A New Approach to Detect Similar Proteins from 2D Gel Electrophoresis Images

Nawaz Khan
School of Computing Science
Middlesex University, UK
n.x.khan@mdx.ac.uk

Dr. Shahedur Rahman
School of Computing Science
Middlesex University, UK
s.rahman@mdx.ac.uk

Abstract

Many algorithms are available for quantitative and qualitative analysis of protein spot in gel electrophoresis images and majority of the these algorithms use geometric and image processing techniques to match protein spots. These algorithms do not take into consideration the electrophoretic mobility of the proteins and they only match similar protein spots rather than matching similar proteins. The approach presented in this paper uses a novel technique based on electrophoretic mobility to match protein spots between source and target images. The algorithm identifies the protein spot in the target image which lies on the same line of path as it is in the source image. A shape matching algorithm using Generalized Hough Transform and Canny Edge Detection method is used to determine the shape variance. The method describes here gives an accuracy of 90% or more to identify the same or similar proteins from the target image. Finally, a dedicated target based database has been created to store a set of finite values of an element spot for correlating the 3D protein structure.

1. Introduction

Gel electrophoresis technique is a fundamental procedure for separating the DNA and proteins from a mixture. This technique is based on the concept of separating the charged particles in gel. Gel electrophoresis analysis is a well-established technique to compare and contrast one protein with another. There are many software available which are used to compare the protein spots. For example, GELLAB system [9,10], uses the point pattern comparison technique to identify the spots of the same protein. Another system is MELANIE which compares spot clusters instead of using simplified point pattern matching technique. The system also defines a probabilistic criterion for the definition of a correct match (Appel *et al.* 1997). These systems work well for the same type of gel electrophoregram where the intensity of the spots does not change. However in both cases if the intensity value of the spots change then it can significantly affect the outcome of the gel image matching.

New algorithms are also emerging for protein spot matching. For example, Panek and Vohradsky 1999, introduced a new algorithm to identify the protein spot. Their approach uses the information from the neighbourhood spots for comparison. A syntactic descriptor characterize the spots and then the positional similarity is derived. Pre-processing phase in their technique also leads to image distortion. Pleibner *et al.* 1999 introduced another algorithm to detect the protein spots using Delaunay Triangulation technique. This method is based on computational geometry and it requires to correct the false results interactively. All these methods are based on geometric computation and image processing techniques and they match similar protein spot rather than matching similar protein. Although the spot in the source and target image can be identical or similar, but still the following parameters can vary:

- background value
- protein spot intensity
- protein spot shape, and
- noise in the image

For example, in figure 1, *triosophosphate isomerase* protein spot lies at the same line of path (*pq*) in both images *(a,b)* but the spot shape, background information and the intensity of the spot vary significantly in this case.

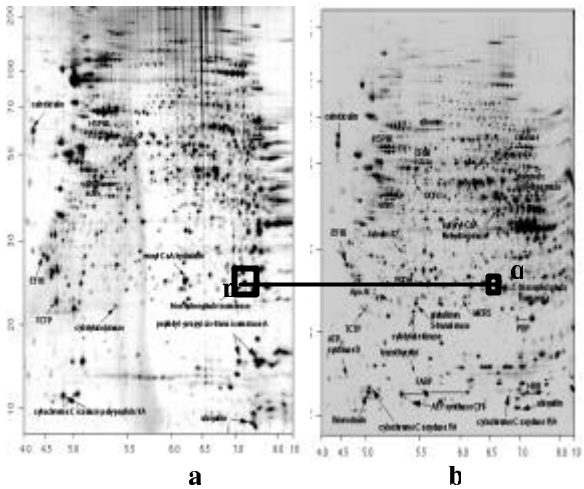


Figure 1. ***Triosophosphate isomerase*** **protein spots in two different images**

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 **IEEE**

This paper presents a novel approach for identifying the identical or similar protein spot in 2D gel electrophoresis images by considering the following factors:

- 2D gel electrophoresis protein spots differ significantly in two different images even when they represent the same protein.
- same or similar protein spots will lie at the same line of path because of their electrophoretic mobility and molecular weight.
- the intensity of the matched regions in both images can be different even though it shows a correct matching.
- the region of similar spot at the target image must lie at the same or different directional vector on the line of path.

