Protein 2D Gel Electrophoresis Images Matching with Maximum Relation Spanning Tree*

Daw-Tung Lin, Juin-Lin Kuo
Department of Computer Science and Information Engineering
Chung-Hua University
30 Tung Shiang, Hsinchu, Taiwan. R.O.C
dalton@chu.edu.tw, juinlin@mail2000.com.tw

En-Chung Lin, San-Yuan Huang
Divisions of Applied Biology and Biotechnology
Animal Technology Institute Taiwan
P.O. BOX 23, Chunan, Miaoli, Taiwan. R.O.C
eclin@mail.atit.org.tw,syhuang@mail.atit.org.tw

Abstract

The electrophoresis chromatography, a popular analysis tool, is able to separate different kinds of protein pro£les. In this paper, we present a robust comparative algorithm Maximum Relation Spanning Tree (MRST) for matching large scale and large sets of two dimensional protein gel electrophoresis (2DGE) chromatography images without the need of a priori landmark. The algorithm not only can handle the conditions of image rotation, shift and reverse, but also can handle fractional mapping problem. In the matching process, we apply fuzzy inference technique to conclude the £nal decision of mapping and location. The proposed system presents up to 94% correct matching performance for 225 2D gel test images. The additive value is the foundation of comparing small gel images to large format gel images and the constitution of searching scheme for a huge two dimensional gel electrophoresis chromatography image database.

1. Introduction

In the research of bio-informatics, proteomic technology plays an important role in protein expression study and analysis. Two dimensional gel electrophoresis (2DGE) chromatography is a popular tool for protein characteristics analysis. 2DGE images analysis is the £rst step of the whole study procedure. Issues of 2DGE images analysis have been discussed in some literatures. The problems can be categorized into the following topics: image registration, image distortion correction, spot detection, and spot matching. Image registration is widely used in biomedical imaging, which includes methods developed for automated image labelling and pathology detection in individuals and groups [1]. Image registration is equally important to biological systems, e.g. in proteomic research, 2DGE images is an important tool for investigating differential patterns of qualitative protein expression [2]. Spot detection is a basic procedure for 2DGE images analysis. We must locate the protein spots coordinates and then record or compare the attributes. There has been many methods proposed for protein spots detection, including Gaussian £tting [3], Laplacian £tting [4], Histogram [5] and Watershed Transformation [6, 7]. The common feature of those algorithms is spatial domain single processing. The aim of the segmentation process is to de£ne the location, true boundary and intensity for each spot. Image registration and spots detection are important procedure in the whole 2D gel image analysis scenario. Image registration and spot detection are common issues and has been studied thoroughly in the past.

Spot matching is a challenging problem in 2D gel comparison. We must £nd the different spots between standard gel and relative gel. A few algorithms have been proposed to solve this problem, for example Restriction Landmark Genomic Scanning (RLGS) [3, 8, 9] and Fuzzy Cluster [10]. RLGS compares the protein based on constructing of computer graphs and landmark. It has the drawback that users need to allocate the landmark manually. Fuzzy Cluster algorithm uses the relation between two protein spots and calculate the similarity. Fuzzy Cluster method also can calculate the similarity fast, but if there have hundreds or thousands protein spots in two gel images, this method will fail without using a large amount of features.

We have also investigated and implemented several methods to construct a framework of 2DGE image analysis system [11]. For spot detection, we choose the Watershed Algorithm to segment the protein spots for the sake of low intensity and process speed, and to adapt to local area

*This work was supported in part by the National Science Council, Taiwan, R.O.C. (NSC91-2745-E-216-001, NSC89-2313-B-059-045, NSC90-2313-B-059-002).

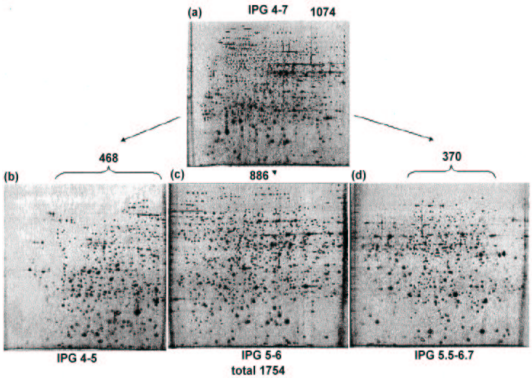

Figure 1. Illustration of gel images combination [12]: (a) IPG 4-7 gel image; (b) IPG 4-5 gel image; (c) IPG 5-6 gel image; (d) IPG 5.5-6.7 gel image.

intensity. We provide a novel Maximum Relation Spanning Tree (MRST) to overcome the problems of RLGS and Fuzzy cluster method. The algorithm not only can handle the rotation, shift and reverse condition, but also can handle fractional mapping problem. In the matching process, we apply fuzzy inference to correct the variation, which is generated from rotation, shift and reverse condition. In addition, this method doesn't need landmark allocated in a priori by users.

