Grassfire Spot Matching Algorithm in 2-DE

Yun-Kyoo Ryoo[1], Chan-Myeong Han[2], Ja-Hyo Ku[3], Dae-Seong Jeoune[4],
and Young-Woo Yoon[2]

*[1] Department of Medical Computer Science, Daegu Health College
#15 Youngsong-ro, Buk-gu, Daegu, 702-722,
[2]Department of Computer Engineering, Yeungnam University
#280 Daehak-ro, Gyeongsan, Gyeongbuk, 712-749, Republic of Korea
[3] Department of Computer Engineering, Kyungwoon University
#730 Gangdong-ro, Sandong-myun, Gumi, Gyeongbuk, 730-739, Republic of Korea
[4] Department of Media Design, Daegu Future Colleage
#114 Mirae-ro, Gyeongsan, Gyeongbuk, 712-716, Republic of Korea
kyoo@dhc.ac.kr[1], cmhan@yu.ac.kr[2], ywyoon@yu.ac.kr[2], jhku@iwk.ac.kr[3],
dsjeoune@dmc.ac.kr[4]*

Abstract

Grassfire method for spot matching is proposed based on similarity comparison of topological patterns for neighbor spots. Grassfire method is an algorithm where spot matching is performed as if fire spreads all around on grass. Spot matching starts from a seed spot pair confirmed as a matched pair of spots and spot matching spreads to the direction where the best matching result is produced. In this paper, the simple type of grassfire method where a seed spot pair is manually selected and spot matching is conducted under the circumstances without outlier spots is studied to examine the potential of grassfire method. The proposed method outperforms matching methods by random combination of spots in terms of speed and accuracy.

Keywords*: grassfire, spot matching, neighbor spot, topological pattern, seed spot pair*

1. Introduction

Two-dimensional electrophoresis(2-DE) is a widely used method for protein separation used in the field of Proteomics [1, 2]. The basic principle of electrophoresis is to move proteins to their positions by isoelectric point and molecular weight of protein on two-dimensional gel. Various types of spots in shape and size are seen all over the gel after the gel electrophoresis is finished and the spots are separated proteins. Positions where spots are staying are crucial clues for identifying spots.

It is needed to study expression, extinction and change of proteins from a certain tissue in protein research. Different environments make different proteins expressed in the same tissue and reference gel and target gel are compared to track down changes on the constitution of proteins. Reference gel is a standard sample of a tissue under the normal environment and target gel is a sample to be tested to study differential protein expression or diagnose diseases.

Manual comparison of two gels is a very time-consuming and boring process because thousands of proteins are usually included in one gel. Automating analytical processes of 2-DE is required for this reason [3]. 2-DE is a very simple experimental method but relatively huge variation is involved in the result. Considerable differences in the result are created even though the same experimental tools and the same sample are used in the same laboratory.

Variation of result in inter-laboratory experiments is even more increasing. Positional variation is the most common variation and it is one of the main causes which make it more challenging to automate the analytical processes. Nevertheless, many tries to solve these problems have been made and many methods have been proposed as the result. Analytical processes of 2DE are categorized into two stages. One is "spot detection" where spots and background are distinguished from digitalized gel image and the other is "spot matching" where spots from reference gel correspond to spots from target gel if two spots are the same protein.

For spot matching, earlier techniques require extensive user involvement, especially in initial spot pairing (GELLAB [4, 5]). Various transform functions are used as a basis for spot matching, among the most popular are piecewise bilinear mapping [6-8], Delaunay triangulation [9-12], and radial basis functions [13, 14]. Most of these transform functions rely on pertinent assumptions that are not satisfied by actual gel images. Among the most common problems are spot overlapping, light spot handling, and noise.

In this paper, a new method for spot matching is proposed based on similarity comparison of topological patterns for neighbor spots [15]. The idea of the proposed method is from the way grass fire spreads all around. A pair of spots from reference gel and target gel is selected as a location of ignition where fire starts or matching starts and matching is performed as if fire spreads all around grass.

2. Definition

Spot matching starts with two sets of 2D points, $P=\{p_1, p_2, \ldots, p_m\}$ *and* $Q=\{q_1, q_2, \ldots, q_n\}$ where centroids of spots from reference gel image $pi=(x_i, y_i)$ and centroids of spots from target gel image $q_j=(x_j, y_j)$. Spot matching is to find the maximum set of one to one matching pairs between P and Q, $M=\{(p_{i1}, q_{j1}), (p_{i2}, q_{j2}), \ldots, (p_{il}, q_{jl})\}$ where $p_{il} \in P$, $q_{jl} \in Q$, $m \neq n$, $l \leq min(m, n)$.