2. Methodology

The method described in this paper has used a novel technique to determine the position of similar or identical protein spot in target image which lies on the line of path. For shape matching the method interprets the pixel regions in terms of directional vector, intensity, contour description, and its electrophoretic mobility. It identifies a set of pixels with its centre and draws a polygonal curve around these set of boundary pixels to determine the shape which establishes the region of interest. It then searches for the same spot at the same or least variance position in the target image. The shape of the spot 'ρ' needs to be matched with the shape 'λ' which can be found at the same or least variant position in the target image. To find out the identical or the similar spot in the target image the following criteria needs to be fulfilled:

1. Intersecting the spots in the target image with the same line of path as it is used for the source image spot
2. Determining the spot of interest in the target image by examining the directional vector
3. Determining the shape of the source spot for resemblance matching with the target spot shape.

To achieve an 'exact match' the objective is to identify the spot in the target image which is closely resemblance to the source image spot features.

2.1 Determining the position of the protein spot in the source image

The source image *S* from which a particular spot of protein will be detected is set to a user defined reference point with equal width and height. This assumption is applied to all the source images. The source image is also divided into four quadrants, two upper regions (*a* and *b*) and two lower regions (*c* and *d*) with the horizontal and the vertical axes intersecting at the central point of these quadrants (Fig. 2). Any protein spot '*p*' of the electrophoregram will lie in any one of these quadrants.

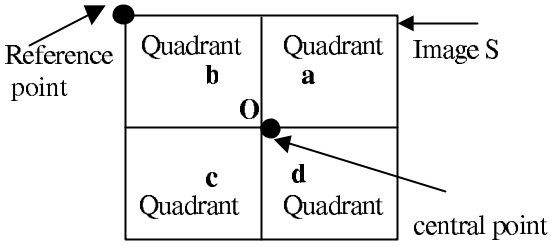


Figure 2. Source image divided into four quadrants

If we consider this '*p*' as any point on the vertical plane, we can define a vector on this plane to determine the orientation of the spot. The vector (δ_s, θ_s) to reach point '*p*' from centre of the image will produce the specific orientation of the spot (Fig. 3), where δ_s and θ_s are the path length and the corresponding angle respectively.

The directional vector, δ_s, from the centre point to point '*p*' is determined as follows:

$$|\delta_s| = \sqrt{(x_o - x_p)^2 + (y_o - y_p)^2} \quad \ldots\ldots\ldots \text{ i}$$

where (x_o, y_o) is the coordinate of the central point *O* and (x_p, y_p) is the coordinate of point *p*. The angle with horizontal axis, θ_s, is determined as follows:

$$\theta_s = \text{arctg}\left(\frac{y_p - y_o}{x_p - x_o}\right). \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{ ...ii.}$$

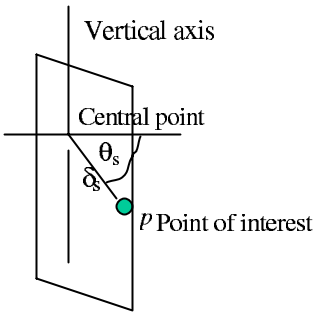


Figure 3. Angle produced with the horizontal axis for any point of interest on the vertical plane.

δ_s and θ_s will determine the orientation of point '*p*'. A region of interest is drawn around the point to determine its mean pixel value (M_s) (Fig. 4).

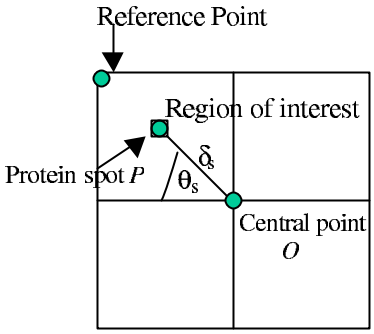


Figure 4. Region of interest of point *p* in the source image.

2.2. Defining the Region of Interest

The region of interest is defined by a set of boundary contour points *Xs* which are selected interactively by the user where the contour points, *M*, is defined as $M = \{X_1, X_2, ..., X_n\}$. Each contour point of *M* is defined with reference to the centre point Y_0. For each of these points, a vector *r* is defined. The largest *r* is taken as the radius and a circle with radius *r* is drawn which covers the whole spot. A rectangle of *2r* width and height is then drawn as the region of interest (Fig. 5).

