the modified block matching technique explained as follows.
- The system refines the coordinates in order to match exactly spots' centres. They are also the centres of the blocks into which the image is divided (the whole image is not divided into blocks, unlike in Block-Matching algorithms; only some blocks are marked: as many as the spots extracted). To perform this task LM fitting has been used. Therefore, it is necessary to define an area that might contain, with high probability, the centre of the protein. The coordinates of the centre of that area are also extracted and used as the initial estimation in LM fitting. That area is determined using wave search, which is carried out as follows: a minimum radius is established. The initial coordinates represent its centre. This radius is increased (to a maximum) until it reaches the rim (an intensity value up to a certain grey threshold) and contains the protein centre (an intensity value less than a certain grey threshold). The minimum intensity point in the delimited area is used as the provisional centre for each iteration.
- Once this area has been established, the centre of the protein is located using LM fitting. It was necessary to find a fitting function for the proteins' grey level. Then, an inverted bidimensional Gaussian function was chosen as it is a continuous function which is adequate to describe the distribution of the molecule intensity. This model determines the protein centre (the darkest value and therefore the lowest grey level) as the minimum in the inverted Gaussian curve. The fitting function undergoes a rotation geometrical transformation because the molecule can turn towards different directions.
- The protein centre adjusted in the previous phase is the central position of the search block for the modified Block-Matching algorithm. The size of the block is defined by specifying its dimensions using the Block Size parameter (in pixels). Using each block as a centre, a search region is defined. This region is demarcated by a maximum range of displacement (Search region).
- In the search region, a spiral exploration strategy has been followed. Thus, starting at the block coordinates of the first image, hops determined by the Search hop parameter are performed following a spiral. Each hop represents one pixel, in such a way that the whole space of states will be explored obtaining always the best possible value for the comparison criterion, the Pearson correlation coefficient which is a typical statistical coefficient used to calculate the linear relationship between two quantitative variables. It is immune to changes in the medium grey level and it is independent of the value scale used. It has also proved to be robust when there is noise.
- Once the correspondence with a block in the test image has been established, the position of the centre of this destiny block will be refined in order to match the protein centre, as was done at the beginning for the reference image. The output is showed in the Figure 2.

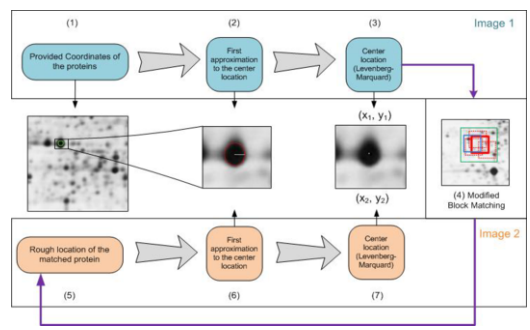

Figure 1 –Modified block-matching diagram

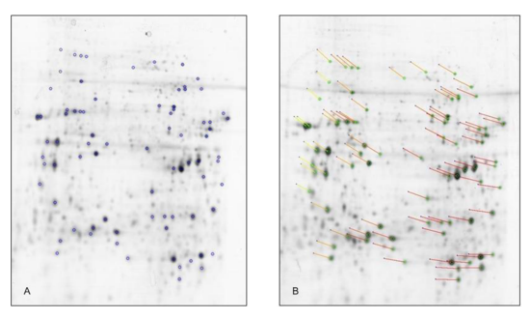

Figure 2 – System output. A) Marked molecules. B) market molecules in the test image, with a vector that indicates not only direction, but also the magnitude of the displacement in relation to the corresponding protein in the reference image

Results

The proposed system has been implemented using Open Source Computer Vision Library (OpenCV) [21] and has been tested in order to compare our results to those of other articles.

The Table 1 show the results, represented as success rates, of the comparison between the proteins in the reference and the test image. In order to understand better the results, it is necessary to define the following terms:

- It is considered a success when the centre calculated by the algorithm is included within the perimeter of a protein in the test image. In addition, this protein must match, according to the expert's criterion, its corresponding protein of the reference image, based on its spatial location and characteristics.
- It is considered failure when the calculated centre is out of the body of the protein which matches the reference one (according to the expert's criterion).

Test

In order to check the accuracy of the method presented in this paper, the results obtained were compared to those obtained in [18]. This work shows the success rates of a proposed hybrid method and those obtained with an approach based only on intensity. In [18], their results are compared to the results obtained in a previous work [17], which includes a hybrid approach and an intensity-based one.