In thesis [15], two sets of neighbor spots for two spots to be matched are defined and topological patterns formed by them are compared to determine whether the two spot correspond to each other. It is very important which graph must be applied in a set of spots because it directly affects the definition of neighbor spots, which means different graphs define different neighbor spots. Figure 1 presents an example where 3-NNG is applied to a set of spots, set V. Modified 3-NNG is used here so that edges can be bi-directed. A graph is defined by a set of spots and a set of edges between vertexes and it can be notated as formula (1), (2) and (3).

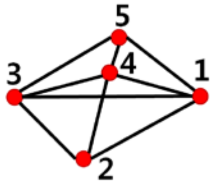


Figure 1. Example of a Graph with Three Neighbor Nodes(3-NNG)

$$V = \{1,2,3,4,5\} \tag{1}$$

$$\begin{aligned} E = \{&(1,2), (1,3), (1,4), (1,5), (2,1), (2,3), (2,4), (3,1), (3,2), (3,4), \\ &(4,1), (4,2), (4,3), (4,5), (5,1), (5,3), (5,4)\} \end{aligned} \tag{2}$$

$$G = (V, E) \tag{3}$$

Edges between spots are created by the definition of a graph after it is applied to a set of spots. Neighbor spots are defined by whether there is an edge between two spots. Spot *v* has spot *u* as a neighbor spot if an edge exists between *v* and *u* and the definition of neighbor spot is notated as formula (4). The number of neighbor spots for spot *v* is called "degree of spot *v*" and it is notated as formula (5). "*N*" in the notation "$N_G(v)$" is the first letter of the word "Neighbor" and 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.*

$$N_G(v) = \{u \mid vu \in E\} \tag{4}$$

$$\deg_G(v) = | N_G(v) | \tag{5}$$

In Figure 1, neighbor spots for spot *5* are as formula (6) and the degree of spot *5* is as formula (7)

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

$$\deg_{3-NNG}(5) = | N_{3-NNG}(5) | = 3 \tag{7}$$

There are many graph theories but Delaunay triangulation, Gabriel graph, relative neighbor graph and k-NNG are frequently used for spot matching problems. In this paper, 5-NNG is used for the definition of neighbor spot based on thesis [16].

3. Spot Matching using Grassfire Method

3.1. Topological Transform of Neighbor Spots

In Figure 2, correspondence test between spot p_i from reference gel and spot q_j from target gel is performed by evaluating similarity of topologies formed by neighbor spots of p_i, $N_{5\text{-}NNG}(p_i)$ and neighbor spots of q_j, $N_{5\text{-}NNG}(q_j)$[15]. Topological comparison of neighbor spots is made not by superimposing two central spots (p_i, q_j) but by transforming coordinates of neighbor spots of q_j considering translation, scale and rotation parameters to neighbor spots of p_i so that the same conditions of comparison are applied to both topologies.

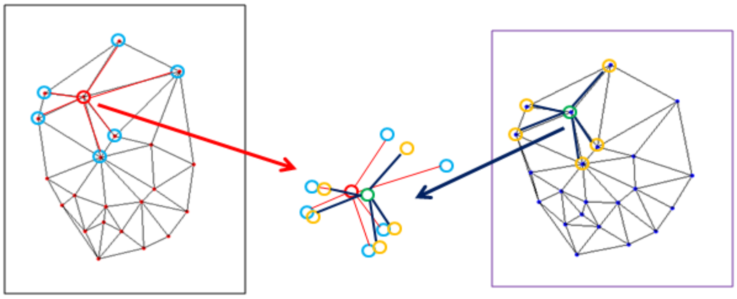


Figure 2. Spot Matching using Topological Comparison of Neighbor Spots

Two pairs of spots are needed to calculate similarity transform parameters; translation, scale and rotation. One is selected from a pair of central spots (p_i, q_j) by assuming they correspond to each other because the spots are being tested for whether they are in the relationship of correspondence. It is called “central spot pair” (*CSP*). The other is selected among matched pairs of neighbor spots. All the possible combination of neighbor spots, $PC=\{(np_i, nq_j) \mid np_i \in N_{5\text{-}NNG}(p_i), nq_j \in N_{5\text{-}NNG}(q_j)\}$ are tested because matching neighbor spots is not performed and correspondence of neighbor spots is not known yet. The pair with the highest similarity is finally selected as pivot spot pair (*PSP*) among them. *PSP* is a very important pair because it is used with *CSP* for calculating similarity transform parameters.