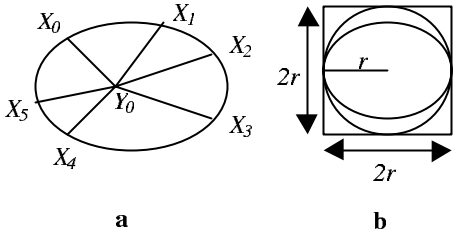


Figure 5. (a) A set of points defined by the user, (b) Defining the region of interest.

2.3. Matching the selected protein spot in the target image

The target image '*T*' will be loaded from local or public domain databases at user defined reference point. To locate the protein spot on the target gel electrophoregram at the same position and orientation, image *T* is also divided into four quadrants. The path length, δ_T and the angle with the horizontal axis, θ_T are determined for the target image using the previous equations (i and ii). Region of interest is drawn at the same position and orientation and the mean value, M_T is calculated. If the mean value M_T matches with the mean value M_s at the same position and orientation, then the point will be considered as matched point. The target point will be considered as 'matched point' when:

i. $\theta_s = \theta_T$;ii. $\delta_s = \delta_T$ and iii. $M_s = M_T$

In other situation where

i. $\theta_s = \theta_T$;ii. $\delta_s = \delta_T$ and iii. $M_T < M_s$ or $M_T > M_s$ (where M_T is within a threshold value), it will be considered as 'spot found' and it might be the same or similar protein. The mean value of the region of interest may vary due to the variation in the protein concentration or variation in the image brightness/contrast. If no spot is detected on the line of path it will then continue to search for the protein spot in the neighbourhood area.

2.4. Searching for the protein spot in the neighbourhood area

The line of path through the region of interest is chosen for the neighbourhood spot matching operation assuming that all identical or similar protein spots will lie on this path. The assumption is based on the fact that the electrophoretic mobility and molecular weight of the same or similar protein do not vary. The line of path drawn at θ_T angle with δ_T length is parallel to the horizontal axis and it goes through the region of interest on image *T*. This line will intersect any similar point along its path. The line will be a 'non emptied' line of path if it intersects at least at one spot of the protein along its path (Fig. 6).

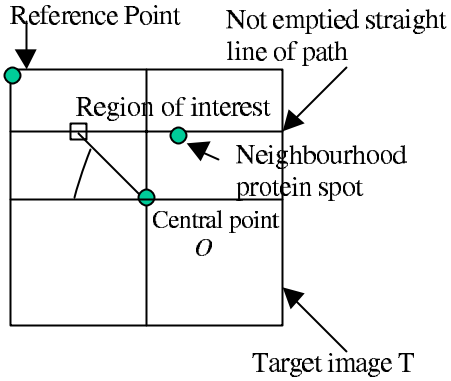


Figure 6. Non emptied straight line of path in the target image to determine the neighbourhood protein spot

The directions of search for the neighbourhood spot are defined as follows:

i. If region of interest lies at quadrant 'b' then it will look in the right direction along the path (towards quadrant 'a')
ii. If region of interest lies at quadrant 'a' then it will look in the left direction along the path (towards quadrant 'b')
iii. If region of interest lies at quadrant 'c' then it will look in the right direction along the path (towards quadrant 'd')
iv. If region of interest lies at quadrant 'd' then it will look in the left direction along the path (towards quadrant 'c')

Figure 7 illustrates the directions of search for the neighbourhood spot.

2.5 Selecting the best matched spot

i. Variance analysis: The line of path is drawn to isolate all those spots from target image which lies on the line of path. A vector for all these spots are stored. The vector $\vec{v}$ consists of pixel value MT_n, length δ_T, angle θ_T and the coordinates of the spots. $\vec{v}$ can be expressed as

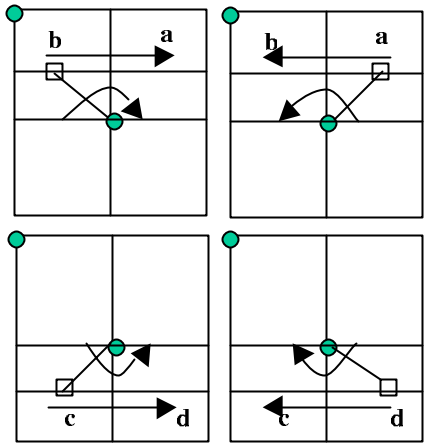


Figure 7. Directions of search for the neighbourhood spot.

