Fuzzy Similarity Measure and Fractional Image Query for Large Scale Protein 2D Gel Electrophoresis*

Daw-Tung Lin[1], Juin-Lin Kuo[2], En-Chung Lin[3], and San-Yuan Huang[4]

[1] Department of Computer Science and Information Engineering
National Taipei University, 151, University Rd., San Shia, Taipei, 237 Taiwan
[2] Biomedical Engineering Center, ITRI
Hsinchu, Taiwan
[3] Department of Animal Science
National Taiwan University, Taipei, Taiwan
[4] Divisions of Applied Biology and Biotechnology
Animal Technology Institute Taiwan, P.O. BOX 23, Chunan, Miaoli, Taiwan

Abstract. Spot matching is a challenging problem in two dimensional protein gel electrophoresis (2DGE) chromatography images analysis. It is necessary to provide a robust solution to the problem of querying and matching large scale for various size of 2DGE images. In this paper, we have developed a novel maximum relation spanning tree (MRST) algorithm which is capable of performing fast and accurate matching without the need for landmarks to be manually selected. In the matching process, we employ fuzzy similarity measuring technique to conclude the final decision of matching and location. The resultant system performs up to 94% correct matching for 225 2DGE test images. The additive value is the foundation of querying fractional gel images with large format gel images database.

1 Introduction

In the research of protein expression analysis, two-dimensional gel electrophoresis (2DGE) chromatography is a popular tool for investigating differential patterns of qualitative protein expression [1]. The problems can be categorized into image registration, image distortion correction, spot detection, and spot matching. Two dimensional spot matching of two non-uniform images is an NP-hard problem [2]. Their computation is non-deterministic polynomial time. A few algorithms have been proposed and tried to solve this problem, for example Restriction Landmark Genomic Scanning (RLGS) [3–5] and Fuzzy Cluster [6]. RLGS compares the protein using construction of computer graphs and landmark. Fuzzy Cluster method uses the relation between two protein spots and calculates the similarity. For two gel images with n and m spots, the worst case upper bound of computation complexity is $O(n^2m^2)$ arc pairs and $O(n \log m)$ for measuring the pattern

* This work was supported in part by the National Science Council, Taiwan, R.O.C. grants NSC90-2313-B-059-002, NSC91-2745-E-216-001, and NSC93-2213-E-216-016

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1366–1372, 2005.

similarity of each pair [7]. In this paper, we propose a novel Maximum Relation Spanning Tree (MRST) and integrate fuzzy inference technique to solve the matching problem. We can use this method to find the gel image which contains or is similar to the small and fractional query image, and to locate the area residing in the large scale images. In addition, this method is fully automated and does not need landmark allocated in a priori by users.

2 Features Extraction and Fuzzy Similarity Measure

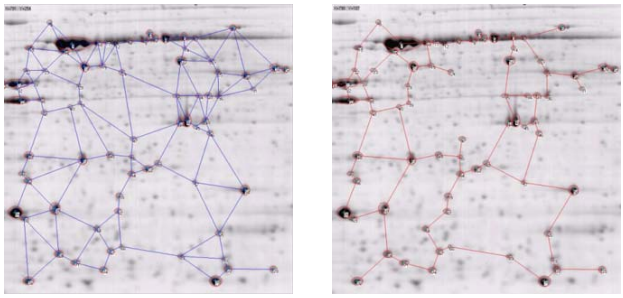
We have to construct features of the protein spots which are invariant to intensity bias and geometric distortions due to the casting, polymerization and running procedure of the gels [7]. In this research, we apply the computer graphics theory to construct and to extract the features from 2DGE images [8–10]. We have selected the Gabriel Graph (GG) [11] and the Relative Neighborhood Graph (RNG) [3] as the feature construction models because the variation of point's feature is more obvious than that of the others. The Gabriel graph P , denoted by $GG(P)$, has its region of influence over the closed disk having segment $\overline{uv}$ as diameter. That is, two vertices $u, v \in S$ are adjacent if and only if

$$D^2(u,v) < D^2(u,w) + D^2(v,w), \ \ for\ all\ w \in V,\ w \neq u,\ v. \tag{1}$$