2. Features Extraction with Gabriel Graph and Relative Neighborhood Graph

In the early stages of the biological experiment with long range pH gradient, the protein spots were found usually too closed or overlap with each other. This is a major problem for protein purging. Therefore, the biologists would prefer processing with narrow range of pH gradient in the same gel areas separately and then compose several gel sub-images to analyze protein as a whole. As we can see in Fig. 1, the image on the top is composed from three partially overlapped images shown below it. We can use the combined gel image to identify more useful protein spots instead of using the overlapped and blurred protein spot in wide range IPG gel image. Due to this constraint, we provide a novel comparative method Maximum Relative Spanning Tree (MRST) for matching the small part of image with it's original gel image. Thus, we can use this method to £nd the original gel image form small and fractional images and allocate them by using this protocol.

Generally speaking, there is no £x model of protein spots distribution, nor £x size and could reside everywhere in the gel image. We cannot compare gels only by the information of protein spots location or by image intensity. According to the de£cient feature in gel image, we have to construct the unique feature of the protein spots. In this paper, we £rst apply the computer graphics theory to construct and to extract the features for gel images comparison [13, 14, 15, 16]. The proposed MRST algorithm is then utilized to compute the similarity between features with fuzzy inference rules We will discuss these issues in the following sections.

There have been several kind of graph computation methods studied by researchers and scientists such as minimum spanning tree (MST) [17], relative neighborhood graph (RNG) [18], Gabriel graph (GG),Delauney triangular graph (DT) [19] and etc.. These are simple, undirected, straight-lined, connected and planar graphs. Furthermore, these graphs will keep features unchanged no matter with the transformation of rotation, shift and reverse. Therefore, every point in the graph has a set of unique feature from graph viewpoint.
We have selected the Gabriel Graph (GG) and the Relative Neighborhood Graph (RNG) as feature constructive models because the variation of point's feature is more obvious than that of the others. We use two graph models GG and RNG to construct the features and to compare gel images. GG provides the major connection relation. We begin the de£nition of the theoretical and geometric terminologies in the next paragraph.

A graph $G = (V, E)$ consists of a £nite non empty set $V(G)$ of vertices, and a set $E(G)$ of unordered pairs of vertices known as edges. An edge $e \in E(G)$ consisting of vertices u and v and is denoted by $e = \overline{uv}$; u and v are called the endpoints of e and are said to be adjacent vertices or neighbors. The degree of a vertex $v \in V(G)$, denoted by $deg_G(V)$, is the number of edges of $E(G)$ which have v as an endpoint. A path in a graph G is a £nite non-null sequence $P = v_1v_2...v_k$ where the vertices $v_1v_2...v_k$ are distinct and $\overline{v_iv_{i+1}}$ is an edge for each $i = 1, ..., k-1$. The vertices v_1 and v_k are known as the endpoints of the path. A cycle is a path whose endpoints are the same. A graph is connected if, for each pair of vertices $u, v \in V$, there is a path from u to v . Gabriel Graph and Relative Neighborhood Graph are then de£ned in following sections:

2.1 Gabriel Graph

The Gabriel graph P , denoted by $GG(P)$, has its region of in¤uence the closed disk having segment $\overline{uv}$ as diameter. That is, two vertices $u, v \in S$ are adjacent if and only if

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

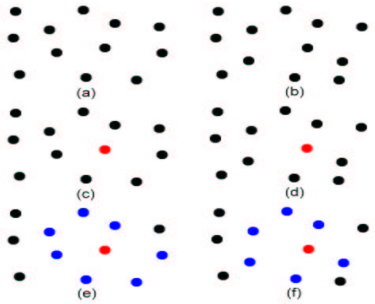

Figure 2. Illustration of major spots and satellite spots.

