International Journal of Software Engineering and Its Applications
Vol.7, No.5 (2013), pp.87-98
http://dx.doi.org/10.14257/ijseia.2013.7.5.09

A Spot Matching Algorithm using the Topology of Neighbor Spots in 2D-PAGE Images

Chan-Myeong Han[1], Dae-Seong Jeoune[2], Hwi-Won Kim[3],
and Young-Woo Yoon[4]
[1,4]Department of Computer Engineering, Yeungnam University
#280 Daehak-ro, Gyeongsan, Gyeongbuk, 712-749, Republic of Korea
[2]Department of Media Design, Daegu Future Colleage
#114 Mirae-ro, Gyeongsan, Gyeongbuk, 712-716, Republic of Korea
[3]Education Center for IT, Kyeongbuk College
#77 Daehak-ro, Yeungju, Gyeongbuk, 750-712, Republic of Korea
[1]cmhan@yu.ac.kr, [2]dsjeoune@dmc.ac.kr, [3]khw@kbc.ac.kr, [4]ywyoon@yu.ac.kr

Abstract

In this paper, a new spot matching algorithm is proposed. The algorithm compares topological patterns on two central spots to be matched that are selected each from reference and target 2D-PAGE images. Similarity transform is applied to one of two patterns in order to consider and correct global and local distortion before comparison of topological patterns. Matching between neighbor spots from two patterns is performed and similarity is evaluated using the normalized Hausdorff distance(NHD). Finally, matching of spots is determined by the number of neighbor matched pairs, the number of outlier spots and the NHD in turn. The proposed algorithm shows good results, even better matching performance when the relative topology is well preserved and no outlier is included.

***Keywords:** spot matching, neighbor spot, 5-NNG, spot topology, pivot spot, normalized Hausdorff distance*

1. Introduction

It is found that whole genome sequence cannot explain life phenomena enough and has a lot of limit to find disease related genes after successful genome sequencing of over 40 species [1]. Studies on proteins and interactions among them are considered as one of key fields because genes are expressed into proteins through mRNAs. Proteomics is the large-scale study of learning functions of proteins and the very basic processes are used to identify proteins included in cells. 2D-PAGE(two-dimensional polyacrylamide gel electrophoresis) is the most popular analysis method in Proteomics [2-4]. In the study of proteins, spot matching is a main bottleneck and the implementation of fast and precise spot matching algorithm with no intervention is the most essential part to upgrade Proteomics one level up [5]. However, it is impossible to produce the same gel image each time due to many experimental parameters even if the same sample is used for a couple of electrophoresis experiments. It makes spot matching more difficult.

For this reason, we propose a new spot matching algorithm of 2D-PAGE images which mimics human recognition process and treats spot matching as point pattern problems using graph theory. The algorithm determines matched pairs by comparing topological patterns of two central spots. The definition of neighbor spots and a method

ISSN: 1738-9984 IJSEIA

for similarity measure of topological patterns are proposed to produce better matching performance.

2. Previous Research

Spot matching by centroids of spots can be considered as point pattern matching problems [6]. The typical spot matching in 2D-PAGE gel image is a method by landmarks which are manually defined. Spots around landmarks are matched in turn [7-11]. Piecewise bilinear mapping is obtained using manual landmarks [8]. Initial matching is performed with landmarks and next matching is performed with best matching of neighbor spots [10]. Some methods enable users to check and correct the matching results [11]. A majority of conventional software use the landmarks that are defined manually. Nevertheless, a process of manually defined landmarks has high error rates because it is tedious and makes humans tired.

In recent, a method called iterative closest point is proposed as an automated protein spot matching [12], where spot matching is performed according to distances between matched pairs of spots from two sets of spots. Thus, parameters of non-linear transformation can be acquired. The calculated parameters are used in transforming gels non-linearly and distances between spot pairs are recalculated and the condition of converge is tested. ICP is to repeat a series of these processes. Euclidean distance and shape context distance are used as a distance measure. It assumes that 2-DE gel images are under non-linear deformation but they are actually under non-linear deformation only locally.