$$\vec{v} = \begin{bmatrix} MT1 & \delta T1 & \theta T1 & x1 & y1 \\ MT2 & \delta T2 & \theta T2 & x2 & y2 \\ . & . & . & . & . \\ MTn & \delta T3 & \theta T3 & xn & yn \end{bmatrix} \quad \ldots\ldots\ldots \text{ iii}$$

To identify the best spot on the line, the variances of the spot position and the orientation are calculated with respect to the protein spot p in the source image. Let us assume that the length, angle and the mean variance are δ_v, θ_v and M_v respectively, then

$$|\delta_v| = \delta_T - \delta_s$$
$$|\theta_v| = \theta_T - \theta_s$$
$$|M_v| = M_T - M_S$$

The cumulative variance Δ is calculated as follows:

$$\Delta_i = (\delta_v + \theta_v + M_v)_i \quad \ldots\ldots\ldots\ldots \text{ iv.}$$

where $i = 1$ to n and n is the number of spots.

The variances for each spot in the target image that lies on the line of path are calculated. The spot with minimum variances (Δ_{min}) is considered as the 'best possible matched' with that of the source image.

The spot with the least variance (Δ_{min}) is considered to be the 'best matched spot'.

ii. Shape contour matching: Shape contour in the source image is approximated by the sample points M selected interactively (see section 2.2) by the user. Every point of M can be described with respect to some reference point y inside the contour through a vector $r_i = y - x_i$ (Fig. 5, section 2.2). All the vector r and the points M are stored in a M-R table. An edge image I' is drawn using the values from the M-R table and the edge image, I', is mapped to the target region of interest image. This is used to determine the shape of the target spot. The edge of the shape, using *Canny* edge detector (Canny, 1986), is used to determine the boundary of the target spot. The pixel positions and r' are determined for the target protein spot with reference to the region of interest (ROI) centre point. These values are stored in M'-R' table. The shape variation of the spots on the same line of path in the source and target image is compared by using the M-R and M'-R' table. The shape matching results are considered 'similar' if edge ρ of source image consisting X_i to X_{i+n} and edge λ of target image consisting X'_i to X'_{i+n} are within an error tolerant limit.

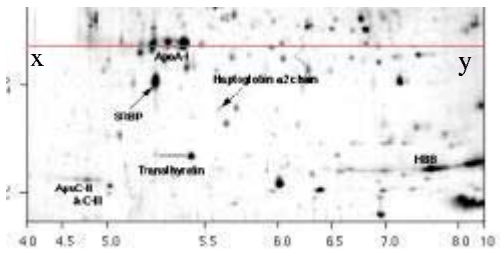
2.6. Retrieving 3D structure of a protein

A dedicated, task-specific and declarative database (Raghavan and Garcia-Molina, 2001, Khan and Rahman, 2002) is created to store the gel spot information. The set of spots in each get electrophoresis is labelled and each spot is associated with a finite number of values. A class is formed using these labels and values. Each entry in the class is of the form (L,V) where L is a label and $V=\{v_1,\ldots,v_n\}$ is a set of values. Each v_i represents a value that could potentially be assigned to an element E, if label (E) matches L. In our case element E is a spot which corresponds to the specific 3-D structure of a protein. An image feature extractor, M_f, is used to look for the value v_i to search the corresponding element E.

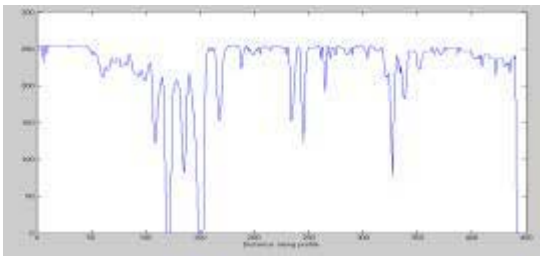
3. Experiments and results

3.1. Identifying a spot along the line of path

Figure 8a is an image of gel electrophoresis. A line of path (x,y) is drawn on the image. To identify the spots on the line of path, the pixel values which intersect the line of path are determined. Nearest neighbourhood interpolation technique is used to find the intensity value for each point along the line of path. It has been established (see the plot (Fig 8b)) that the protein spots which lies on the line of path show a low intensity value. Hence, it can be assumed that the lower intensity values of the pixels can be chosen as the protein spots on the line of path. An appropriate threshold value for low intensity is used to determine these spot on the line of path.