where $D(u,v)$ denotes the distance of $\overline{uv}$. The relative neighborhood graph of P, denoted by $RNG(P)$, has a segment between points u and v in P if the intersection of the open disks of radius $D(u,v)$ centered at u and v is empty. Equivalently, $u, v \in V$ are adjacent if and only if

$$D(u,v) \leq max[D(u,w), D(v,w)], for\ all\ w \in V, w \neq u, v. \tag{2}$$

Thus, RNG is a subset of a GG and is relatively transformation insensitive compared with its superset [3]. Examples of Gabriel graph and Relative neighborhood graph constructed from one gel image are shown in Fig 1(a) and (b), respectively. Geometrical spot matching relies on the similarity of the features extracted from the structured graphs. After we have constructed the proximity graphs, we continue to extract the features of the spots. For each node on both



(a) Gabriel graph (b) Relative neighborhood graph

Fig. 1. Examples of graph representation for a 2D gel image

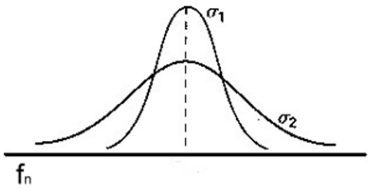


Fig. 2. Illustration of membership function

Gabriel Graph and Relative Neighborhood Graph, three features are obtained: f_1 degree of each protein spot, f_2 angle of connected edges, and f_3 Euclidean distance between protein spots. The conventional direct superimpose matching is not appropriate due to imperfect 2D electrophoresis technique [7]. An adaptive decision method is essential to examine and measure the similarity from the above-mentioned spot pair features. We decide to apply the fuzzy inference to develop our comparative framework. Due to the difference in each local area of the 2DGE spots geometric relation, we use a stylized gaussian membership function as shown in Fig. 2. The nth feature of spot s in the sample gel image is defined as $f_n(s)$, and the corresponding feature of spot r in the reference gel image is defined as $f_n(r)$. Let μ_{f_n} be the fuzzy membership function of the similarity measure between the sample image and reference image according to the nth feature:

$$\mu_{f_n}(s,r) = e^{-\frac{(f_n(s)-f_n(r))^2}{2\sigma^2}}, \tag{3}$$

where σ denotes the variance of the feature f_n between spots. The function is illustrated in Fig 2 in which σ_1 and σ_2 denote different local intensity. In this figure, we can see that different sets of spots will have different kind of membership functions constructed by different σ. With three different features, we calculate three fuzzy relations $\mu_{f_1}, \mu_{f_2}, \mu_{f_3}$ for distance, degree, and angle on the Garbriel graph, respectively. To aggregate three fuzzy measurements, a weighted mean value [12] is computed and defined as the closure measurement:

$$R(s,r,\omega_{f_1},\omega_{f_2},\omega_{f_3}) = \frac{\omega_{f_1}\cdot\mu_{f_1}(s,r) + \omega_{f_2}\cdot\mu_{f_2}(s,r) + \omega_{f_1}\cdot\mu_{f_3}(s,r)}{3}, \tag{4}$$

where w_{f_i} is the weight of the corresponding feature and $\Sigma_{i=1}^{3}\omega_{f_n} = 3$. The weights can be set optimally according to learning mechanism. Finally, we can choose the maximum relationship from the spot pairs and proceed to the next comparison procedure.

3 Fractional Matching with Maximum Relation Spanning Tree

In order to compare the similarity between two gel images, we have developed a maximum relation spanning tree (MRST) algorithm, in which the minimum

distance derived from the Minimum-Cost Spanning Tree [13] is replaced by the proposed maximum fuzzy relation. We calculate the relationship between protein points using the features of the structured graphs and find the maximum relationship of fuzzy inference. The MRST algorithm is elucidated as follows. When we implement this algorithm, we separate the process into two parts: global matching and local matching.

```
Maximum Relation Spanning Tree Algorithm
MRST() {
 If node tree T is empty
   Insert a new anchor point pair with the maximum fuzzy relation;
   MRST();
 else if  anchor point pairs are not empty
          Find next anchor point pair in the satellite spots;
          MRST();
      else terminate;
 Compare the matched area;
 If the difference is less than 10%
   then it is matched
   else match fail;
}
```