The test bed was built by using the images of G.-Z. Yang (Royal Society/Wolfson MIC Laboratory, Department of Computing, Imperial College of Science, Technology and Medicine, London). This test bed consists of a group of images of 2D gels from different types of tissues and different experimental conditions.

The test bed used in the present work was built grouping these images in pairs in order to perform their alignment. Every couple was assigned a complexity level according to the criterion of an expert. As a result, the following groups were obtained:

- Five pairs of images where the visual correlation of the proteins between every reference image and test image is simple.
- Five pairs of images where most of the proteins can still be visually correlated, but, in this case, it is harder to establish the correlation.
- Five pairs of images of high complexity, where most of the proteins cannot be visually correlated

The comparison function, block size, search region and search method parameters were chosen according to previous tests on synthetic images. The values chosen for the parameters were Pearson Correlation as comparison function, wave search as search method, block size of 75x75 in easy and medium complexity images and 150x150 in complex images and search region of 75x75 in easy and medium complexity images and 125x125 in complex images.

We have selected [17] and [18] because their tests use images extracted from the Wolfson MIC Laboratory test bed. However, in both, only one pair of images was selected for each type of complexity (no results of high complexity images were published). In these works, 208 proteins were selected in order to test the results in low complexity images, and for the medium complexity pair 158 proteins were selected. After that, the number of proteins correctly and incorrectly identified was obtained. The success rate was then calculated. In order to fairly compare the results obtained in [17] and [18] to the results obtained in our work, the same pairs of images were used. As they do not show which proteins have been selected, it was carried out as follows: for the low complexity pair, 212 identifiable molecules were found. From these, 208 were selected. As a result, at the most, 2.4% of the molecules may vary. From the medium complexity pair, 158 proteins were selected from the 160 identifiable ones. In this case, at the most, 1.27% the proteins may vary. As these rates are minimal, the results can be considered comparable. The pair of high complexity images chosen in [17] was also selected, despite none of the papers having mentioned any related results. We have chosen 55 identifiable proteins for this pair. This number is not very high because most of the proteins have a high overlapping rate that prevents establishing a reliable correspondence.

Table 1 – Comparative with other methods

	Easy			Medium			Complex		
	n_{cor}	n_{inc}	%	n_{cor}	n_{inc}	%	n_{cor}	n_{inc}	%
[17] Intensity	187	21	89.9%	137	21	86.7%	-	-	-
[17] Hybrid	201	7	96.6%	149	9	94.3%	-	-	-
[18] Intensity	200	8	96.2%	150	8	94.9%	-	-	-
[18] Hybrid	203	5	97.6%	153	5	96.8%	-	-	-
Modified BM	207	1	99.5%	154	4	97.4%	47	8	85.4%

Test results

The number of correctly (n_{cor}) and incorrectly (n_{inc}) identified proteins are presented in Table 1 with the corresponding success rates.

The results obtained in the present work are slightly better than those obtained in the approaches published in [17] and [18]. Even more, these two works do not present results for high complexity images, as they do not consider them suitable due to their complexity. There are several differences between these methods and the one presented here: in [17] and [18], a manual selection process of certain proteins from the reference image and the corresponding ones in the test image is performed. This user-supplied matching information is used to refine the registration in regions where the information about intensity is not enough. Then, the alignment is performed on the whole gel. However, in this work, the process is fully automatic and the system performs the association between the proteins of interest, it does not carry out the whole gel alignment.

Conclusion

The aim of the present work has been to obtain an effective method in order to calculate the correspondence between the proteins of 2D gel images obtained using an electrophoresis process. An approach based on regional fitting techniques has been developed. The generic Block-Matching technique was modified in order to apply it to the 2D gel scenario. The success rates obtained after executing the required tests using real biomedical images were higher than the rates reached by previous works using the same test bed. This method performs the identification over each protein separately, and it does not perform the whole alignment. Thus, the identification could be performed over a subset of proteins because many times only some proteins are important for establishing a diagnosis. Since this method works at spot level, it could be easier to transfer protein labels between images, so the manual labelled process would be avoided. In this way, it could be very useful for protein information retrieval systems.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Innovation (Ref TIN2006-13274), grant (Ref. PIO52048 and RD07/0067/0005) funded by the Carlos III Health Institute, grant (Ref. PGDIT 07TMT011CT) and (Ref. PGDIT08SIN010105PR) from the General Directorate of Research, Development and Innovation of the Xunta de Galicia and grant (2007/127 and 2007/144) from the General Directorate of Scientific and Technologic Promotion of the Galician University System of the Xunta de Galicia. The work of José A. Seoane is supported by an Isabel Barreto grant from the General Directorate of Research, Development and Innovation of the Xunta de Galicia.

