International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 22 (2019) pp. 4119-4125
© Research India Publications. http://www.ripublication.com

A Fast Spot Matching Method for Determining Landmark Spot Pairs in 2D-PAGE

Chan-Myeong Han
Manager,
Raonsys Co. Ltd.
Daegu Metropolitan City, Republic of Korea.
ORCID: 0000-0002-9492-6685

Yun-Kyoo Ryoo
Associate Professor,
Dep't of Medical Computer Science, Daegu Health College
Daegu Metropolitan City, 41453, Republic of Korea.
ORCID: 0000-0001-7846-2182

Dae-Seong Jeoune*
Director,
Technology Research Center, A1 Engineering Inc.
Gyeongsan, Gyeongbuk, 38542, Republic of Korea.
ORCID: 0000-0003-1669-5700

Abstract

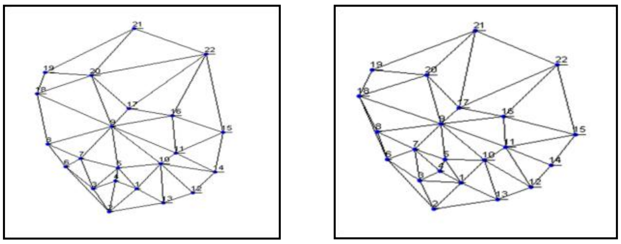
The grass fire spot matching algorithm in 2D-PAGE is fast and reliable, but it is required to determine one pair of spots called “seed spot pair” as starting point. The seed spot pair must be true-positive, otherwise it starts off on the wrong foot when the starting point is not right. Detecting the seed spot pairs in manual is most obvious, but it is not a good idea in automated spot matching process. Landmark spot pairs refer to a set of pairs that are easily identified visually or mathematically as definitely matched spot pairs. Only one spot pair with the highest reliability is selected and used as the seed spot pair among them. There are two kinds of algorithms proposed for detecting landmark spot pair. One is to find landmark spot pairs from various or multiple graphs and then combine the results to find the most reliable pair. The other is to compare only the sub-point patterns consisting of the first and second neighbors, and then scores them based on pattern similarity to select best one. Both algorithms have a problem that the amount of computation is higher than 2D-PAGE spot matching itself because they put a lot of emphasis on reliability. In this paper, a novel method is proposed, which determines the smaller number of reliable landmark pairs. However, it is much faster than the previous algorithms due to a lot of reduction in the number of comparison.

Keywords: 2D-PAGE, Spot Matching, Grassfire Algorithm, Landmark Spot Pair, k-NNG

INTRODUCTION

It is found that the whole genome sequence cannot explain life phenomena enough and has a lot of limitation 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. Proteomics is the large-scale study of learning functions of proteins and the very basic process is to identify proteins included in cells. The two-dimensional polyacrylamide gel electrophoresis, so called 2D-PAGE as an abbreviation, is the most frequently used method in proteomics [2, 3].

When analyzing images from two-dimensional gels, there is a reference gel image that represents the distribution of a sample of proteins in reference conditions, i.e. normal or healthy status. Proteins in the reference gel image are labelled and their spatial locations are confirmed manually. In the meanwhile, there are test gel images where spatial locations and kinds of proteins are unknown. Comparison between a test gel image and the reference gel image is performed in order to establish the correspondence between proteins, which is called ‘spot matching’. The unknown proteins from the test gel image can be identified by matched proteins from the reference gel image. The differences of protein composition between the reference gel image and the test one are very crucial clues for diseases or the mechanism of protein expression [4].



(a) Reference gel (b) Target gel

Figure 1. Two Point Sets with Different Local Distortions

The spot matching performed in 2D-PAGE is not a simple task. The spots generated from the result of 2D-PAGE include a lot of local distortions as well as global distortions. The Figure 1 shows how different they look when they have local distortions. The reference gel (a) and the target gel (b) are the same point set but it is not easy to identify spot pairs simply by superimposing two gels due to local distortions. For this reason, the false-positive detection rate is quite high, which would lead to a lot of errors even when a human perform the process.