PSP is usually selected among pairs which are actually in the corresponding relationship among *PC* and the matched pair of neighbor spots which produces the highest similarity in topological comparison is determined as *PSP*. Pairs which are not in the corresponding relationship among *PC* can be filtered easily and quickly by limiting rotation parameter below 15 degrees because severe rotation never happens in 2-DE.

After two pairs $CSP=(p_c,q_c)$, $PSP=(p_p, q_p)$ for getting similarity parameters are determined, formula (8) and formula (9) are used respectively for calculating scale parameter s and rotation parameter θ. Spots of CSP and PSP are centroids of spots on two dimensional plane and can be presented as $p_c=(x_{pc}, y_{pc})$, $q_c=(x_{qc}, y_{qc})$, $p_p=(x_{pp}, y_{pp})$ and $q_p=(x_{qp}, y_{qp})$.

$$s = \frac{\sqrt{(x_{pc} - x_{pp})^2 + (y_{pc} - y_{pp})^2}}{\sqrt{(x_{qc} - x_{qp})^2 + (y_{qc} - y_{qp})^2}} \tag{8}$$

$$\theta = a\tan(\frac{y_{qp} - y_{qc}}{x_{qp} - x_{qc}}) - a\tan(\frac{y_{pp} - y_{pc}}{x_{pp} - x_{pc}}) \tag{9}$$

$$\begin{bmatrix} x'_{qj} \\ y'_{qj} \end{bmatrix} = s\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}\begin{bmatrix} x_{qj} - x_{qc} \\ y_{qj} - y_{qc} \end{bmatrix} + \begin{bmatrix} x_{pc} \\ y_{pc} \end{bmatrix} \tag{10}$$

Transformed coordinates of neighbor spots of q_j are obtained using formula (10) by assigning s and θ from formula (8) and formula (9). These transformed coordinates are the result of making neighbor spots of q_j move to their best positions for topological comparison.

3.2. Evaluation of similarity for topological patterns of neighbor spots

Topological pattern in point matching problem means a topology formed by a set of neighbor spots. Similarity of topological patterns for two central spots can be used in spot matching. There are many methods in evaluating similarity of topological patterns and Hausdorff distance is the most popular method. Hausdorff distance can be formulated as in formula (11). Matched pairs of neighbor spots are found and Euclidean distances between two spots from the matched pairs are calculated respectively. Hausdorff distance is the maximum distance among them.

$$\begin{gathered} h(NP, NQ) = \max_{npi \in NP}(\min_{nqi \in NQ}(d(np_i, nq_i))) \\ where, NP = N_{5-NNG}(p_i), NQ = N_{5-NNG}(q_j) \end{gathered} \tag{11}$$

Matched pairs of neighbor spots are acquired by generating all of the possible combination $PC'=\{(np_i, nq'_j) \mid np_i \in N_{5\text{-}NNG}(p_i), nq'_j \in similarity\_transform(N_{5\text{-}NNG}(q_j), s, \theta, p_c)\}$ and getting Euclidean distances for them. Pairs whose distance is the shortest are selected as matched pairs one by one. It should be noted that neighbor spots of q_j are not original ones but ones transformed by similarity transform. Two spots from one matched pair cannot correspond to other spots any more once they are matched because spot matching in 2-DE is one to one matching. It shows how to get matched pairs of neighbor spots and Hausdorff distance.