A method by hierarchical structure and minimization of energy is proposed [11]. The proposed algorithm for spot matching is an integration of the hierarchical-based and optimization-based methods. The hierarchical method is first used to find corresponding pairs of protein spots satisfying the local cross-correlation and overlapping constraints. The matching energy function based on local structure similarity, image similarity and spatial constraints is then formulated and optimized. There is a trial which uses a quadratic assignment formulation together with a correspondence estimation algorithm based on graph matching which takes into account the structural information between the detected spots [13]. Similarly some studies propose matching methods motivated by the preservation of topology. To compare the similarity of topology patterns, distances and angles among neighbor spots are compared [14].

3. Proposed Spot Matching Algorithm

3.1. Neighbor Spots

Two point patterns of reference gel and target gel are defined as $P= \{p_1,p_2,p_3,\ldots,p_m\}$, $Q=\{q_1,q_2,q_3,\ldots,q_n\}$, where $p_i=(x_i, y_i)$ and $q_j=(x_j , y_j)$ are the coordinates of the points in the x-y plane. Spot matching is to find a set of correspondence $\{(p_{i1}, q_{j1}), (p_{i2}, q_{j2}), \ldots , (p_{il}, q_{jl})\}$, where $m\neq n$, $l\leq n$ or $l\leq m$. Spots without correspondence are called outliers. After spot detection, each point of a given gel is represented as a certain graph with vertices, also called nodes or spots, and edges connection between them. For any two vertices in a graph, if there exists an edge connected between them, it is referred to as they are *adjacent* or *neighbor* and defined as shown in equation (1), where *Graph* is the name of the applied graph type.

$$N_{Graph}(v) = \{u \mid vu \in G\} \tag{1}$$

Different graph types form different edges for the same point pattern. As for the node *v* surrounded by its neighbor nodes, it is called *central node* or *central spot*. Here, we define the degree of node *v* as the number of neighbor nodes connected to node *v*, shown in equation (2).

$$Deg(v) =| N_{Graph}(v)\} \tag{2}$$

The definition of neighbor nodes depends on what kind of graph to be used for edges. Edges are defined by graph theory and neighbor nodes are determined by the edges. Different results are generated if different graphs are applied to the same point pattern. Therefore, it is necessary to find out which graph is the best for spot matching for 2D-PAGE images. In this paper, topological patterns of neighbor spots are utilized for spot matching. The definition of graph is tightly related to the definition of neighbor spots. Graph type affects the performance of spot matching and proper graph must be selected for spot matching. In the literature of reference [6], various types of graph such as Gabriel graph, Delaunay triangulation, 4-NNG, 5-NNG and 6-NNG were tested and 5-NNG was chosen as the best graph for spot matching. Therefore, 5-NNG is used to define neighbor spots in our spot matching algorithm.

3.2. Algorithm

A spot matching algorithm by topological patterns of neighbor spots estimates correspondence with similarity of patterns of neighbor spots $N_{Graph}(v)$ located around the central spots. Two sets of spots distributed around the spots to be matched are used when they are asked to match spots in 2D-PAGE. The proposed method in this paper is based on the human recognition process. If central spots p_i and q_j from reference gel and target gel are given, two sets of neighbor spots, $N_{Graph}(p_i)$ and $N_{Graph}(q_j)$ are extracted as depicted in Figure 1 and similarity between two patterns is compared to decide whether two central spots are a good match or not.