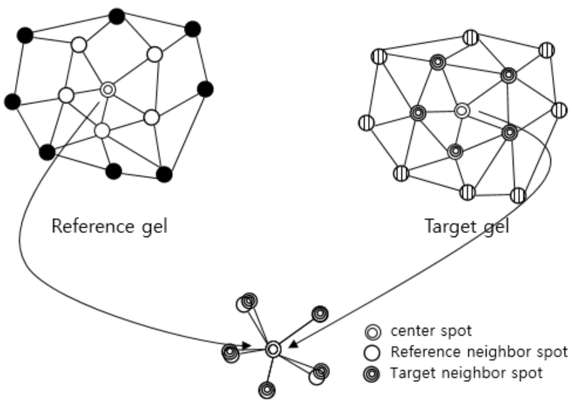


Figure 2. Spot Matching by Topological Patterns

Han *et al.* developed a method to compare the similarity among sub-patterns using the nearest neighbor graph, so called *k*-th NNG, to solve the problem depicted in the Figure 2. In the process of comparing using similarity, two subgraphs are superimposed onto the central spots as the center as in the Figure 3(a). A pair of spots other than the center pair is selected and one of the two subgraphs is rotated to adjust the rotation disparity as shown in the Figure 3(b). The pair used in aligning two subgraphs is called pivot pair. Finally a scale factor correction is applied to adjust the size as showed as in the Figure 3(c) [5]. Nevertheless, false-positive matching pairs are detected because this method takes advantage of comparisons for all the combinations of sub-patterns without considering topological locations of spots [6].

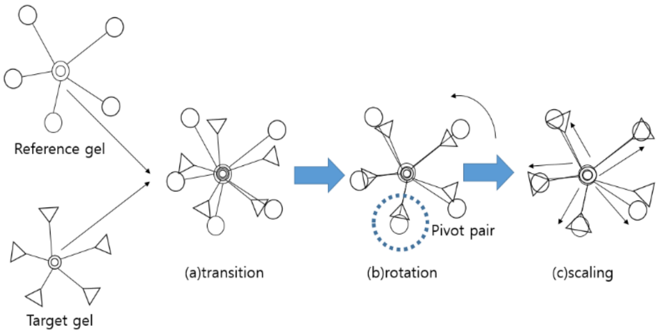


Figure 3. Similarity Transform

In the literature [7], Ryoo *et al.* proposed so-called the grass fire spot matching algorithm. It focuses on how the algorithm works to match the spots sequentially, as if a fire on the grass is spreading around as shown in the Figure 4. Since a 2D-PAGE image contains a large number of spots in it, there exist naturally many patterns which appear similar topologies of sub-patterns by chance. The grassfire algorithm solves false-positive detection problem by performing matching only near the previously matched place. It makes the matching speed drastically fast. However, it needs a single starting point called the ignition point.

