Fully-Automated Spot Recognition and Matching Algorithms for 2-D Gel Electrophoretogram of Genomic DNA

Katsutoshi Takahashi [1]
sltaka@jaist.ac.jp

Masayuki Nakazawa [2]
nakazawa@ishikawa-pc.ac.jp

Yasuo Watanabe [3]
watanabe@infor.kanazawa-it.ac.jp

Akihiko Konagaya [1]
kona@jaist.ac.jp

[1] Japan Advanced Institute of Science and Technology (JAIST) Hokuriku
Asahidai 1, Tatsunokuchi, Ishikawa 923-1211, Japan
[2] Ishikawa Polytechnic College, Yuigaoka I 45-1, Anamizu, Ishikawa 927-0024, Japan
[3] Kanazawa Institute of Technology, Ohgigaoka 7-1, Nonoichi, Ishikawa 921-8501, Japan

Abstract

We have developed the fully-automated algorithms for processing 2-D gel electrophoretograms based on RLGS (restriction landmark genomic scanning) method; one for fully-automated spot recognition from RLGS electrophoretogram and another for fully-automated pairwise matching of the spots found on such 2-D electrophoretograms. Without any human interaction, several thousands of spots on a 2-D electrophoretogram, including hidden spots found at the shoulder of large spots, can be identified correctly by applying our spot recognition algorithm, except for only a few true-negative and false-positive spots. Once the locations and intensities of the landmark spots are correctly recognized automatically, our pairwise spot matching algorithm reliably and rapidly identifies equivalent pairs of spots found on the nonlinearly distorted RLGS electrophoretograms in the fully-automatic way, i.e., the boring and annoying spot landmarking process is unnecessary. At the beginning of the spot matching process, most suitable pair of corresponding spots is searched automatically, then the other equivalent pairs of spots are identified. With our powerful image processing algorithms, it is possible to detect DNA molecular changes such as deletions, additions, amplifications or DNA methylations occurring at or near to the restriction enzyme cleavage sites by means of comparing large amount of RLGS electrophoretograms, without any visual inspection and human interaction.

1 Introduction

The sequencing projects and cDNA projects have been producing large sequence data during their process. As a consequence, several complete sequences of microbial genomes have been recently established and the complete sequence of human genome is predicted to be solved by early in the next century. The next goal of human genome project is the identification of all of the 10^5 genes coded on the sequence and understanding their function, based on the high-quality human genomic sequence data and massive EST (expressed sequence tag) data.

With the complete sequence data, the existence of genes (or protein coding regions) on the sequence can be predicted and identified by listing up all Open Reading Frames (ORFs) followed by the sequence similarity search, matching with motif dictionary and so forth. It is, however, rather difficult to evaluate the function of the newly identified genes, when it shows no strong similarity with the gene which function is already known.

In order to explore the functions of the new genes, several experimental techniques are available. DNA chip technology [8, 21, 26, 28] is the one of the most powerful techniques to detect individual genetic differences, genetic mutations or to investigate the tissue specific expression patterns of the genes. Positional cloning [2, 6] and Positional candidate [7] methods are also the powerful experimental strategy to identify genes or gene products which correspond to specific phenotype. Southern hybridization based method [5, 20] and PCR based method [16] have been developed and applied to identify pathogenic genes and gene products which cause genetic diseases.

Restriction Landmark Genomic Scanning (RLGS) method [11, 12] is the high-speed multiplex technique alternative to the Southern hybridization-based or the PCR-based techniques which detect mutations occurring at the particular regions of genomic DNA molecules. By applying RLGS method, mutations occurring at or near to the restriction landmarks can be scanned throughout the whole genome of any organism. It has been effectively applied to various lines of genetic analyses such as map-based identification of the important genes [12, 13] or to detect genetic variations [1].