a



b

Figure 8: Identifying a spot along the line of path using the low intensity values

3.2. Identifying a spot of interest in the target image

A set of ten images (400x500) has been used for this experiment. Number of images have been created synthetically with the same background, intensity values and spot size to test the concept. Also a limited number of spots have been created in these target images where a selected number of spots lie on the same line of path as it is in the source image (Fig. 9). The objective here is to see if it retrieves the correct target image when it matches the selected spot from the source image to the corresponding spot in the target image. The spot positions are varied in the synthetic images. The orientation of the spots (angles θ_T) and the path distance (δ_T) of the spots are known (Table 1).

When a spot is selected interactively in the source image it successfully retrieves the corresponding images when the orientation of the source and target spot matches. The retrieval accuracy is 100% in this test.

The objective for the next phase of the test is to look for similar spots in the neighbourhood assuming that an exact match has not been found in the first case. In this test, images have been selected in such a way so that the exact spot corresponding to the source spot does not exist on the line of path. It retrieves all target images which contain the neighbourhood spot that are considered as similar protein spots with regard to the source image protein spot. Four nearest spots at different orientations for each image are used. When the test retrieves the target image, a region of interest is drawn on the target image. The orientation of the spots (the angle and distance path) are recorded and they are compared against the pre-determined known set of values. The result shows an accuracy of 1:8 match which indicates that it only picks up the best neighbourhood spot out of the eight possibilities (Fig. 10).

3.3. Matching on 2D gel electrophoresis image

This part of the experiment is carried out on real gel electrophoregram images. The gel electrophoregrams are not pre-processed for unique background or to eliminate light spots which have higher grey scale values. A source image is chosen and a protein spot is selected which is to be matched with the same or similar protein spot in the target image. The line of path and the region of interest are drawn. The target image is then loaded and the matching operation is carried out to identify the same or similar spot in the target image (figure 11).

Table 1. Vectors of each spot on experimental image to determine the least variance

Spot	Angle	Length To spot	Mean grey value	Total value	Vari-ance	Least Varia-nce	Com-ment
Source spot	58.76	71.3	0	130.0			
1	28.36	128.4	0	156.7	26.6		
2	31.63	116.28	0	147.9	17.84		
3	39.49	95.9	0	135.4	5.34		
4	45.95	84.8	0	130.8	0.74		
5	58.76	71.3	0	130.0	0	0	Spot Of interest

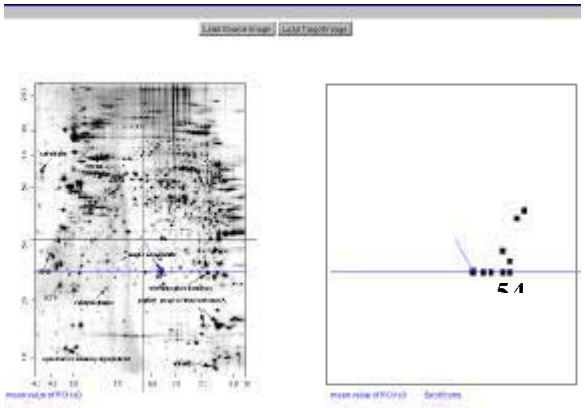


Figure 9. Identifying the spot at the same orientation as it is in the source image

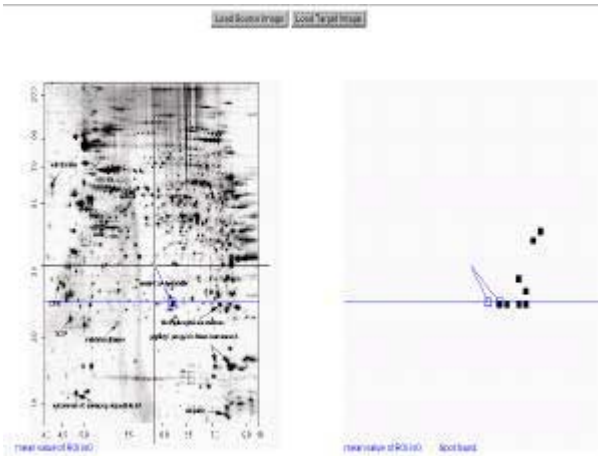
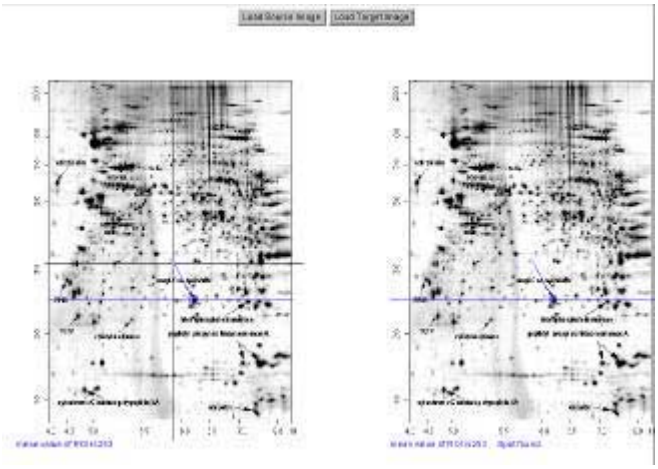