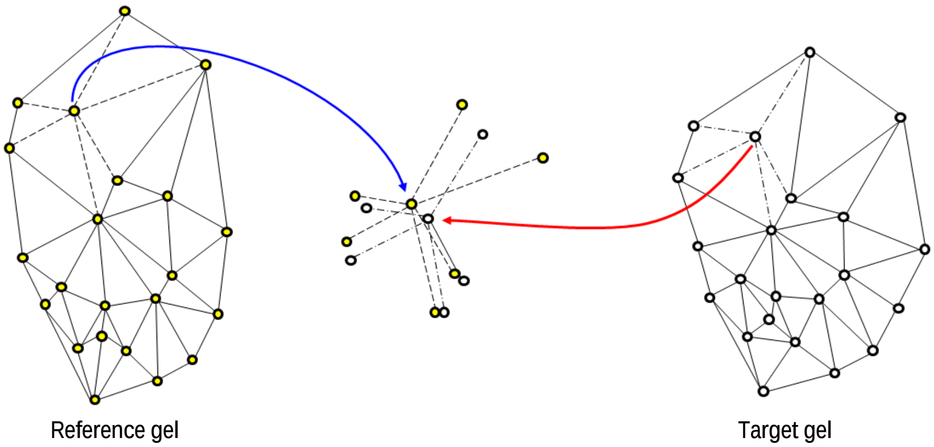


Figure 1. Spot Matching by Neighbor Spots

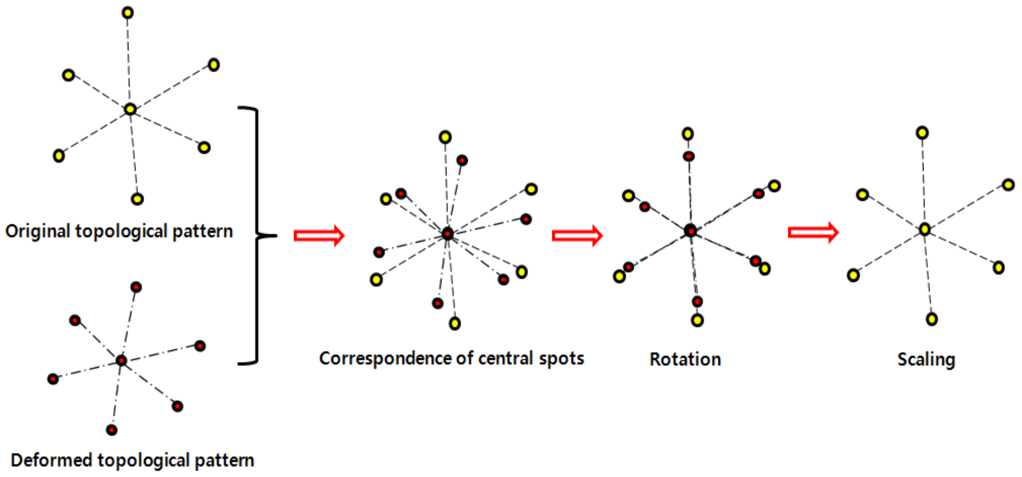

Original topological pattern
Deformed topological pattern
Correspondence of central spots
Rotation
Scaling

Figure 3. Similarity Transform

For correspondence estimation, topological patterns for two central spots are compared. They must be compared in their best state. The best state means the best positions of spots to be compared. Figure 3 shows how to move spots to their best positions. A series of transformation such as correspondence of central spots, rotation and scaling should be applied before comparison of topological patterns. Deformed topological pattern in Figure 3 is a pattern transformed with different rotation and scale parameters from original topological pattern. It is not good to compare two patterns without any change as in correspondence of central spots of Figure 3. Deformed topological pattern can be compared after being transformed with proper rotation and scale parameters. After correspondence of central spots, rotation and scaling, the deformed pattern becomes the exactly same pattern as original topological pattern and it concludes that two central spots are matched for two patterns are equal.

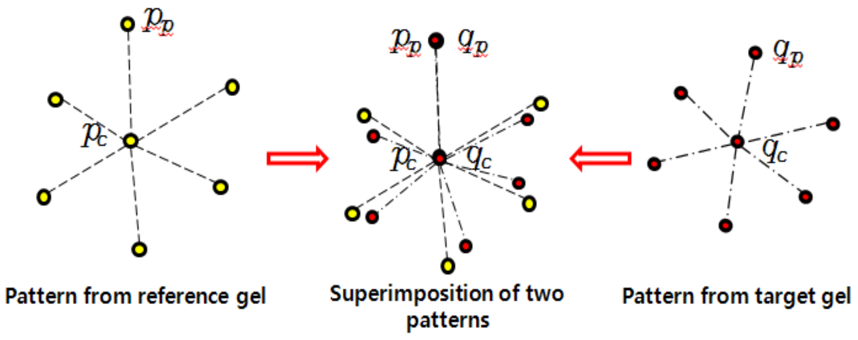


Figure 4. Correspondence of Topological Patterns

Transform involved with three parameters of transposition, rotation and scale is called *similarity transform*. Two pairs of spots whose matching is confirmed are needed to evaluate parameters of similarity transform as in Figure 4. The central spot pair (p_c, q_c) is automatically selected for it is under assumption that central spots are matched and the other pair can be chosen from one of neighbor spot pairs between $N_{Graph}(p_c)$ and $N_{Graph}(q_c)$. In