3.1 Global Matching

In this step we find the initial anchor point pair automatically by comparing the fuzzy relation of all possible corresponding anchor pairs between the sample gel image and reference gel image. Substituting the feature of degree from the profiles of GG and RNG into the similarity measure of fuzzy membership function (Equation 3), we can obtain two fuzzy relationship values R_{GG} and R_{RNG}, respectively. If the product $R_{GG} \times R_{RNG} \geq 0.7$, then we treat it as a candidate anchor point pairs with larger fuzzy similarity relationship. The candidate pair with the maximum fuzzy relation will be selected with higher priority in the local matching process.

3.2 Local Matching

Once the anchor point is located, we start to apply the maximum relation spanning tree algorithm on the Gabriel graph. The fuzzy similarity measure (Equation 4) of three features (degree, distance, and angle) of the Gabriel graph is computed. If the fuzzy membership is greater than 0.6, the graph is extended. This algorithm will proceed recursively until all the spot pairs produced by Gabriel matching is completed. The flow chart is depicted in Fig. 3. Through this process, we will find all similar spot pairs between two gel images.

4 Simulation Results

We have implemented and tested the proposed system. The experiments were based on fifteen 2D protein gel profiles (image size: 1498 x 1544) of porcine

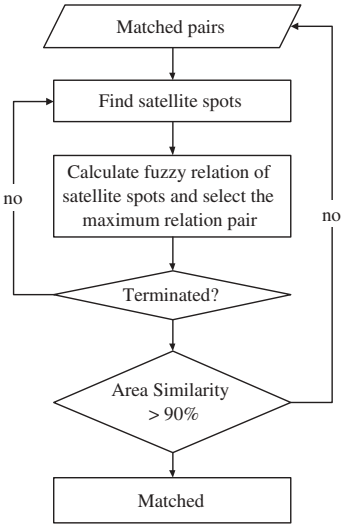


Fig. 3. The flowchart local matching of the protein 2D gel pattern matching process

testis obtained from the Bioinformatics Center for Swine Research in Tropical Area at the Animal Technology Institute Taiwan (ATIT). The experiment data set is constructed from these fifteen gel images. The data set contains totally 225 gel images as follows: 15 original gel images (1498×1544), 135 fractional gel images composed of nine different sizes of fractional images chopped randomly from each of 15 original gel images (listed in Table 1), and 75 rotated images: 45 gel images obtained from the original gel images by rotating in 90°, 180°, and 270° degrees, respectively, and 30 gel images obtained from the original gel images by flipping horizontally and vertically, respectively. We have done the test on fractional matching by using various size of segmented image samples (135 images) to perform query in the original gel images (15 images). The correct match is up to 94%. The results are detailed in Table 1. In order to simulate the situations of image rotation, reverse, and translation, we have tested 75 different modified gel images with five situations mentioned above. The ratio of correct matching is 100%. To further confirm the capability of fractional matching, we have also used the rotated fractional gel images and to perform searching in the original large-scale gel images. One of the results is demonstrated in Fig. 4 where one fractional sample gel image of size of 200×200 is rotated or flipped into five images with different conditions (rotated in 90°, 180°, 270°, flipped horizontally ,and vertically) shown on the left hand side in Fig. 4. With these six small images, we tried to search in the original large 2D gel image database. The location of correct matching is identified in the rectangle on the right-hand side in Fig. 4. Only Pánek and Vohradský reported their matching accuracy is 98% but based on one gel image [14]. To justify the advantage of the proposed work, we need to make comparison with other methods. However, the quantitative information of fractional matching performance is not available from the literature survey.