In the RLGS procedure, purified genomic DNA is cleaved with a landmark restriction enzyme (A) such as *Not*I. The cleavage sites are end-labeled with radioisotope and the labeled DNA fragments are further digested by restriction enzyme (B) such as *Eco*RV, which recognition site appears more frequently in the genome than of restriction enzyme A. The mixture of the DNA fragments are then subjected to high-resolutional two-dimensional (2-D) electrophoresis. After the first-dimensional agarose gel electrophoresis, the DNA fragments are in-gel digested with a restriction enzyme (C) such as *Hinf*I, then fractionated by second-dimensional polyacrylamide gel electrophoresis. The autoradiography of the dried up gel yields RLGS electrophoretogram (hereafter referred to as RLGS profile) on which several thousands of landmark spots appear. When appropriate restriction enzyme is chosen as a landmark enzyme, genetic alteration occurring at the regions which have important role in gene regulation, such as famous CpG island [4], can be detected effectively. Furthermore, it also offers an analysis of differential methylation of CpG-rich landmark sites associated with genomic imprinting and the changes in CpG methylation of specific loci [15], if methylation-sensitive landmark enzymes such as *Not*I are employed.

Despite the wide applicability and usefulness of RLGS-based genetic study, as mentioned above, the laborious image analysis of RLGS profiles makes the high-speed multiplex genomic scanning method impractical. In RLGS profiles, some irregular features in spot shape can be observed very frequently. Some landmark spots have the long-tailed shape, and other spots show flat shape caused by signal saturation, in addition to the strongly drifted un-uniform background patterns. Because of such irregularities, spot recognition by means of neither visual inspection nor conventional computer algorithms [10, 22] can be employed. Furthermore, in principle, the RLGS profiles take non-linear distortions, i.e., the whole patterns of the RLGS profiles do not coincide even if the profiles are derived from exactly the same genomic DNA. Therefore it is also difficult to match the equivalent spots on the related RLGS profiles by means of the simple overlay technique or conventional image matching algorithms [3, 9, 17, 22].

2 Previous Work

To overcome the difficulties in the analysis of RLGS profiles, we have developed the automated computer algorithms for spot recognition and spot pattern matching [27], and implemented as the RLGS image processing system so-called **DNAinsight** [23]. By applying the spot recognition algorithm, the locations and intensities of several thousands of spots are correctly recognized in spite of the irregularities in spot shape, since ring operator was adopted as the spot detector [23, 27]. Though the algorithm was highly efficient and fast, several tens of false-positive (ill-recognized) spots and true-negative (un-recognized) spots were still remained. Therefore, with this image processing system, such ill-recognized and un-recognized spots should be revised by hand with the graphical user interface provided. This spot revising process was the most time-consuming during the processing of RLGS profiles. Moreover, in principle, no 'hidden' spots, which could be found at the shoulder of neighboring large spots, can be recognized at all, because such hidden spot does not show any peak at its location. Such hidden spots should also be identified in order to make the subsequent spot matching process being accurate.

With **DNAinsight** the equivalent pairs of landmark spots can be identified automatically from two related RLGS profiles. In this computer system, pairwise matching problem of two strongly disordered RLGS profiles was treated as the pattern matching problem of two structured graphs, that is, the Delaunay net and the relative neighborhood graph [24] (hereafter referred to as DN and RNG, respectively,) and solved as the breadth-first search of corresponding arcs in the two structured graphs. The algorithm was very fast and accurately identified the equivalent pairs of landmark spots [23]. Despite the automatic spot matching feature of the algorithm, the initial equivalent pair of spots should be chosen from each RLGS profiles by visual inspection with the graphical user interface **DNAinsight** system provides. This is also the boring, annoying and time-consuming process when large amount of RLGS profiles are being treated simultaneously.

In order to realize the truly automated and objective processing and comparison of RLGS profiles by computer, the above time-consuming, subjective and interactive procedures should be eliminated entirely. Hence we have developed the fully-automated algorithms for spot recognition and spot pattern matching. The improved algorithms are described in this paper in detail.

3 Fully-Automated Algorithms

As described above, the important points to improve our previous work and to accomplish the truly automated and objective processing of RLGS profile are **(i)** automated removal of the spots ill-recognized, and **(ii)** automated detection of the hidden spots and un-recognized spots. Our new approach to overcome these problems is the so-called "Gaussian modeling" of the landmark spots. Once the possible spot location is detected by means of the elliptic ring operator [23, 27], each spot on a RLGS profile is fitted with Gaussian-type function. The spots failed to be fitted by any Gaussian-type functions are treated as the false-positive spots and will be taken away from the list of the valid spot. The existence of hidden spot or un-recognized spot is identified based on the difference between the given RLGS profile and the Gaussian-modeled profile.