Figure 4, central spot pair *(p_c, q_c)* and neighbor spot pair *(p_p, q_p)* are selected for evaluating similarity transform parameters and two patterns are superimposed after the pattern $N_{Graph}(q_c)$ is transformed according to the evaluated parameters. Neighbor spot pair is called *pivot pair* because it is selected as a standard pair. Superimposition of two patterns in Figure 4 shows that central pair *(p_c, q_c)* and pivot pair *(p_p, q_p)* are exactly superimposed and other spots moved around according to parameters derived from central pair and pivot pair.

Pivot pair can be one of neighbor spot pairs. The problem is that these neighbor pairs are not known until pattern comparison ends. All possible combinations of pairs between $N_{Graph}(p_c)$ and $N_{Graph}(q_c)$ can be generated and each combination is used as pivot pair. Final pivot pair is derived from a neighbor pair where the degree of similarity is the highest. If the rotation parameter is evaluated over than 15 degree while comparison is performed, it must be dropped because 2D-PAGE gel images do not have severe rotation parameter and a neighbor pair with severe rotation has high possibility of a false neighbor pair.

Equation (3) represents how to calculate new coordinates of neighbor spots of q_j with scale parameters s and rotation parameters θ. The newly transformed coordinates make neighbor spots of q_j move to their best positions to be matched in similar geometric conditions.

$$\begin{bmatrix} x'_{q_j} \\ y'_{q_j} \end{bmatrix} = s \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_{q_j} - x_{q_c} \\ y_{q_j} - y_{q_c} \end{bmatrix} + \begin{bmatrix} x_{p_c} \\ y_{p_c} \end{bmatrix} \tag{3}$$

$$where,\ s = \frac{\sqrt{(x_{p_c} - x_{p_p})^2 + (y_{p_c} - y_{p_p})^2}}{\sqrt{(x_{q_c} - x_{q_p})^2 + (y_{q_c} - y_{q_p})^2}},$$

$$\theta = \tan^{-1}\left(\frac{y_{q_p} - y_{q_c}}{x_{q_p} - x_{q_c}}\right) - \tan^{-1}\left(\frac{y_{p_p} - y_{p_c}}{x_{p_p} - x_{p_c}}\right)$$

Matched pairs of Neighbor spots between $N_{Graph}(p_c)$ and $N_{Graph}(q_c)$ must be determined before similarity of two patterns is calculated for matched pairs are used in evaluation of similarity. Figure 5 shows that two patterns are superimposed on the axis of the central pair *(p_1, q_1)* after similarity transform by the central pair *(p_1, q_1)* and the pivot pair *(p_5, q_5)*. Dotted circles represent matched neighbor pairs where Euclidean distances of two spots are the shortest. Distances for all possible combinations of neighbor spots between $N_{Graph}(p_c)$ and $N_{Graph}(q_c)$ are calculated and one-to-one matched pairs are formed in order of shorter distance. Outlier spot, p_6 does not have correspondence for the degree of p_c is greater by one than the degree of q_c. Outliers can be produced even if the degree of p_c is equal to the degree of q_c. Threshold distance value for neighbor matched pairs is defined and two spots are not accepted as a matched pair if distance between them is greater than the threshold.