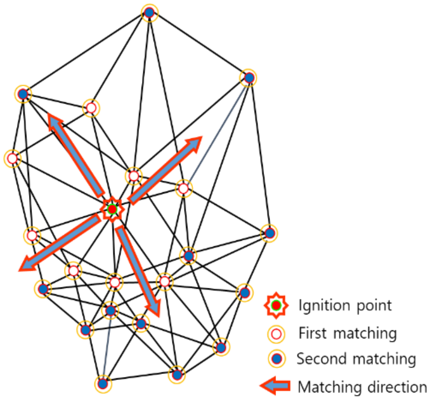


Figure 4. Concept of the Grassfire Spot Matching

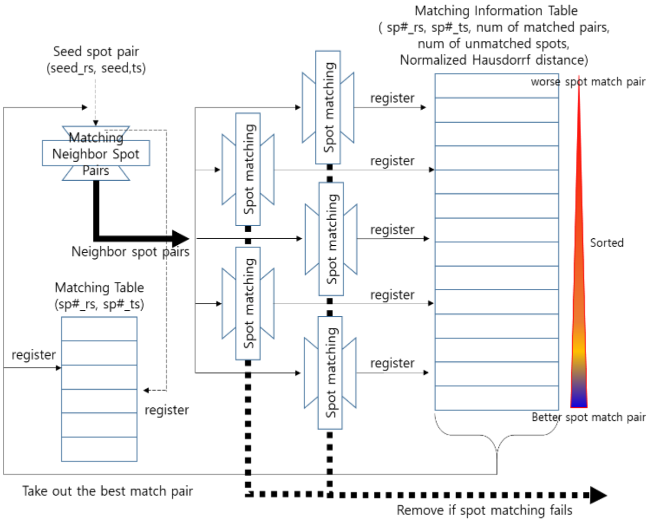


Figure 5. Architecture of the Grassfire Algorithm

The grassfire algorithm starts from a pair of seed spot and the neighbor spot pairs are obtained with respect to this points. They go through spot matching processes to get spot matching results such as the number of matched pairs, the number of unmatched spots and the normalized Hausdorff distance. The neighbor spot pairs with the matching results are saved into the matching information table as shown in the Figure 5. The neighbor spot pairs with poor matching results are chosen less because they are not thought to be the right match. The poor matching results are defined with thresholds for three parameters mentioned above.

For the next step, the best neighbor spot pair is taken out from the information table. It is registered in the matching table which stores final matching results and it is used as the center spot pair. Spot matching is continued and neighbor spot pairs are obtained. For every neighbor spot pair, spot matching with neighbor spot pairs as the center spot pairs is performed and the pairs which satisfy the threshold are saved into matching information table as in the Figure 5. The matching results can be doubly checked with the help of the matching table. If a matched neighbor pair is already registered in the matching table or it conflicts with information in the matching table, it should be discarded before being saved into the matching information table. The same process is repeated until the matching information table becomes empty with no entry [5].

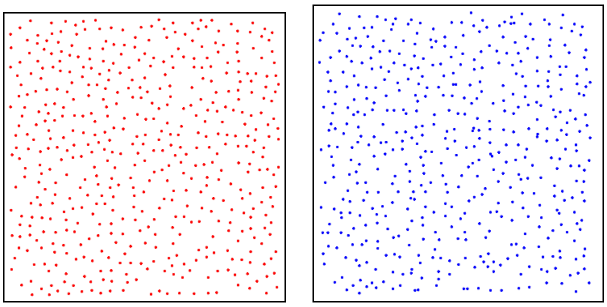
The ignition point must be a true-positive pair of matched spots. It is an easy task to detect it manually, but it is not recommended because a huge amount of gel images should be matched. It should be determined automatically in the algorithm itself to make the whole matching process automatic. So far, there are two algorithms proposed for automatic seed spot pair detection. One is to detect spot pairs using several k-th NNG iteratively and select one spot pair with the highest matching frequencies [8]. The other method is to detect spot pairs using sup-patterns of the first and the second neighbors and select one spot pair with the best matching score [9].

Both algorithms are very meaningful in that they automate the detection process of seed spots. On the other hand, the amount of computation for seed spot detection is much larger than the main process of spot matching. For this reason, it is necessary to improve the process to speed up the overall spot matching.

MATERIALS AND METHOD

A. Method for Landmark Spot Pairs using Multiple k-NNG Graphs

The nearest neighbor graph(NNG) for a set P with n objects in a metric space is a directed graph with its vertex set P and directed edges from p to q whenever q is a nearest neighbor. The k-th nearest neighbor graph (k-th NNG) is a graph in which two vertices p and q are connected by an edge, if the distance from p to q is the smallest among k distances from p to other objects in P.



(a) Reference gel (b) Target gel

Figure 6. Synthesized 2D-PAGE Images with 500 Pixels Each for Experiment

Spot matching is performed for two gel images, reference and target as shown in the Figure 6, after configuring NNG graphs from 5-th to 10-th as in the literature [3]. Spot pairs are cross checked for each k-th graph shown in the Table 1. Each spot pair has matching frequencies from 6 to 0. It has score 6 if they are successfully matched for all the six kinds of k-th graphs where k=5, 6, 7, 8, 9 and 10. While it has score 0 if it fails to be matched for all kinds of graphs. Spot pairs with score 6 are considered as the most reliable spot pairs, hence one of them would be selected as seed spot pair.