In order to realize completely automatic comparison of RLGS profiles, the initial equivalent pair of landmark spots should be selected automatically, prior to applying our previously reported graph matching algorithm [23, 27]. This can be achieved by the heuristic search of the possible combination of the spots in two RLGS profiles. With considering the featured point pattern represented by the RLGS spot location and intensity (hereafter referred to as RLGS pattern) as a structured graph, fitness of each pair of nodes in both graphs is examined by some fitness function, which accounts for the similarity of the central node and all nodes connected to by any arcs in the graph. Additionally in our new spot pattern matching algorithm, we employed the DN as the guide for matching any spots in two RLGS profiles, though the RNG was used in our previous spot pattern matching algorithm mainly because of the computational time requirement.

The fully-automated algorithms to identify landmark spots from a RLGS profile and to compare two RLGS patterns are described in the following subsections.

3.1 Fully-Automated Spot Recognition

The fully-automated recognition of spot locations and their intensities consists of the following steps:

STEP1: Preprocessing for image enhancement and smoothing. Let the preprocessed image be $\phi(x, y)$.

STEP2: Applying background normalization operation [22] to $\phi(x, y)$. Let the resultant image be $\psi(x, y)$.

STEP3: The overall background level t of $\psi(x, y)$ was determined by means of conventional smoothed density histogram method. The binary image $f(x, y)$ of $\psi(x, y)$, which gives the possible spot domains, was obtained as

$$f(x,y) = \left\{ \begin{array}{ll} 1, & \text{if } \psi(x,y) \geq t \\ 0, & \text{otherwise} \end{array} \right. .$$

STEP4: Applying a ring operator to $\phi(x,y)$ on the domain $\{(x,y)|f(x,y)=1\}$ to detect the local maxima independently of background density. Suppose that

$$C(x,y) = \{(u,v)|(u-x)^2 + (v-y)^2/\alpha^2 \leq r_M^2\}$$

and

$$R(x,y) = \{(u,v)|r_m^2 \leq (u-x)^2 + (v-y)^2/\alpha^2 \leq r_M^2\}$$

where α is the ratio of the minor axis to major axis for an ellipse. Then the output of a ring operator is defined by

$$h(x,y) = \max_C \phi(x,y) - \max_R \phi(x,y).$$

Here, the elliptic ring operator was adopted in order to allow efficient detection of flat shaped spots which have long tails in the first dimensional electrophoresis axis x.

The recognized spots are labeled for their identification. Let the recognized location of the spot i be $({s^x}_i, {s^y}_i)$.

STEP5: Spot domain identification. Gray levels of the preprocessed image $\phi(x,y)$ are carefully analyzed and each pixel in the image is classified into one of the spot domains with the following steps:

STEP5-1: Labeling pixels in the preprocessed image $\phi(x,y)$ whose gray levels are the highest through the image. The labeled pixels are either local maxima or in the regions of local flatness inside the flat shaped spots. Each labeled pixel is then classified into the spot domain D_i, which consists of adjoining pixels. Here, pixels on the domain $\{(x,y)|\psi(x,y) \leq t\}$ are ignored.

STEP5-2: A second label marks unlabeled pixels whose gray levels are the highest among all of the unlabeled pixels. The second labeled pixels are then classified into their adjoining spot domains which appeared in the previous steps, with enlarging each spot domain. This classification is continued until no more second labeled pixels adjoin any spot domains. Then, the residual unclassified pixels are offered to define new spot domains.

STEP5-3: The above step is repeated until any unlabeled pixels cannot be found on the image.

Here, let the spot domain corresponding to the spot i be D_i.