Figure 10. Identifying the neighbourhood spots at the least variance position



a b

Figure 11. Matching spot in the target image at the same location as it is in the source image. [a] Source image and [b] spot found in the target image.

The experiment used ten gel electrophoregram images and it successfully identified spot at the same orientation in different target images. In this case, the best nearest similarity value is used to choose the best matching. For similarity search, the variance value is used. This is carried out by determining the vector $\vec{v}$ and the variances Δ_i as described in equation (iv). The spot with least variance is chosen as the same or similar protein spot on the line of the path. Overall performance shows about 90% successful identification of spots at least variance position in the target image (figure 12).

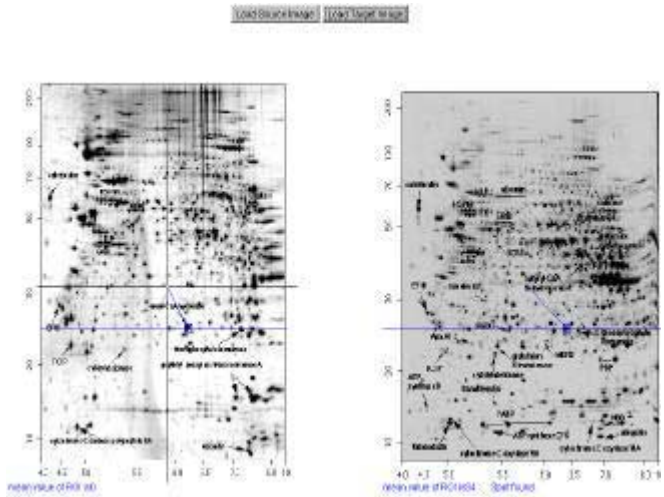
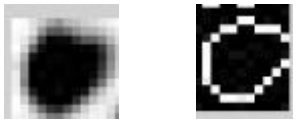


Figure 12. Identifying the neighbourhood spot the in target image.

3.4. Shape comparison

In this experiment a single spot is chosen and the variation between source and target spot are calculated to determine the similarity of the spots. The difference matrix is used to calculate the variance. For example, the following figures (fig 13) show the variation result of 94 %. Shape variance within an offset is taken as similar shape. Table 2 illustrates the parameters used for the comparison.



a b

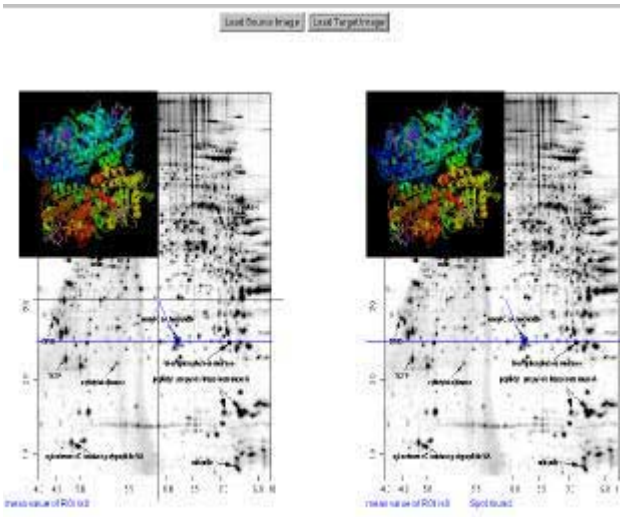
Fig 13: Source spot (a) and detected spot (b) in the target image for shape comparison.