Comparison of Hausdorff distances for spot matching is valid under the same scale factor. If the scale factors are different, Hausdorff distances must be normalized to make a fair comparison as in formula (12) [15]. Unmatched spots can exist in matching of neighbor spots because degrees of p_i and q_j might be different. The problem is that normalized Hausdorff distances (*NHD*) cannot be a correct criterion for comparison of similarity when unmatched spots are more than the number of matched pairs. *NHD* is the maximum distance among those of matched pairs for neighbor spots and it might be very short by chance under the circumstances of many unmatched spots. Therefore, three factors; the number of matched pairs (*NOMP*), the number of unmatched spots (*NOUS*) and *NHD* must be considered together with the priority order of *NOMP*, *NOUS* and *NHD* [15].

$$
\begin{aligned}
& NHD(NP, NQ) = \frac{h(NP, NQ)}{PD} \\
& where,\ \ NP = N_{5-NNG}(p_i), NQ = N_{5-NNG}(q_j) \\
& \qquad PD = dist(p_c, p_p)
\end{aligned}
\tag{12}
$$

3.3. Grassfire Method

Once how to compare topological patterns of neighbor spots is settled, spot matching can be performed with all of the possible combination of central spots, *(p_i, q_j)* as in matching of neighbor spots. The only difference is that Euclidean distance is a measure for matching of neighbor spots and similarity is for matching of central spots. However, matching methods by random combination have computational burden and higher similarity in some randomly combined pairs might be produced than that of real matched pair in rare cases.

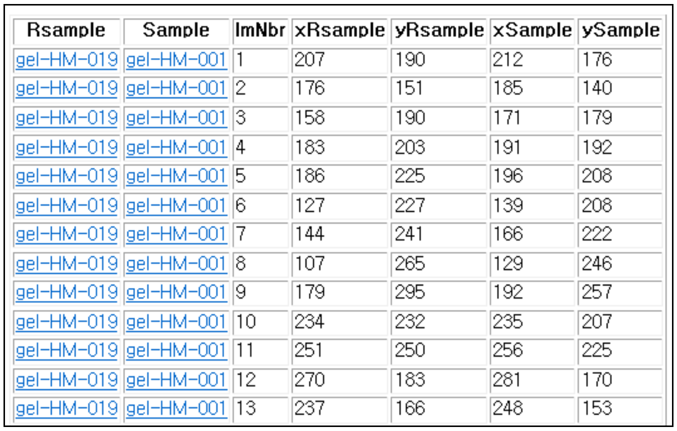
In this paper, grassfire method is proposed to solve the two problems. The Idea of grassfire method is from the way fire spreads all around on grass after fire starts at the center. Matching is performed as fire spreads in all directions from one location. One matched pair is needed as a location of ignition where matching starts. It is named “seed spot pair (*SSP*)”. One of advantages in grassfire method is that previous result of matching can be used as a hint for next stage of matching and it helps the algorithm produce more accurate result in a shorter time than matching methods by random combination.

There are two key issues in grassfire method. One is how to determine *SSP* and the other is which direction to make fire spread. First, *SSP* can be found manually or automatically. Many theses on landmark spot can be referred for the automated selection of SSP. Secondly the direction of matching is a hot issue where many methods can be considered. It means there is a lot of room for studying. This paper tests a method where SSP is selected manually and fire spreads in the direction of the best result of matching to examine the usability of grassfire method.

4. Experiment and Results

In 2-DE, spot detection must be preceded before spot matching. Centroids of spots obtained from the stage of spot detection are very important information for spot detection. The stage of spot detection is omitted for objective evaluation of grassfire method as spot matching because spot detection is error-prone and it affects spot matching to a great extent.

In the experiment, data set from the web site[1] is used. This set has 128 pairs of gels and each gel has approximately 22 manually matched pairs of spots. A text file named "landmark.tbl" can be downloaded from the web site and matching information of spots between reference gel and target gel is shown as in Figure 3.



Rsample	Sample	lmNbr	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 3. Format of Landmark.tbl

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 numbers are assigned for two spots of a matched pair and matching can be checked right if two matched spots have the same spot number. The program language perl is used to implement the proposed algorithm and python with turtle graph library is used to visualize the matching results. Matched pair number 1 is designated as SPP for all of 128 gel pairs.

Table 1. Summary of Experimental Results

Measures	Values
Total number of gel pairs	128
Total number of spot pairs	2,763
Number of detected spot pairs	2,762
Detection rate	99.96%
Matching accuracy	100%

Table 1 shows that 2,762 pairs are detected as matched pairs among 2,763 of total number of pairs. Detection rate and matching accuracy are 99.96% and 100%, respectively. The detection rate means total number of detected pairs including false positive and true positive divided by 2,763 of total number of matched pairs. Matching accuracy means a rate of the number of right matched pairs among the detected pairs.