STEP6: Gaussian-type function centering at $({s^x}_i, {s^y}_i)$,

$$g_i(x,y) = a_i \exp\{-(x-{s^x}_i)^2/2{S^x}_i - (y-{s^y}_i)^2/2{S^y}_i\},$$

is fitted onto the background normalized image $\psi(x,y)$. The fitting parameters a_i, ${S^x}_i$ and ${S^y}_i$ are adjusted so as to minimize the error function

$$E_i = \sum_{(x,y)\in D_i} \Big[\log\{\psi(x,y)\} - \log\{g_i(x,y)\}\Big]^2.$$

STEP7: The fitted Gaussian-type functions are subtracted from the background normalized image $\psi(x,y)$. The residual image $r(x,y)$ is defined as

$$r(x,y) = \psi(x,y) - \sum_i g_i(x,y).$$

STEP8: Spot detection from the residual image $r(x,y)$. The ring operator is again applied to $r(x,y)$ to detect the additional hidden spots.

STEP9: Gaussian-type function centering at each newly recognized spot is fitted onto the residual image $r(x,y)$, to give the each set of fitting parameters a_i, ${S^x}_i$ and ${S^y}_i$.

STEP10: All of the fitting parameters a_i, ${S^x}_i$ and ${S^y}_i$ are refined so as to minimize the following error function,

$$E = \sum_{(x,y)} \{\psi(x,y) - \sum_i g_i(x,y)\}^2.$$

After this refinement process, integrated intensity of the spot i, I_i, is calculated as

$$I_i = \sum_{(x,y)} a_i \exp\{-(x - {s^x}_i)^2/2{S^x}_i - (y - {s^y}_i)^2/2{S^y}_i\}.$$

The flow of this algorithm is summarized in Fig. 1 with typical intermediate images which occur during the processing of a RLGS profile.

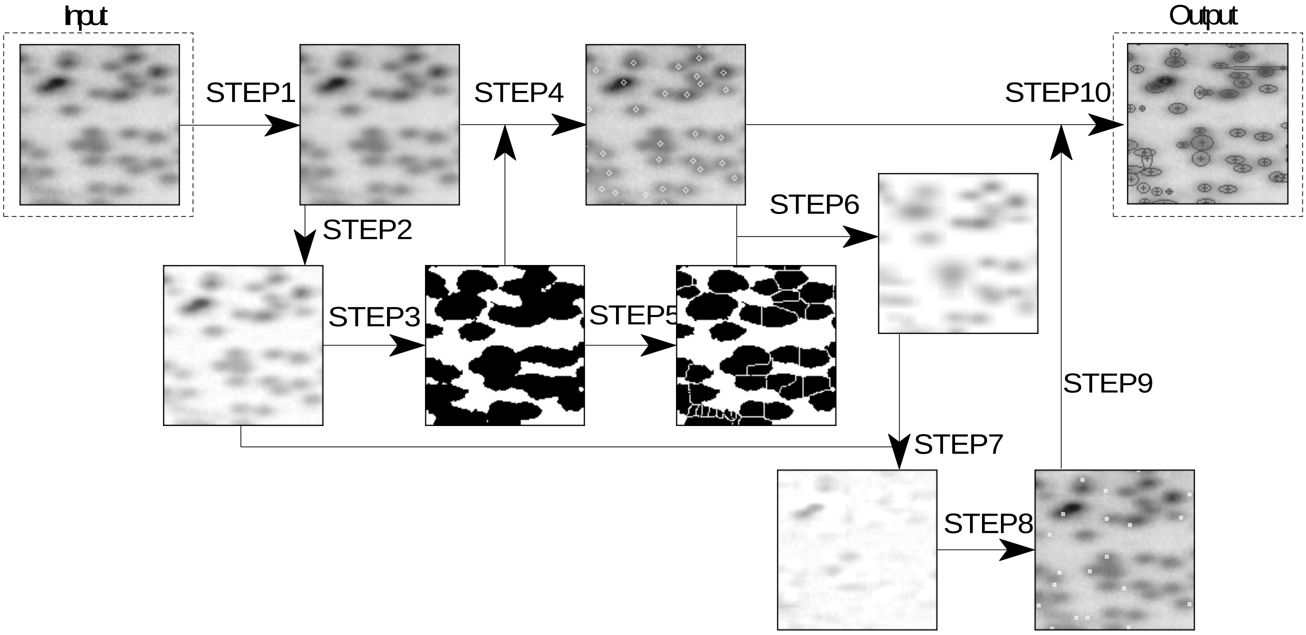


Figure 1: The flow of the fully-automated spot recognition algorithm is summarized with the typical intermediate images; **STEP1**: preprocessing; **STEP2**: background normalization; **STEP3**: binarization; **STEP4**: spot detection by applying ring operator; **STEP5**: spot domain identification; **STEP6**: fitting of single Gaussian-type function onto each spot; **STEP7**: subtraction of each fitted Gaussian-type function from background normalized image; **STEP8**: hidden spot detection from the residual image; **STEP9**: fitting of single Gaussian-type function onto each hidden spot detected; **STEP10**: non-linear least-square fitting of linear combination of Gaussian-type functions onto the background normalized image. For details, see text.