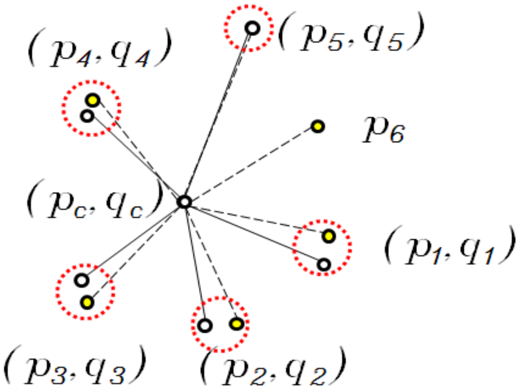


Figure 5. Matching Process of Neighbor Spots

3.3. Similarity Measure

Similarity can be calculated using neighbor matched pairs. The simplest way is to get an average of distances of all the matched pairs. The less this average is, the higher similarity is. However, Hausdorff distance is the most popular when getting similarity of patterns, which is *"the maximum distance of a set to the nearest point in the other set"*. More formally, Hausdorff distance from topological pattern of p_c to toplogical pattern q_c is a maximin function, defined as equation (4), where p_i and q_j are spots from patterns of p_c and q_c, respectively, and $d(p_i, q_j)$ is Euclidean distance function between these points.

$$h(P,Q) = \max_{p_i \in N_{Graph}(p_c)}(\min_{q_j \in N_{Graph}(q_c)}(d(p_i, q_j))) \qquad (4)$$

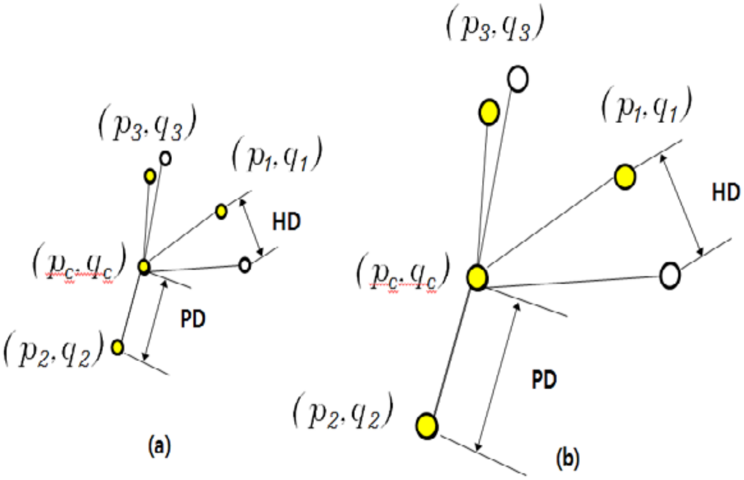


Figure 6. Normalization of Hausdorff Distances with Different Scale

Hausdorff distance, however, does not consider scale parameter and the calculated distances are different. As depicted in Figure 6, the scale of pattern (b) has twice compared to the pattern (a). Hausdorff distance of (b) is greater than that of (a) even though they are the same pattern comparison. Therefore, Hausdorff distance must be normalized by dividing with pivot distance. *Pivot distance(PD)* is a distance between central spot p_c and pivot spot p_2. The *normalized Hausdorff distance(NHD)* is defined as equation (5).

$$NHD = \frac{h(p_i,\ q_j)}{PD} \tag{5}$$

It often happens that the number of outliers is greater than the number of neighbor matched pairs and very short Hausdorff distance is evaluated due to very small number of the matched pairs. Many number of outlier spots means that there is less possibility where two central spots correspond to each other. For this reason, NHD is not enough as a criterion of matching for two spots. Therefore, three criteria, in this paper, are employed for better matching performance as follows; the number of the matched neighbor pairs, the number of outlier spots and the NHD, also summarized in Table 1.

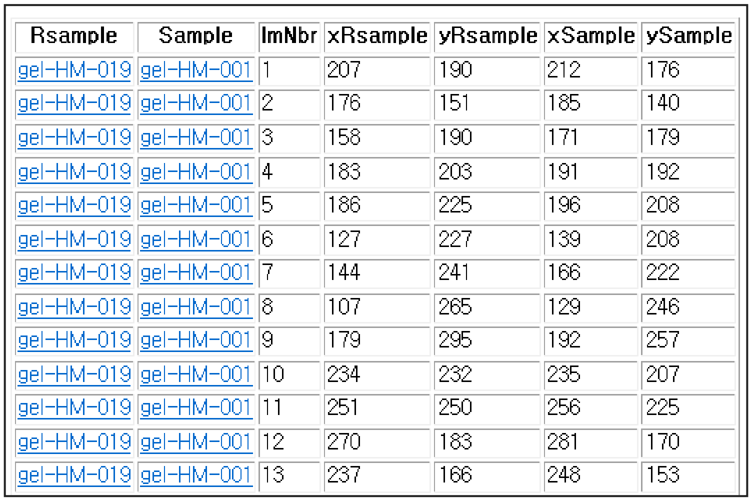
Table 1. Decision Criteria with Priority