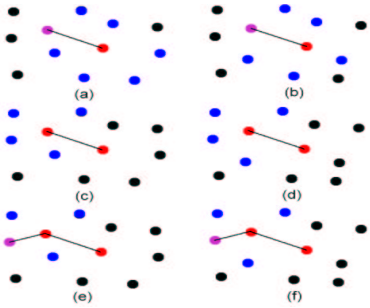

Figure 3. Illustration of maximum relation spanning tree matching pairs.

2.2 Relative Neighborhood Graph

Given a set P of points in R^2 , 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. This region of in¤uence is referred to as the lune of u and v . Equivalently, $u, v \in S$ are adjacent if and only if

$$D(u, v) \leq max[D(u, w), D(v, w)] \\ , for\ all\ w \in S, w \neq u, v. \quad (2)$$

2.3 Feature Extraction

After we have constructed the proximity graphs, we will continue to extract the features of the spots on protein gel images. In the above proximity graphs, we can obtain the following features:

- Degree of each protein spots,
- Angle of connected edges, and
- Distance between protein spots.

These features will not be changed in the circumstance of shift, rotation and reverse due to the advantage of graph properties. We will calculate the features for every protein spot based on these information and utilize them as a framework of spot matching comparison in the next section.

3 Fractional Spot Matching with Maximum Relation Spanning Tree

In order to compare the similarity between two gel images, we have developed the maximum relation spanning tree (MRST) a comparative framework for this task, as in which the minimum distance derived from the minimum spanning tree is replace by the maximum relation. We will calculate the relationship points as their features and £nd the maximum relation protein spot pair as basic information for image matching. The algorithm will be terminated if there is no referable pair in the spot pair sets. The illustration of the MRST algorithm is shown in Fig 2 and Fig. 3 and is discussed as follows.

The MRST algorithm is based on Gabriel graph. We use the connected condition of Gabriel graph to de£ne whether a spot can join and to be compared as shown in Fig. 2. Figure 2(a) and (b) denote two protein spots sets, (c) and (d) denote the initial comparative pair (plotted in red dots) found by global matching, (e) and (f) denoted the satellite spots (plotted in blue dots) found by Gabriel graph connected condition. We name the red spot and the blue spots as the major spot satellite spots, respectively. Starting from any two major spots in the standard gel image (target) and relative gel image (test sample), we will travel through all of the satellite spots in the standard gel image and calculate the relationship with all of the satellite spots in relative gel image. As illustrated in Fig. 3 with connected edges, we will £nd the maximum relation for every satellite spot pairs recursively. Figure 3(a) and (b) denotes the maximum relation matching pair between two images to be computed, (c) and (d) shows that the satellite spots are used as the center point and to £nd the next pair, (e) and (f) illustrates that recursive processing and £nd the next pair. From these pairs, we chose the maximum relation pair to be the major spot pair for comparison and store the rest of high relation pairs into the temporary stack. If we cannot £nd further major spot pair, choose the the largest relation pair in the temporary stack and process continuously. When all of the spots in the standard gel image have been traversed, the computation will be terminated and the matching results will be obtained. When we implement this algorithm, we separate the process into two parts: global matching and local matching.

3.1 Global matching

In this step, we need to £nd the initial comparative pair. We compare all possible pairs and £nd the maximum relation between each other. Euclidean distance of three different features (spot degree, angle, and distance as mentioned in Section 2.3) is computed as the similarity measure between spot pairs and then sort them by their relation values, then recorded into the temporary stack.

3.2 Local matching

After the searching of the similar pairs has been completed, we start to process the maximum relation spanning tree. The algorithm is elucidated as following:

$MRST()$
{
$If\ node\ tree\ is\ null.$
    $Inserting\ new\ comparative\ pair.$
    $MRST()$
$else\ if\ comparative\ pairs\ is\ not\ empty$
    $Find\ next\ comparative\ pair\ in\ the\ satellite\ spots.$
    $MRST()$
$else\ if\ temporary\ stack\ is\ empty$
    $Function\ will\ be\ terminated.$
}

3.3 Fuzzy Inference

Direct superimpose matching is not appropriate due to the imperfect 2D electrophoresis technique. 2D gel patterns usually present various kinds of transformations such as distortion, translation and the variation of protein properties. An adaptive decision method is essential to examine and conclude the similarity measure from the above-mentioned spot pair features. We decide to applying the fuzzy inference to develop our comparative framework. By transferring the crisp relation into fuzzy relation. Using fuzzy relation to compare satellite spots is more suitable.