Table 1. Spot Matching Results using Six Kinds of k-th Nearest Neighbor Graphs

Spot Pairs (p, q)	Spot Matching Results using k-NNG						Matching Frequency
	k=5	k=6	k=7	k=8	k=9	k=10	
(301, 301)	x	x	x	x	x	x	6
(335, 335)	x	x	x	x	x	x	6
(225, 225)	x	x	x	x	x	x	6
(393, 393)	x	x	x	x	x	x	6
(15, 15)	x	x	x	x	x	x	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(94, 94)	x	x	x	x	-	x	5
(165, 165)	x	x	-	x	x	x	5
(96, 96)	x	x	x	x	-	x	5
(463, 463)	x	x	-	x	x	x	5
(21, 21)	x	-	x	x	x	x	5
(26, 26)	-	x	x	x	x	x	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(198, 198)	-	-	x	x	x	x	4
(111, 111)	x	-	x	-	x	x	4

B. Method for Detecting Landmark Spot Pairs using the 1st and the 2nd Neighbors

There is five neighbor spots for every spot in 5-th NNG. For the neighbor spots, 5-th NNG can be applied once again. The neighbor spots are called “the first(1st) neighbor spots” and the neighbor spots generated from the 1st neighbor spots are called “the second(2nd) neighbor spots”. In some cases, one spot can be the 1st neighbor spot and the 2nd neighbor spot at the same time.

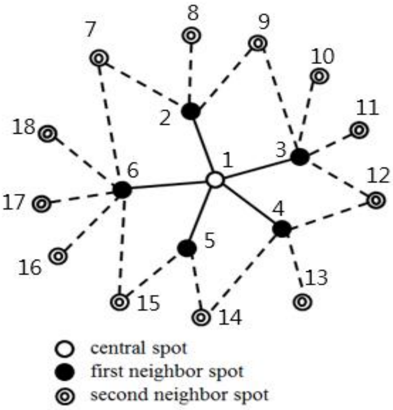


Figure 7. Sub-graph Configuration for a Central Spot Together with the 1st and the 2nd Neighbor Spots

For every spot as a central spot, each together with the 1st and the 2nd neighbor spots forms the topological pattern called "sub-graph" as shown in the Figure 7. Two sub-graphs, one from reference gel and the other from target gel, are compared to confirm the true-positive match for the central spot by measurement of similarity using the number of matched neighbor pairs, the number of unmatched spot and the normalized Hausdorff distance [9]. The best spot pair having the most matched neighbor spot pairs and the least normalized Hausdorff distance can be used as the seed spot pair. With this method, many true-positive spot pairs securely can be obtained. But, the burden of computation is huge because it performs matching process for all the possible combination of spot pairs even though only one best spot pair is needed.

C. Mathematical Definition for the 2nd Neighbor Spots

The 2nd neighbors and the number of the 2nd neighbors can be defined mathematically [7]. Neighbor spots are defined by whether there is an edge between two spots. A certain spot *v* has a spot *u* as a neighbor spot if an edge exists between *v* and *u*. Thus, the definition of neighbor spot is notated as shown in the equation (1). The number of neighbor spots for a spot *v* is called "degree of spot *v*" meaning the number of the edges that a spot *v* is connected to. So, it can be expressed as shown in the equation (2) As for the term "$N_G(v)$", the symbol "*N*" means "Neighbor spot" and the subscript "*G*" is the name of graph to be applied. The name of graph must be specified because the definition of neighbor spots depends heavily on the graph theory.

$$N_G(v) = \{u \mid vu \in E\} \tag{1}$$

$$\deg_G(v) = \mid N_G(v) \mid \tag{2}$$

For the sub-graph given in the Figure 2, the equations (1) and (2) can be more specified by example. The equation (3) denotes that the neighbor spots for the spot 1 as central spot are spots 2, 3, 4, 5 and 6. And the number of neighbor spots for the central spot 1 is 5 as shown in the equation (4).