Criteria	Priority	Precedence
Number of the matched neighbor pairs	0	more >> fewer
Number of outlier spots	1	more << fewer
Normalized Hausdorff distance	2	longer << shorter

4. Experimental Results

In 2D-PAGE, spot detection must be preceded before spot matching. In the stage of spot detection, centroids of spots are very important information for spot detection. This paper does not consider and omits spot detection stage because objective evaluation of algorithm for spot detection is also error-prone and it affects spot matching to a great extent. Data set called *"Human leukemias"* and *"HL-60 cell lines"* from the web site[1] is used. As an example, the set *"Human leukemias"* has 128 pairs of gels and each gel has 22 manual-matched pairs of spots or so.

[1] From the web site *http://www.lecb.ncnifcrf.gov/2DgelDataSets*, test data sets on 2D-PAGE can be obtained for convenience in experiments.



Rsample	Sample	ImNbr	xRsample	yRsample	xSample	ySample
gel-HM-019	gel-HM-001	1	207	190	212	176
gel-HM-019	gel-HM-001	2	176	151	185	140
gel-HM-019	gel-HM-001	3	158	190	171	179
gel-HM-019	gel-HM-001	4	183	203	191	192
gel-HM-019	gel-HM-001	5	186	225	196	208
gel-HM-019	gel-HM-001	6	127	227	139	208
gel-HM-019	gel-HM-001	7	144	241	166	222
gel-HM-019	gel-HM-001	8	107	265	129	246
gel-HM-019	gel-HM-001	9	179	295	192	257
gel-HM-019	gel-HM-001	10	234	232	235	207
gel-HM-019	gel-HM-001	11	251	250	256	225
gel-HM-019	gel-HM-001	12	270	183	281	170
gel-HM-019	gel-HM-001	13	237	166	248	153

Figure 7. Format of the File "landmark.tbl"

A text file named "landmark.tbl" is available from the web site and matching information of spots between reference gel and target gel is shown in Figure 7. *Rsample* and *Sample* are names of reference gel and target gel respectively and *ImNbr* is a series of matching numbers. *xRsample*, *yRsample*, *xSample* and ySample are the central coordinates of spots from reference gel and target gel. The corresponding two spots are on the same line, which means that the two spots consist of a single matched pair. The file "landmark.tbl" has information in one piece on 128 pairs of gels and 128 files are separated from it. Each gel pair has one-to-one matched pairs and there is no outlier. The same spot number is assigned for two spots of matched pairs and matching can be checked right if spots with the same spot number are matched.

The program language *perl* is used to implement the proposed algorithm and *python* with turtle graph library is used to visualize the matching results. Table 2 shows the summary of experimental results for the given data sets. The detection rate means total number of detected pairs including correct and false matches divided by the total number of pairs and the matching accuracy denotes ratio of the number of correct matched pairs over all of the detected pairs.

Table 2. Summary of Experimental Results

Measures	Data sets	
	Human leukmias	*HL-60 cell lines*
Total number of gel pairs	128	112
Total number of spot pairs	2,763	2,422
Detected spot pairs	2,716	2,191
False detected spot pairs	19	3
Detection rate (%)	98.30	90.46
Matching accuracy (%)	99.30	99.86

Through examining the results, the reason that leads to false match is analyzed. First, the difference between two coordinates of central spots from reference gel and target gel is too small to distinguish each other. Second, the spots are densely located in a certain local area and topology. Furthermore, topology of spots does not conserved due to the inherent characteristics of obtaining 2D-PAGE images. Figure 8 shows the topological patterns of the gel pair in *"HL-60 cell lines"*, where false matching occurred in the area of dotted box.