Due to the difference in every local area, we use a stylized membership function. The feature of the spots in the standard gel image is de£ned as f^s and the feature of the spots in relative gel image is de£ned as f^r. Let the function be:

$$R(f_n^{S_a \to R_b}) = e^{-[\frac{(f_n^{S_a} - f_n^{R_b})}{2\sigma^2}]^2} \quad (3)$$

where R denotes the membership function and σ denotes the variance of the feature f_n between spots. The MRST algorithm utilizes this fuzzy inference system to compute the relations of every spots whether it is the neighbor of the center spot. With three different features, we calculate three different relations for these features and obtain the weighted average value [21]. The total relation R_{total} is de£ned as:

$$R_{total} = \frac{\omega_{f_1} \cdot R_{f_1} + \omega_{f_2} \cdot R_{f_2} + \omega_{f_1} \cdot R_{f_3}}{3} \quad (4)$$

w_{f_i} is the weight of the corresponding feature. Finally, we can choose the maximum relationship from the spot pairs and proceed to the next comparison procedure. This algorithm will be process recursively until all of the spot pairs produced by the Gabriel matching is completed. Through this process, we will £nd all similar spot pairs between two gel images. We will also label the matched spots corresponding to their relative spots and their matching area. As shown in Fig. 4, the left window displays the source image to be matched, the square on the right-hand side indicates the matched area in a large scale gel image. If no complete fractional match were available, this algorithm could present the possible match point pairs with corresponding matching labels as shown in Fig. 5. As we can observe in Fig 5, each spot in the rectangle area has two labels indicating the mapping relation between the standard (target) pattern and the corresponding spot label of the test pattern. The left label indicates the spot number in the target gel image. The right label followed by arrow sign denotes the matched spot label in the source (test) image.

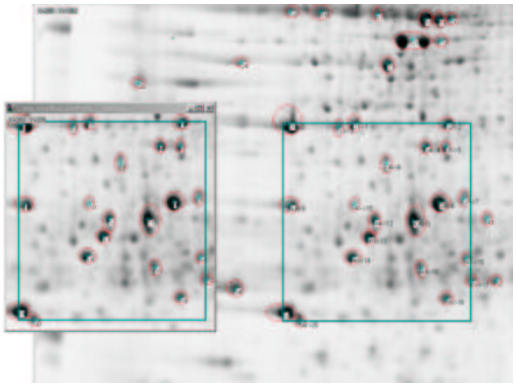

Figure 4. Result of fractional matching.

4 Simulation Results

We have implemented the proposed 2DGE images analysis system, and demonstrated the results of gel matching (global matching and fractional matching) in this section. We have obtained 15 gel images (1498 x 1544) from the Animal Technology Institute Taiwan (ATIT) as the test images. We utilized them to construct the experiment data set. The data set contains totally 225 gel images as illustrated in the following: 15 original gel images (1498 x 1544), 135 fractional gel images composed from

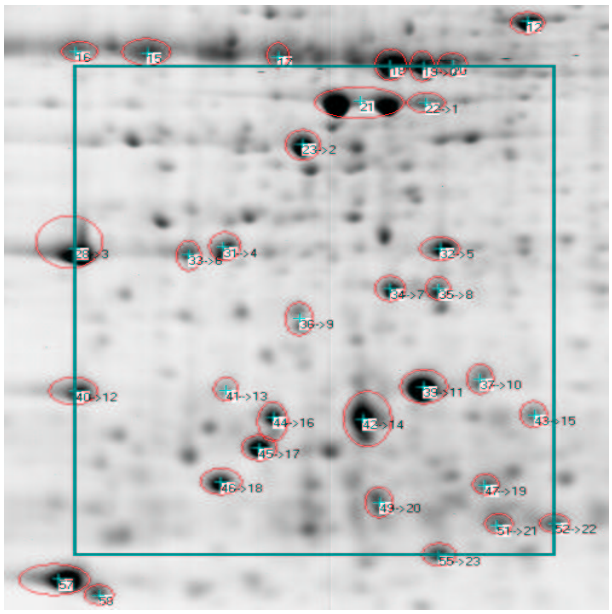

Figure 5. Relation of pattern mapping between two similar fractional gel images.