$$N_{5-NNG}(1) = \{2,3,4,5,6\} \tag{3}$$

$$\deg_{5-NNG}(1) = \mid N_G(1) \mid = 5 \tag{4}$$

As the same manner, the mathematical definition for the 2nd neighbors can be given as the equations (5) and (6). As for the notation "$SN_G(v, u)$", "*SN*" means "the second neighbor". And, the variables *v* and *u* are the central spot and the first neighbor spot, respectively. The 2nd neighbor spots are spots 7, 8 and 9 for the central spot 1 and its 1st neighbor spot 2 as shown in the equation (7). The equation (8) shows the number of the 2nd neighbors for the central spot 1 and its 1st neighbor spot 2 is 3. Here, we should recognize that the number of the 2nd neighbor spots is not five when the graph applied is 5-NNG. This is why the central spot itself and the spots defined as the 1st neighbor spots from the other spots are excluded by definition. The equation (9) denotes total number of the 2nd neighbor spots for all the 1st neighbor spots for a given spot *v*.

$$SN_G(v,u) = \{w \mid uw \in E, w \neq v, w \notin N_{5-NNG}(v)\} \tag{5}$$

$$\deg 2_G(v,u) = \mid SN_G(v,u) \mid \tag{6}$$

$$SN_{5-NNG}(1,2) = \{7,8,9\} \tag{7}$$

$$\deg 2_{5-NNG}(1,2) = \mid SN_{5-NNG}(1,2) \mid = 3 \tag{8}$$

$$\deg 2_{5-NNG}(v) = \sum_{u \in N_{5-NNG}(v)} |SN_{5-NNG}(v,u)| \tag{9}$$

D. Screening Candidates using the Number of the 2nd Neighbors

The method proposed in the literature [9] is the best algorithm for getting the seed spot pair except its computation burden. In order to solve the problem, the idea starts to limit candidates only for spots with highly possibilities instead of matching for all the possible combination of spot pairs.

Let's assume two spots *p* and *q* that come from the reference gel and target gel, respectively. The topologies of sub-graphs with respect to *p* and *q* is very different if the spot pair (*p, q*) is not true-positive correspondence when trying to matching them. Topology of sub-graph for spots has a lot of information and it cannot be defined as a single value. Nevertheless, it can provide a hint for the similarity of two topologies such as the total number of the 2nd neighbor spots for *p* and *q*.

It is not always true that two spots (*p, q*) are the correspondence when they have the same total numbers of the second neighbor spots. To the contrary, it is always true that two spots (*p, q*) have the same total number of the second neighbor spots if they are a true-positive pair. This means that they have high possibility for matching and the target spots for *q* can be screened so that the computation burden is reduced efficiently.

E. Algorithm

As for the proposed method, the pseudo code for algorithm implementation is described in the Figure 8. The variables *ref_gel* and *tar_gel* are arrays including all the spot information for the reference gel image and the target one. Here, the spot information is simply composed of a spot number, *x*- and *y*-coordinate as shown in the Table 2. The variables *t_nghbr_ref* and *t_nghbr_tar* are tables or arrays for the neighbor spots for each spot(*rs*) of the reference gel image and each spot(*ts*) of the target gel image, respectively. The function of *get_nghbr_spot(spot_number, spot_array)* is to get the neighbor spots around the *spot_number* using *spot_array* which includes all the spot information. And the function of *get_sec_nghbr_spot(spot_number, first_ngbhr_array, spot_array)* is to get the second neighbor spots with respect to the *spot_number* with the first neighbor spots around the *spot_number* and *spot_array*.

```
01: for each spot rs ∈ ref_gel do
02:  for each spot ts ∈ tar_gel do
03:   t_nghbr_ref ← get_nghbr_spot(rs, ref_gel)
04:   t_nghbr_tar ← get_nghbr_spot(ts, tar_gel)
05:   for each spot rrs ∈ t_nghbr_ref do
06:    t_nghbr_ref2 ← t_nghbr_ref2 +
                 get_nghbr_spot(rrs, ref_gel)
07:    num_nghbr2_ref ← get_sec_nghbr_spot(rs,
              t_nghbr_ref, ref_gel)
08:    end for
09:   for each spot tts ∈ t_nghbr_tar do
10:    t_nghbr_tar2 ← t_nghbr_tar2 +
11:                  get_nghbr_spot(tts, ref_gel)
12:    num_nghbr2_tar ← get_sec_nghbr_spot(ts,
                    t_nghbr_tar,tar_gel)
13:   end for
14:    if num_neighbor2_ref <> num_neighbor2_tar then
15:     break;
16:     do_spot_matching(rs, reference_gel, ts, target_gel)
17:  end for
18: end for
```