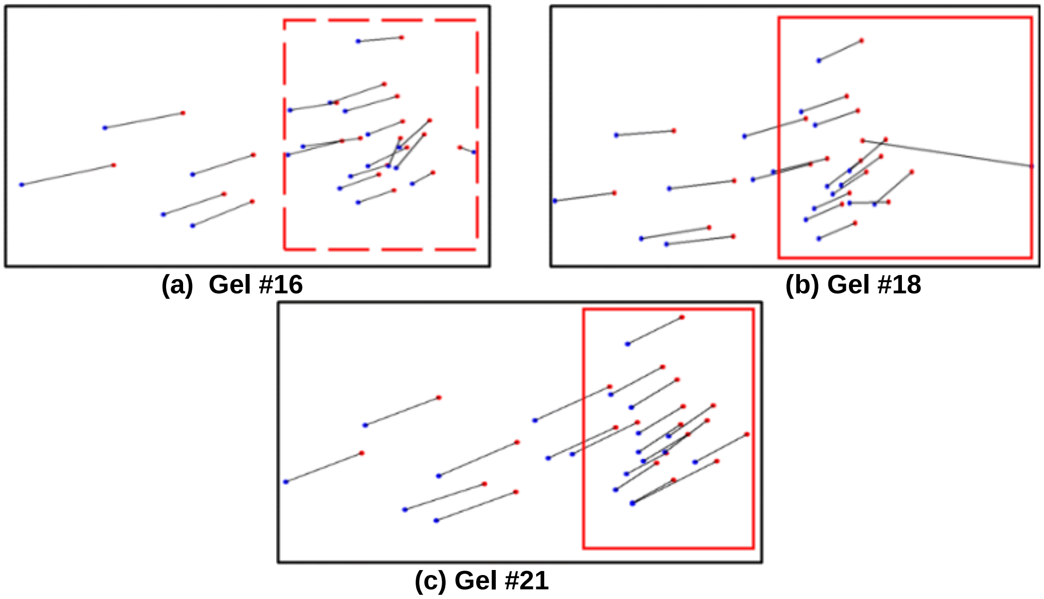


Figure 8. Topological Patterns of the Gel Pair in "HL-60 cell lines"

5. Concluding Remarks

In this paper, an algorithm for spot matching in 2D-PAGE images is proposed and is based on the definition of neighbor spot and the similarity of topological patterns. It defines a graph on neighbor spots and 5-NNG was proved to be the most proper graph for spot matching among various graph types. The proposed algorithm is capable of matching all of spots robustly under large global and local distortion. It results in outstanding output, especially under circumstances where the topological patterns of spots are well preserved and the corresponding two gels contain no outliers and missing spots.

Further research should be continued to enhance spot matching problems with outliers. In this case, virtual missing spots can be inserted to use the proposed method without any change. The proposed algorithm needs to be modified to produce better matching results when used together with various graph types and new determining criteria of similarity measure.

References

[1] Y.-S. Hwang and J.-H. Lee, “Matching spots in Electrophoresis Images by Topology Preserving Relaxation (in Korean)”, The Korean Institute of Information Scientists and Engineers: Software and Application, vol. 39, no. 6, **(2012)**, pp. 436-443.

[2] J. L. Harry, M. R. Wilkins and B. R. Herbert, “Proteomics: Capacity versus Utility”, Electrophoresis, vol. 21, **(2000)**, pp. 1071-1081.

[3] P. H. O'Farell, "High resolution two-dimensional electrophoresis of proteins", Biological Chemistry, vol. 250, **(1975)**, pp. 4007-4021.

[4] R. Westermeier, "Electrophoresis in Practice: A Guide to Theory and Practice", John Wiley & Sons Inc., **(1993)**.

[5] M. Daszykowski, E. Mosleth Faeregestad, H. Grove, H. Martens and B. Walczak, "Matching 2D Gel Electrophoresis Images with MatLab 'Image Processing Toolbox'", Chemometrics and Intelligent laboratory Systems, vol. 96, **(2009)**, pp. 188-195.