In the above algorithm, steps 1 to 5 are almost identical to our previously reported spot recognition algorithm [23, 27], which identifies the location of the landmark spots in spite of the strongly irregular spot shape and the strongly drifted un-uniform background pattern. The subsequent steps 6 to 10, newly introduced in this paper, realize the automatic ill-recognized spot removal and hidden spot identification. In addition to such functions, integrated intensity of landmark spot can be estimated accurately even if the spot has irregular shape or overlaps with other spots.

3.2 Fully-Automated Spot Matching

To compare two RLGS profiles and to identify equivalent pairs of spots from both RLGS patterns, we first represent the RLGS pattern as the structured graph, that is, the Delaunay net (DN) [24]. Here, we employed the DN because the Delaunay triangulation is known as locally equiangular [18] and the DN is the supergraph [25] of the Gabriel graph, the relative neighborhood graph and the minimal spanning tree, which have been used to compare protein 2-D electrophoresis gels [19] or used in the field of genetic taxonomy. The featured point pattern matching problem is, then, treated as the graph matching problem. To solve the problem, we adopted an algorithm in which the DN of the reference RLGS pattern is used as a guide for matching any nodes in the DN of object RLGS pattern against those in the reference DN.

With initial points in both reference and object graphs given, the graph matching starts with these initial points and progresses in breadth-first manner, minimizing the following evaluation function for the depth k.

$$E_k = \sum_i \Big[\{1 - S(\boldsymbol{a}_i^{(k)}, F(\boldsymbol{a}_i^{(k)}))\}^2 + \{1 - S(\boldsymbol{P}_r(\boldsymbol{a}_i^{(k)}), \boldsymbol{P}_o(F(\boldsymbol{a}_i^{(k)})))\}^2\Big],$$

where for the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$S(\boldsymbol{u}, \boldsymbol{v}) = \frac{(\boldsymbol{u}, \boldsymbol{v})}{|\boldsymbol{u}||\boldsymbol{v}|},$$

$\boldsymbol{a}_i^{(k)}$ is the i-th vector corresponding to the directed arc to be traversed, and F is the mapping from arcs of the reference DN to those of the object DN. In addition, $\boldsymbol{P}_r(\boldsymbol{x})$ is the vector representing a referenced subimage centering at the destination point of $\boldsymbol{x}$ and $\boldsymbol{P}_o(\boldsymbol{x})$ means the similar vector with respect to the object subimage. Here, the mapping F which minimizes the above evaluation function E_k gives the optimal spot correspondence in the reference and object RLGS patterns.

Besides, our algorithm eliminates redundant paths during the traverse, by using thresholds of the difference in the corresponding arcs, i.e,

$$\begin{aligned}
\tau_a &\geq \cos^{-1} S\{\boldsymbol{a}_i, F(\boldsymbol{a}_i)\}, \\
\tau_l &\geq \Big||\boldsymbol{a}_i| - |F(\boldsymbol{a}_i)|\Big|, \\
\tau_e &\geq \{1 - S(\boldsymbol{a}_i^{(k)}, F(\boldsymbol{a}_i^{(k)}))\}^2 + \{1 - S(\boldsymbol{P}_r(\boldsymbol{a}_i^{(k)}), \boldsymbol{P}_o(F(\boldsymbol{a}_i^{(k)})))\}^2.
\end{aligned}$$

Here, we should choose the respective values as the thresholds, taking statistics of a variety of patterns.