5. Conclusion

In this paper, grassfire method is proposed and it shows better performance in speed and accuracy than that of the previous research [15]. Previous results of matching are used as a guide or a hint for the next stage of matching and it enables the proposed method to produce more reliable result very quickly. Furthermore, the proposed method outperforms methods by

[1] *http://www.lecb.ncnifcrf.gov/2DgelDataSets*

random combination of pairs because directions where fire spreads or matching is performed are determined toward the best result of matching. Grassfire method is tested under circumstance without outlier spots in this paper but it is expected to present excellent performance in the case of spot matching with outlier spots. The future work will be to elaborate on grassfire method to work well in spot matching with outlier spots.

References

[1] P. H. O'Farrell, Journal of Biolological Chemistry, vol. 250, **(1975)**, pp. 4007-4021.

[2] J. Klose, Humangenetik, vol. 26, **(1975)**, pp. 231-243.

[3] 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.

[4] P. F. Lemkin and L. E. Lipkin, "Computers and Biomedical Research", vol. 14, **(1981)**, pp. 355-380.

[5] P. F. Lemkin and L. E. Lipkin, "Computers and Biomedical Research", vol. 14, **(1981)**, pp. 407-446.

[6] S. Veeser, M. J. Dunn and G. Z. Yang, Proteomics 2001, vol. 1, **(2001)**, pp. 856-870.

[7] J. Salmi, T. Aittokallio, J. Westerholm and M. Griese, Proteomics 2002, vol. 2, **(2002)**, pp. 1504-1515.

[8] J. S. Gustafsson, A. Blomberg and M. Rudemo, Electrophoresis 2002, vol. 23, **(2002)**, pp. 1731-1744.

[9] K. P. Pleißner, F. Hoffmann, K. Kriegel and C. Wenk, Electrophoresis 1999, vol. 21, **(1999)**, pp. 2637-2640.

[10] A. Efrat, F. Hoffmann, K. Kriegel and C. Schultz, Journal of Computational Biology, vol. 9, no. 2, **(2002)**, pp. 299-315.

[11] K. Kaczmarek, B. Walczak, S. de Jong and B. G. M. Vandeginste, Journal of Chemical Information and Computer Science, vol. 42, **(2002)**, pp. 1431-1442.

[12] K. Kaczmarek, B. Walczak, S. de Jong and B. G. M. Vandeginste, Journal of Chemical Information and Computer Science, vol. 43, **(2003)**, pp. 978-986.

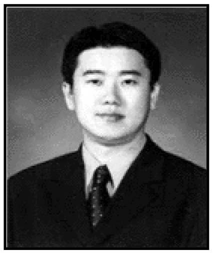
[13] M. Rogers, J. Graham and R. P. Tonge, British Machine Vision Conference, **(2004)**.

[14] M. Rogers, J. Graham, R. P. Tonge, IEEE International Symposium on Biomedical Imaging, **(2004)**.

[15] C.-M. Han, S.-Y. Suk, H.-W. Kim and Y.-W. Yoon, "A Study For Point Pattern Matching Using Local Matching Based On Neighbor Points(in Korean)", Institute of Embedded Engineering of Korea, National Symposium on Embedded Technology, **(2011)**.

[16] C.-M. Han, S.-Y. Suk, M.-A. Kim and Y.-W. Yoon, "A Study of Neighbour Point for Point Pattern Matching," ISET, **(2011)**.

Authors



Yun-Kyoo Ryoo received a bachelor's degree at the department of business administration, Yeungnam University, Korea, in 1995 and M.S. degree at the department of computer engineering, Kyungil University, Korea, in 1997. At present, he has been Ph.D. program at the department of computer engineering, Yeungnam University, Korea. Since 1998, he is working as a professor at the department of medical computer science, Daegu Health Collge, Korea. His research interests are image processing, pattern recognition, and bioinformatics.

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.

Ja-Hyo Ku received B.S., M.S. and Ph.D. degree at the department of computer engineering from the Yeungnam University, Korea, in 2000, 2002, and 2008 respectively. Since 2012, he has been a professor at the computer engineering department of Kyungwoon University, Korea. His current interests are smart mobile, convergence computing, and privacy protection.

Dae-Seong Jeoune achieved M.S. and Ph.D. degree at the department of computer engineering, Yeungnam University, Korea, in 1996 and 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 of YNU, Korea in 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.

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.