The original proteomic images used in this work are courtesy of Prof. G.-Z. Yang, Royal Society/Wolfson MIC Laboratory, Department of Computing, Imperial College of Science, Technology, and Medicine, London/UK.

References

[1] Görg A, Obermaier Ch, Boguth G, Harder A, Scheibe B, Wildgruber R, Weiss W: The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000, 21:1037-1053.

[2] Nonlinear Dynamics. http://www.nonlinear.com/products/progenesis

[3] Decodon. http://www.decodon.com/Solutions/Delta2D.html

[4] Genebio. http://www.genebio.com/products/melanie/index.html

[5] Peiβner K, Hoffmann F, Kriegel K, Wenk Carola, Wegner S, Sahlström A, Oswald H, Alt H, Fleck E: New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. Electrophoresis 1999, 20:755-765.

[6] Almansa A, Gerschuni M, Pardo A, Preciozzi J: Processing of 2D Electrophoresis Gels. Iccv 2007 – Workshop on Computer Vision Applications for Developing Countries. 15 October 2007.

[7] Kaczmarek K, Walczak B, De Jong S, Vandeginste BGM: Feature based fuzzy matching of 2D electrophoresis images. J of Chem Inf Comput Sci 2002, 42:1431-1442.

[8] Smilansky Z: Automatic registration for images of two-dimensional proteins gels. Electrophoresis 2001, 22:1616-1626.

[9] Veeser S, Dunn MJ, Yang GZ: Multiresolution image registration for two-dimensional gel electrophoresis. Proteomics 2001, 1:856-870.

[10] Dowsey AW, Dunn MJ, Yang GZ: Automated image alignment for 2D gel electrophoresis in high-throughput proteomics pipeline. Bioinformatics2008, 24(7):950-957

[11] Dowsey AW, English J, Pennington K, Cotter D, Stuehler K, Marcus K, Meyer HE, Dunn MJ, Yang GZ: Examination of 2-DE in the Human Proteome Organisation Brain Proteome Project pilot studies with the new RAIN gel matching technique. Proteomic 2006, 6:5030-5047.

[12] Besl PJ, McKay ND: A Method for Registration of 3D Shapes. IEEE Trans on Pattern Anal and Mach Intell 1992, 14(2):239-256.

[13] Shi G, Jiang T, Zhu W, Liu B, Zhao H: Alignment of two-dimensional electrophoresis gels. Biochem and Biophys Res Commun 2007, 357:427-432.

[14] Rogers M, Graham J: Robust and accurate registration of 2-D electrophoresis gels using point-matching. In IEEE transactions on image processing. Vol 16. Edited by IEEE, New York: 2007, 3:624-635.

[15] Sorzano COS, Thévenaz P, Valdés I, Beloso A, Unser M: Elastic Image Registration with Applications to Proteomics. INFORMATION OPTICS: 5th Int Workshop on Information Optics. AIP Conf. Proceedings. Edited by AIP. Vol 860, 2006: 300-309.

[16] Levenberg K: A Method for the Solution of Certain Non-Linear Problems in Least Squares. The Quarterly of Applied Mathematics 1944, 2: 164–168.

[17] Rohr K, Cathier P, Wörz S: Elastic registration of electrophoresis images using intensity information and point landmarks. Pattern Recognit, 2004, 37: 1035-1048.

[18] Wörz S, Winz ML, Rohr K: Geometric alignment of 2D gel electrophoresis images using physics-based elastic registration. In Proc. IEEE Int Symposium on Biomedical Imaging: From Nano to Macro. Paris, France, May 2008.

[19] Stiller C, Konrad J: Estimating motion in image sequences, a tutorial on modeling and computation of 2D Motion. IEEE Signal Processing Mag 1999, 16(4):70-91.

[20] GelPix. http://www.vcu.edu/csbc/msrsbc/instruments_gelpix.html

[21] SourceForge. http://sourceforge.net/projects/opencvlibrary