Even with the above automatic spot pattern matching algorithm, the initial equivalent pair of landmark spots, from which the breadth-first graph search begins, should still be chosen from each spot pattern. To accomplish the completely automatic spot pattern matching, an automated initial equivalent spot search algorithm is newly introduced in this paper. In our algorithm, the initial points in both reference and object DNs are heuristically searched by using the fitness function defined as follows.

Let F_l^m be mapping from arcs connected to the l-th node in the reference DN to those connected to the m-th node in the object DN, which minimizes the function E,

$$E = \sum_i \Big[\{1 - S(\boldsymbol{a}_i^l, F_l^m(\boldsymbol{a}_i^l))\}^2 + \{1 - S(\boldsymbol{P}_r(\boldsymbol{a}_i^l), \boldsymbol{P}_o(F_l^m(\boldsymbol{a}_i^l)))\}^2\Big],$$

where $\boldsymbol{a}_i^l$ is the i-th vector corresponding to the directed arc connected to the l-th node in the reference DN. The fitness function $f(l, m)$ which accounts for the equivalence of the l-th node in the reference DN and m-th node in the object DN is defined as

$$f(l, m) = \sum_i \{e_i(l, m)\}^{-1},$$

where

$$e_i(l,m) = \begin{cases} \infty, \text{if } i\text{-th arc in the reference DN have no corresponding arc in the object DN} \\ \{1 - S(\boldsymbol{a}_i^l, F_l^m(\boldsymbol{a}_i^l))\}^2 + \{1 - S(\boldsymbol{P}_r(\boldsymbol{a}_i^l), \boldsymbol{P}_o(F_l^m(\boldsymbol{a}_i^l)))\}^2, \text{otherwise} \end{cases} .$$

All possible pairs of the nodes in the reference and object DNs are examined with this fitness function, then the node pair which shows the highest similarity is chosen as the initial equivalent pair of spots. From this initial pair the breadth-first graph search begins and progresses automatically to give the all equivalent landmark spots in the two RLGS profiles.

4 Experiments and discussion

We have implemented the fully-automated algorithms for spot recognition and spot matching on DOS/V computer running with Linux. To evaluate the accuracy and performance of our algorithms, several experiments were carried out using the RLGS profiles derived from the genomic DNA extracted from the colon normal-tissue (hereafter referred to as Normal profile) and the colon tumor-tissue (hereafter referred to as Tumor profile): *Not*I, *Pvu*II and *Pst*I were employed as restriction enzymes A, B and C, respectively.