[6] C.-M. Han, S.-Y. Suk, M.-A. Kim and Y.-W. Yoon, "A Study of Neighbor Point for Point Pattern Matching", Proceedings of 6th International Symposium on Embedded Technology(ISET), Jeju, Republic of Korea, **(2011)** May 20-21.

[7] S. Veeser, M. J. Dunn and G. Yang, "Multiresolution Image Registration for Two-dimensional Gel Electrophoresis", Proteomics, vol. 1, **(2001)**, pp. 856-870.

[8] J. Salmi, T. Aittokallio, J. Westerholm, M. Griese, A. Rosengren, T. A. Nyman, R. Lahesmaa and O. Nevalainen, "Hierarchical Grid Transformation for Image Warping in the Analysis of Two Dimensional Electrophoresis Gels", Proteomics, vol. 2, **(2002)**, pp. 1504-1515.

[9] J. Garrels, "The QUEST System for Quantitative Analysis of Two-dimensional Gels", Journal of Biological Chemistry, vol. 264, no. 9, **(1989)**, pp. 5269-5282.

[10] K. Pleiβner, F. Hoffmann, K. Kriegel, C. Wenk, S.Wegner, A. Sahlstrom, H. Oswald, H. Alt and E. Fleck, "An Alternative Approach to Deal with Geometric Uncertainties in Computer Analysis of Two-dimensional Electrophoresis Gels", Electrophoresis, vol. 21, **(1999)**, pp. 2637-2640.

[11] T. Srinark and C. Kambhamettu, "An Image Analysis Suite for Spot Detection and Spot Matching in Two-dimensional Electrophoresis Gels", Electrophoresis, vol. 29, **(2008)**, pp. 706-715.

[12] M. Rogers and J. Graham, "Robust and Accurate Registration of 2-D Electrophoresis Gels using Point-Matching", IEEE Transactions on Image Processing, vol. 16, no. 3, **(2007)**, pp. 624-635.

[13] A. Noma, A. Pardo and R. M. Cesar Jr., "Structural Matching of 2D Electrophoresis Gels using Deformed Graphs", Pattern Recognition Letters, vol. 32, no. 1, **(2001)**, pp. 3-11.

[14] A. Hukhuu, J.-B. Lee and Y.-S. Hwang, "Automatic Matching of Protein Spots by Reflecting Their Topology (in Korean)", The KIPS Transactions: Part B, vol. 17-B, no. 1, **(2010)**, pp. 79-84.

Authors

Chan-Myeong Han received M.S and Ph.D. degree in image processing from the department of computer engineering, Yeungnam University, Korea, in 2007 and 2013, respectively. His research interests are image processing, image recognition, implementation of embedded system and bioinformatics. He is now working as a teaching assistance at Yeungnam University and plans to start his own business in image processing and image recognition filed.

Dae-Seong Jeoune achieved M.S. and Ph.D. degree at the department of computer engineering, Yeungnam University, Korea, in 1996 and 2002, respectively. After, he has been a full-time instructor at the department of multimedia information science, Daegu Future College, Korea, from 1999 to 2002. He worked at the intermediary non-profit organizations of TIPA and SEDA as a team chief from 2003 to 2010. And he was a visiting professor for planning to manage industry-academia cooperation at the GEERC and LINC, YNU, Korea, for a single year from 2011. Now he is an associate professor at the department of media design, DFC, Korea, since 2012. His research interest includes digital video processing, bioengineering, spot matching, new media, industrial security, informatization policy, etc.

Hwi-Won Kim graduated B.S. and M.S. from the department of electronic engineering, Yeungnam University, Korea, in 1984 and 1987, respectively. Also, He achieved Ph.D. degree at the department of computer engineering, Yeungnam University, Korea, in 2003. Since 1990, he is a professor at the Kyungbuk College, Korea and now charge of the annexed IT Education Center. Computer system and digital video processing are the major fields of his research interest.

Young-Woo Yoon graduated from the department of electronic engineering, Yeungnam University, Korea, in 1972. Also, he was awarded M.S. and Ph.D. degree at the same department and university in 1974 and 1984, respectively. Now, he is a professor of computer engineering department of Yeungnam University, Korea, since 1988. His recent research interest includes digital video processing, bioinformatics, protein spot matching, and biometric recognition.