Figure 8. Pseudo Code for the Proposed Method

Table 2. Example of Spot Information

Number of spot	Reference Gel		Target Gel	
	x	*y*	*x*	*y*
#1	203	42	205	47
#2	166	51	162	54
#3	145	69	141	71
#4	81	88	90	80
#5	215	108	213	105
#6	287	120	281	123
#7	109	120	112	123
#8	186	175	187	174
#9	111	182	120	186
#10	109	195	109	191
#11	25	205	23	205
#12	119	210	117	215
#13	140	214	143	217
#14	181	251	182	257
#15	251	254	251	251

The pseudo code makes the all the combination of *(rs, ts)* from the reference gel and the target gel using two *"for"* statements on line 1 and line 2. The number of second neighbor spots for each spot of the reference gel is obtained from line 5 to line 8. The number of second neighbor spots for each spot of the target gel is obtained from line 9 to line 13. The two numbers of the second neighbor spots for the reference gel and the target gel are compared on the line 14. If they are not equal, the spot matching process is not performed for the combination *(rs, ts)*. The spot matching process is performed only when the two numbers of the second neighbor spots are equal as shown on the line 16 in the Figure 8. It is mostly true that the two spots have the same number of the second neighbor spots if they are the correct matching pair. It doesn't matter that a few spot pairs doesn't follow the rule because the final goal is to get only one seed spot pair among the landmark spot pairs. It can reduce a lot of unnecessary computations by filtering the candidates of spot pairs which are not likely to be matched.

EXPERIMENT AND RESULT

A. Experiment

The matching accuracy is the most important performance in 2D-PAGE spot matching. This paper is based on the method in the literature [9], so all the conditions are set as the same as it. The spots in reference gel are generated randomly. And then, the spots from target gel are generated based on the spots in reference gel by adding some distortions along the Gaussian normal distribution as in the synthesis method of the literature [10]. Five hundreds of spots are generated with the minimum distance of 10 pixels in 512×512 gel size. The Gaussian normal distribution N(0,1) is used when the spots of target gel are transformed.

For each spot *p* in reference gel, the total number of the 2nd neighbor spots is obtained and the same process is taken for each spot *q* in target gel. The number of spot *q* to be tested for matching is counted for each spot *p*. The experiment repeats 10 times for a data set generated under the same condition above. The total number of spot *q* in target gel for all the spot *p* in reference gel is compared with the value 250,000 that is the number of computation 500×500 with no screening measure.

B. Result

The Table 3 shows the results for 10 experiments. The clause "Number of Computations for Comparison" in the second column is the number of times in which tests are performed after removing spot candidates with low possibilities using the proposed method in this paper. The method in the previous research of the literature [9] has 250,000 times without getting rid of any spot. The clause "Computational Reduction Rate" in the third column is the percent value where how much computation is removed compared to the previous method. It can be evaluated by the equation $B = 1\text{-}\{A/250{,}000\times100\}$.

Approximately 80% of computation burden can be removed in the most cases, which means the proposed method is 80% faster than the previous one in other words. Although the result shown is only for one pair of gel images, the consequence would be more meaningful when considering hundreds or thousands of gel pairs are generally tested in 2D-PAGE.