Table 1. The results of fractional matching of different size of images with adapted parameters

	Correct Matching Ratio
Original Images	100 %
1000x1000	100 %
900x900	100 %
800x800	100 %
700x700	100 %
600x600	100 %
500x500	100 %
400x400	100 %
300x300	86.7 %
200x200	53.3 %
Overall	94 %

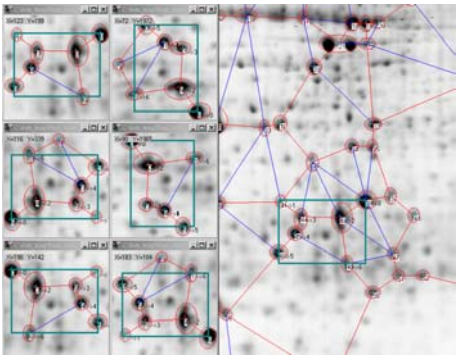


Fig. 4. Result of fractional matching and allocation processing

5 Conclusion

In the research of differential patterns investigation for qualitative protein expression, it is necessary to provide a robust solution to the problem of querying and matching large scale and large sets of protein 2DGE chromatography. In this paper, we have developed a novel, fast, accurate and content-based image matching method MRST utilizing the fuzzy inference technique. We have selected the Gabriel Graph and Relative Neighborhood Graph as the feature construction models. It is expected to compensate the variance of geometric distortions automatically. The proposed method not only can handle the rotation, shift and reverse condition, but can also handle fractional mapping problem. We can use this method to find the gel image which contains or is similar to the small and fractional query image, and to locate the area residing in the large scale images. After all, we can constitute the gel images and protein spots information into the database for further investigation. The proposed system achieves up to 94% correct matching in large-scale gel image searching scenarios. Most importantly, the proposed MRST matching algorithm requires neither the landmarks manually set nor a prior information of gel image alignment.

References

1. M. J. Dunn S. Veeser and G. Z. Yang. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, 1:856–870, 2001.
2. Tatsuya Akutsu, Kyotetsu Kanaya, Akira Ohyama, and Asao Fujiyama. Point matching under non-uniform distortions. *Discrete Appl. Math.*, 127(1), 2003.
3. Y. Watanabe K. Takahashi, M. Nakazawa and A. Konagaya. Fully-automated spot recognition and matching algorithms for 2-D gel electrophoretogram of genomic DNA. In *Genome Informatics Workshop*, pages 161–172, 1998.
4. K. Takahashi, Y. Watanabe M. Nakazawa, and A. Konagaya. Automated processing of 2-D gel electrophoretograms of genomic DNA for hunting pathogenic DNA melecular changes. In *Genome Informatics Workshop*, pages 121–132, 1999.
5. T. Matsuyama, T. Abe, C. H. Bae, Y. Takahashi, R. Kiuchi, T. Nakano, T. Asami, and S. Yoshida. Adaptation of restriction landmark genomic scanning to plant genome analysis. *Plant Molecular Biology Reporter*, 18:331–338, 2000.
6. X. Ye, C.Y. Suen, and E. Wang. M. Cheriet. A recent development in image analysis of electrophoresis gels. In *Vision Interface (VI'99), Trois-Rivieres, CA, 19-21*, pages 432–438, 1999.
7. A.W. Dowsey, M.J. Dunn, and G.Z. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics.*, 3:1567–1596, 2003.
8. A. Efrat, F. Hoffmann, K. Kriegel, and C. Schultz. Geometric algorithms for the analysis of 2D-electrophoresis gels. *Journal of Computational Biology*, 9(2):299–315, 2002.
9. F. Hoffmann, K. Kriegel, and C. Wenk. Matching 2D patterns of protein spots. In *Symposium on Computational Geometry 1998*, pages 231–239, 1998.
10. F. Hoffmann, K. Kriegel, and C. Wenk. An applied point pattern matching problem: comparing 2D patterns of protein spots. In *Discrete Applied Mathematics 93*, pages 75–88, 1999.
11. J. I. Garrels. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *Biological Chemistry*, 254:796l–7977, 1979.
12. G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic - Theory and Appication.* Prentice Hall International Editions, 1995.
13. E. Horowitz, S. Sahni, and D. Mehta. *Fundamentals of Data Structures in C++.* Computer Science Press, 1995.
14. J. Pánek and J. Vohradský. Point pattern matching in the analysis of two-dimensional gel electropherograms. *Electrophoresis*, 20:3483–3491, 1999.