different sizes of fractional images (1000x1000, 900x900, 800x800,700x700, 600x600, 500x500, 400x400, 300x300 and 200x200) chopped randomly from each of 15 original gel images, and 75 rotated images: (1) 45 gel images obtained from the original gel images by rotating in $90°$, $180°$, and $270°$ degrees, respectively, and (2) 30 gel images obtained from the original gel images by ¤ipping horizontally and vertically, respectively. We have done two categories of experiments by using the data set as following:

- Fractional Matching: using various size of fractional image samples (135 images) to perform searching in the original gel images (15 images) and see if the proposed method can £nd the correct match.
- Spots Registration: using the rotated samples (75 images) to match the original gel images (15 images) and see if the proposed algorithm can £nd the correct registration without rotating the samples in *a priori*.

In order to obtain better performance, we tried to adjust two parameters - level of 'opening' and protein size threshold through extensive computer simulations. The 'opening' parameter ranges from 1 to 30, and the protein size threshold is between 0 and 1000. If the gel protein spots intensity in gel image is blurred, we must adjust the parameters to be low enough to retain the small protein spots for matching. On the other hand, if the image background is over-stained, we need to adjust the parameters as high as possible in order to delete the noise. The performance of fractional matching is raised to 94% after the adjustment. The detail results is shown in Table 1. In order to simulate the situations of input images rotation, reverse, and translation, we have tested 75 different modi£ed gel images with 5 situations as shown in Table 2 . The ratio of correct matching is about 80%.

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 %

Table 2. The result of registration for different rotation situations.

	Correct Registration Ratio
Rotated $90°$	80 %
Rotated $180°$	80 %
Rotated $270°$	80 %
Horizontal Reversal	86.7 %
Vertical Reversal	73.3 %
Overall	80 %

To further con£rm 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 (15 images) and see if the proposed method can £nd the correct match. One of the results is demonstrated in Fig. 6 where one fractional standard gel image of size of 200x200 is rotated or ¤ipped into £ve images with different conditions (rotated in $90°$, $180°$, $270°$, ¤ip horizontally ,and vertically) and these fractional images are shown on the left hand side in Fig. 6. By applying these six small images, we tried to search in the original large size 2D gel images and see if the proposed system can £nd exact match. The location of correct matching is identi£ed in the rectangle on the right hand side of Fig. 6.

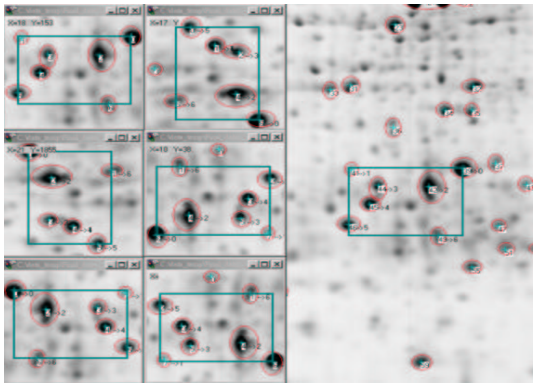

Figure 6. Result of fractional matching and allocation processing.

5 Conclusion

In this paper, we focused on the issues of fractional matching and relocation for the small fraction pattern in the large scale of 2D gel images. We have developed a fast, accurate and content-based image matching method Maximum Relation Spanning Tree (MRST). By using this algorithm, we can easily allocate the protein spots for the fractional images or warp images, and match the original gel images. After all, we can constitute the gel images and protein spots information into the database for further investigation. The proposed system archives up to 94% correct matching in large scale gel image searching scenarios. Most importantly, the proposed MRST matching algorithm does not require neither the landmarks manually set nor a prior information of gel image alignment.

References

[1] X. Y. Wang, D. D. Feng and H. Hong. "Novel Elastic Registration for 2-D Medical and Gel Protein Images". *In Proc. First Asia-Paci£c Bioinformatics Conference (APBC2003)*, Adelaide, Australia. Conferences in Research and Practice in Information Technology, 19. Chen, Y.-P. P., Ed. ACS. 223-226. .

[2] S. Veeser, M. J. Dunn and G. Z. Yang. "Multiresolution image registration for two-dimensional gel electrophoresis". *Proteomics 2001*, 1, 856-870.