Table 3. Experiment results

Iteration of Experiments	Number of Computations for Comparison (A)	Computational Reduction Rate (B) (%)
1	51,709	79.3
2	47,044	81.2
3	48,119	80.8
4	50,710	79.7
5	50,182	79.9
6	48,829	80.5
7	48,429	80.6
8	48,425	80.6
9	49,693	80.1
10	49,422	80.2
AVERAGE	**49,256**	**80.29**

CONCLUSION

The detection of the seed spot pair for the grassfire spot matching algorithm in the previous research is a problem that should be improved in the whole spot matching process because it requires more computation than its main matching process. The matching process for a huge number of combination on reference and target gel images is involved. Besides a gel image consists of hundreds of spots. Therefore, automating the detection process for the seed spot pair is not enough in this environment. The detection should be not only fast but also trustworthy. The methods proposed in the previous research [8, 9] were computationally burdened by testing so much spot pairs needlessly. They sacrificed the cost focusing on the fact that the seed spot must be true-positive spot pair at any circumstances.

In this paper, a fast automatic method for detecting the landmark spot pairs is proposed. It lowers the computational burden by performing the spot matching process only when the total number of the second neighbor spots for every spot in reference gel is equal to that of the spot in target gel. Therefore, it removes the problem in selecting the seed spot pair efficiently.

The future works are as follows. First of all, an empirical study is needed on how much the results are to be changed when data sets generated from various conditions for distortions are applied to the proposed method. Also, it is needed to study the effectiveness of the proposed method when random errors such as rotation and scale are added as in real gel data.

ACKNOWLEDGEMENT

The corresponding author is Dae-Seong Jeoune.

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, pp. 436-443, 2012

[2] J. L. Harry, M. R. Wilkins and B. R. Herbert, "Proteomics: Capacity versus Utility," Electrophoresis, Vol. 21, pp. 1071-1081, 2000

[3] P. H. O'Farrell, "High Resolution Two-Dimension Two-Dimensional Electrophoresis of Proteins," Journal of Biological Chemistry, Vol. 250, No. 10, pp. 4007-4021, 1975

[4] Freire A., Seoane J.A., Rodrıguez A., Ruiz-Romero C., Lopez-Campos G. and Dorado J., A Block-matching based technique for the analysis of 2D gel images, Studies in Health Technology and Informatics, Vol. 160, pp. 1282-1286, 2009

[5] Dae-Seong Jeoune *et al.*, Minutiae-based Fingerprint Matching, International Journal of Applied Engineering Research, Vol. 12, No. 9, pp. 1935-1942, 2017

[6] C.-M. Han *et al.*, A Spot Matching Algorithm using the Topology of Neighbor Spots in 2D-PAGE Images, International Journal of Software Engineering and Its Applications, Vol. 7, No. 5, pp. 87-98, SERSC, September 2013

[7] Yun-Kyoo Ryoo, Chan-Myeong Han, Ja-Hyo Ku, Dae-Seong Jeoune and Young-Woo Yoon, "Grassfire Spot Matching Algorithm in 2-DE", International Journal of Bio-Science and Bio-Technology, Vol. 5, No. 4, pp. 167-174, SERSC, September 2013

[8] Han, C.-M., Jeoune, D.-S. and Ryoo, Y.-K., An automatic detection of landmark spot pairs for iterative spot matching algorithms. Information (Japan). Vol. 18, pp. 1213-1218, 2015

[9] Dae-Seong Jeoune, Chan-Myeong Han and Wook Hyun Kim, "Fully Automated Detection of Landmark Spot Pairs using the Topology of Both the First and the Second Neighbor Spots in Two-Dimensional Electrophoresis", International Journal of Applied Engineering Research", Vol. 11, No. 18, pp. 9448-9454, 2016

[10] D.-S. Jeoune *et al.*, Synthesis of 2-DE Gel Image for Various Spot Matching Applications, International Symposium on Advanced and Applied Convergence (ISAAC 2014), pp. 112-115, 2014