Table 2. Determining the shape variance

Spot	x,y coordinates taken for comparison	r values for the coordinates	Average r	Shape variance %
Source spot	(5,2)	5.385165	4.68772	94.04564
	(4,3)	5		
	(3,6)	4.123106		
	(4,4)	4.242641		
Target spot	(5,3)	4.472136	4.40860	
	(4,3)	5		
	(3,4)	5		
	(4,6)	3.162278		

3.5. Retrieving 3D image

In this experiment, gel electrophoresis image spots are labelled using Melanine software. The following parameters are stored in the target specific dedicated database: all the coordinates, intensity values, positional orientation and the average shape radius of the spots. When a specific spot is selected in the source image, the parameters of that particular spot are then matched within the database for 3d image which corresponds to the selected spot. Figure 14 shows the 3D images that are retrieved from the local dedicated databases for same or similar protein spots.



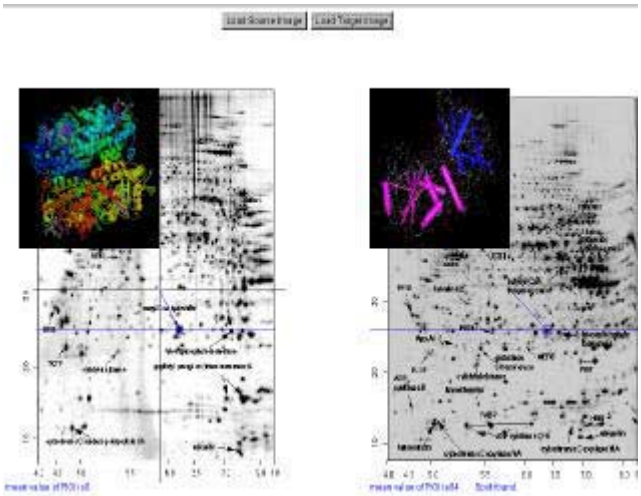


Figure 14. 3D Protein structures retrieved from the dedicated database

4. Discussion and conclusion

This paper has presented an approach for identifying identical or similar spot from target gel electrophoresis image which lies at the same line of path as it is in the source image. A combination of geometric and image processing techniques have been used to identify the spot which matches with the features of the source image spot. Although the technique shows significant accuracy in identifying the identical or the similar protein spot, some false matching were also resulted. The following factors affect the final outcome:

i. image background
ii. the size of the spots, and
iii. the intensity of the spots

Performance can be increased significantly by adjusting the background contrast. Histogram equalisation can be useful tool for creating uniform image contrast. However, a nonlinear transformation can change the shape of the actual image. The contrast of the image background can also be adjusted by estimating the background values using morphological opening. Morphological opening has the effect of removing spot objects (circular shape) with a given predefined radius. But subsequently it also modifies the shape of the objects which therefore leads to image distortion. Binary thresholding and labelling can be used to pre-process the image for improved performance, however it can have a detrimental effect on the actual gel spots. Streaks, twin spots and complex regions can also have a significant effect on the identifying process. The performance can be increased in several fold by using the Brute Force method or LP approach (Alon Efrat *et al*., 2001) which can measure the ellipse encircling the spot. Also the technique described by Kriegel *et al*, 2000 to partition the streaks and the complex regions can have a significant effect when matching with the complex region. Ehrenmann *et al*., 2000 used Generalized Hough Tranform method (GHT) for object shape determination.

GHT in this research is used for determining the contour of the shape and it has then been extended for shape comparison. A dynamic buffer of radius is used for creating the shape in the source image and the target spot boundary coverings are then checked with these values. The novelty of this approach is that the image has not been preprocessed to avoid the image distortion. The spot identifying process described here is dynamic and the image spot is selected interactively by the user (image object keying). The approach has emphasized on extracting the directional vector information from the image. Intensity value of the detected spot shape from the target image is also used as an independent parameter. This helps to come to a conclusion about the spot similarity which enables to identify the protein similarity. The approach is also unique because it searches for the spot only on the line of path and it does not search the whole image. It uses the key concept of electrophoretic mobility of proteins. This approach thus reduces the number of candidate spot to be identified within the image. The research is aiming to apply this approach in the future to initiate the matching process from the local source which will dynamically link the public domain molecular biology databases to extract the matched results or the 3D protein structure. The 3D structure of the protein can then be compared with local protein database for protein variance analysis.