[3] K. Takahashi, M. Nakazawa, Y. Watanabe and A. Konagaya. "Fully-Automated Spot Recognition and Matching Algorithms for 2-D Gel Electrophoretogram of Genomic DNA". *In Proc. of Genome Informatics Workshop*, pp.161-172, 1998. Dec.

[4] "Z3". http://www.compugen.co.il/.

[5] P. Culter, G. Heald, I. R. White and J. Ruan. "A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection". *Proteomics 2003*, 3, 392-401.

[6] L. Vincent and P. Soille. "Watersheds in Digital Spaces: An Ef£cient Algorithm Based on Immersion Simulations". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, June 1991, Vol. 13, No. 6, pp. 583-598.

[7] S.Beucher. "The Watershed Transformation Applied to Image Segmentation". *Cambridge, UK, Scanning Microscopy International*, suppl. 6. 1992, pp. 299-314.

[8] K. Takahashi, M. Nakazawa, Y. Watanabe and A. Konagaya. "Automated Processing of 2-D Gel Electrophoretograms of Genomic DNA for Hunting Pathogenic DNA Melecular Changes". *In Proc. of Genome Informatics Workshop 1999*, pp121-132.

[9] T. Matsuyama, T. Abe, C. H. Bae, Y. Takahashi, R. Kiuchi, T. Nakano, T. Asami and S. Yoshida. "Adaptation of Restriction Landmark Genomic Scanning (RGLS) to Plant Genome Analysis". *Plant Molecular Biology Reporter 18*, 2000, 331-338.

[10] X. Ye, C.Y. Suen, M. Cheriet, E. Wang. "A Recent Development in Image Analysis of Electrophoresis Gels" *Vision Interface (VI'99)*, Trois-Rivieres, CA, 19-21 May 1999, pp. 432-438.

[11] J. L. Kuo, D. T. Lin, E. C. Lin and S. Y. Huang. "Image Analysis System for Protein Two Dimensional Gel Electrophoresis" *16th IPPR Conference on Computer Vision, Graphics, and Image Processing* (CVGIP 2003), pp. 139-146.

[12] H. J. Issaq, T. P. Conrads, G. M. Janini and T. D. Veenstra. "Methods for fractionation, separation and pro£ling of proteins and peptides". *Electrophoresis 2002*, 23, 3048-3061.

[13] A. Efrat, F. Hoffmann, K. Kriegel and C. Schultz. "Geometric Algorithms for the Analysis of 2D-Electrophoresis Gels". *Journal of Computational Biology*, 2002,9(2): 299-315.

[14] F. Hoffmann, K. Kriegel and C. Wenk. "Matching 2D Patterns of Protein Spots". *Symposium on Computational Geometry 1998*: 231-239.

[15] F. Hoffmann, K. Kriegel and C. Wenk. "An applied point pattern matching problem: comparing 2D patterns of protein spots". *Discrete Applied Mathematics 93 (1999)*, 75-88.

[16] Y. Watanabe, K. Takahashi and M. Nakazawa. "Automated Detection and Matching of Spots in Autoradiogram Images of Two-Dimensional Electrophoresis for High-speed Genome Scanning". *ICIP (3) 1997*: 496-499.

[17] D. R. Karger, P. N. Klein and R. E. Tarjan. "A Randomized Linear-Time Algorithm to Find Minimum Spanning Trees". *Journal of the ACM*, March 1995, 42(2):321-328.

[18] J. W. Jaromczyk and G. T. Toussaint. "Relative Neighborhood Graphs and Their Relatives". *Proc. IEEE* (1992), 80 (9):1502X1517.

[19] S. Fortune. "Voronoi Diagrams and Delauney Triangulations". *World Scienti£c, Singapore*, 2nd edition, 1995. Computing in Euclidean Geometry, volume 4 of Lecture notes series on Computing, pages 225–265.

[20] J. W. Jaromczyk and G. T. Toussaint. "Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress". *Technical Report, School of Computer Science, McGill University, Montreal, Canada*, June 2002, SOCS-02.5.

[21] G. J. Klir and B. Yuan. "FUZZY SETS AND FUZZY LOGIC - THEORY AND APPLICATION". *Prentice Hall International Editions*, 1995.