5. References

[1] Alon Efrat, Frank Hoffmann, Klaus Kriegel, Christof Schultz and Carola Wenk, "Geometric Algorithm for the Analysis of 2D Electrophoresis Gels", 2001, Proc. 5th Int. Conference on Computational Molecular Biology, *RECOMB*, 2001, Celera Genomics, pp.114-123

[2] Appel, R.D., Vargas, J.R., Palagi, P.M., Walther, D., Hochstrasser, D.F., "Melanie II, a third-generation software pckage for analysis of two-dimensional electrophoresis images:II Algorithms", *Electrophoresis* 1997, 18,pp. 2735-2748.

[3] Canny, J.F., "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol: 8, 1986, pp.679-698.

[4] Ehrenmann, M., Ambela, D, Steinhaus, P. and Dillmann, R, "A comparison of Four Fast Vision Based Object Recognition Methods for programming by demonstration Applications", *International Conference on Robotics and Automation (ICRA),* April 24-28, 2000, USA, pp.1862-1867.

[5] Josef Panek and Jiril Vohradsky, 1999, Point Pattern matching in the analysis of two dimensional gel electropherograms. *Electrophoresis*, 20, pp.3483-3491.

[6] Khan, N and Rahman, S. "Object Modelling of Gene Mutation Data for Variance Analysis". 2002; *6th World Conference on Systemics, Cybernatics and Informatics (SCI, 2002), SCI in*

Medical and Biology session; Proceedings International Institute of Information Systems (IIIS), USA, Florida, 2002, pp.301-306.

[7]. Klaus-Peter Pleibner, Frank Hoffmann, Klaus Kriegel, Carola Wenk, Susan Wegner, Anders Sahistrom, Helmut Oswald, Helmul Alt and Eckart Fleck, 1999, "New Algorithmic approaches to protein spot detection and pattern matching in two dimensional electrophoresis gel database", *Electrophoresis*, 20, pp.755-765.

[8]. Kriegel, K, Seefeldt, L, Hoffmann, F., Schultz, C., Wenk, C., Regitz-Zagrosek, V. and Fleck, E. "An alternative approach to deal with geometric uncertainties in computer analysis of two-dimensional electrophoresis gels". *Electrophoresis*, 2000, 21, pp.2637-2640.

[9]. Lemkin, P., Lipkin, L., "GELLAB: A computer system for 2D gel electrophoresis analysis. I. Segmentation and preliminaries". *Computers and Biomedical Research* 14, 1981, pp.272-297.

[10]. Lemkin, P., Lipkin, L., "GELLAB: A computer system for 2D gel electrophoresis analysis. II. Spot pairing", *Computers and Biomedical Research* 14, 1981, pp.355-380.

[11] Lemkin, P., Lipkin, L., GELLAB: "A computer system for 2D gel electrophoresis analysis. III. Multiple gel analysis." *Computers and Biomedical Research* 14, 1981, pp.407-446,

[12] Lester, E.P., Lemkin, P.F., Lipkin, L.E., "New dimensions in protein analysis - 2D gels coming of age through image processing," invited paper, *Analytical Chemistry* 53, 1981, pp.390A-397A.

[13] Lemkin, P.F., Lipkin, L.E., GELLAB: "Multiple 2D electrophoretic gel analysis", in *Electrophoresis* '81, R. Allen, Arnaud (eds), W. De Gruyter, New York. 1981, pp.401-411.

[14] Lemkin, P.F., Lipkin, L.E., Lester, E.P., "Some extensions to the GELLAB 2D electrophoresis gel analysis system". Paper given at `Clinical Applications of 2D *Electrophoresis*', Mayo Clinic, Nov. 15-18, 1981. Clinical Chemistry 28, 1982, pp.840-849.

[15] Raghavan, S. and Garcia-Molina, H., "Crawling the Hidden Web". In proceedings *27th VLDB conference*, Roma, Italy, 2001, pp.129-138.

[16] URL:http:\\www.expasy.org/ch2d/, Swiss-2DPAGE, *Nucleic Acid Res*. 28, 2000, pp.286-288.

[17] URL:http:\\www.expasy.org/sw3d/, Swiss-3DIMAGE, *Trends Biochem. Sci.* 20, 1995, pp82-83.