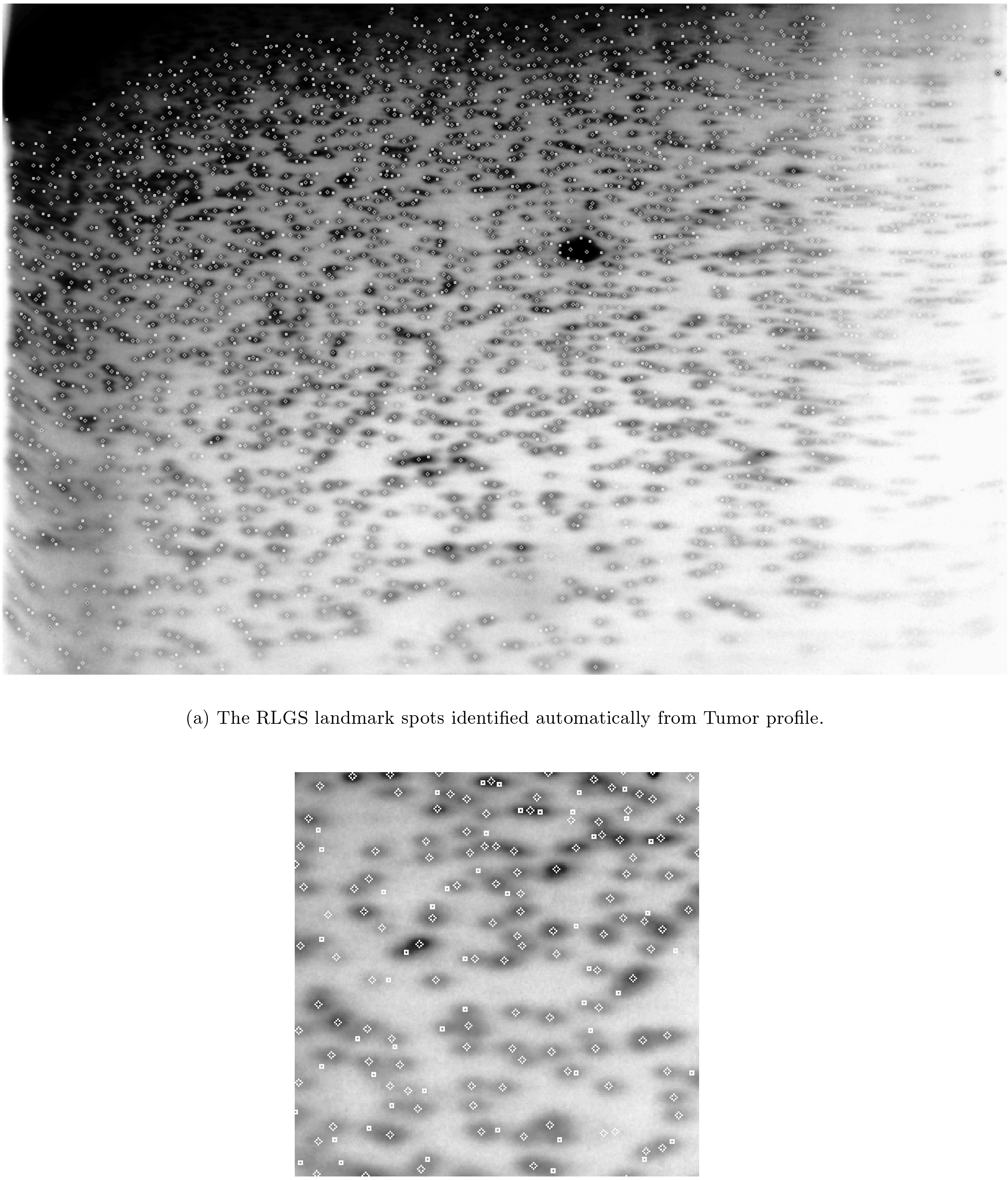
Prior to the experiments, the corresponding X-ray films were digitized with film scanner. In these RLGS profiles, 17 inch × 14 inch sized X-ray film was required to resolve over thousand landmark spots, and gives the image file of 5100 × 4200 pixels with scanning at a resolution of 300 dot per inch. The image file was then re-sampled as to give 1275 × 1050 pixels with 8-bit per pixel. The re-sampled image files were cropped for some convenience and then provided for the consecutive experiments. The image size of the used Normal profile and Tumor profile are 878 × 867 and 1245 × 832, respectively.

4.1 Spot Recognition

The fully-automated spot recognition algorithm was applied to identify landmark spots found on Normal profile and Tumor profile. Fig. 2 shows the locations of the automatically identified 2742 spots overlaid on the image of Tumor profile. As can be seen in this figure, almost of all landmark spots perceptible by visual inspection were correctly identified. It is noteworthy that the hidden spots found at the shoulder part of the neighboring large spots were also recognized correctly by our algorithm, as easily can be seen in Fig. 2b.

The statistics of the automatically identified landmark spots were summarized in Table 1, in which the number of true-negative (un-recognized) and false-positive (ill-recognized) spots are counted by visual inspection. In both cases of Normal and Tumor profiles, only the negligible numbers of true-negative and false-positive spots were counted. These results, consequently, strongly support that our algorithm works very fine with the RLGS profiles which have some irregularities in spot shape and the drifted un-uniform background patterns. Notice that about a thousand of spots, hidden or un-recognized in the first half stage of the algorithm, were recognized based on the differences between background normalized RLGS profile and the Gaussian-modeled profile. As described in later section, such spots, which was never recognized with our previous algorithm [27, 23], show the important role to achieve the accurate spot matching.

Despite the remarkable accuracy in spot identification, our spot recognition algorithm was also very fast. In Table 2, computation time required to process each RLGS profile is shown. Although these results were measured by using the inexpensive DOS/V computer, the image processing of the large RLGS profile was carried out about 4 minutes. The final step in the algorithm was the most time-consuming, but can be finished in practical time.



(a) The RLGS landmark spots identified automatically from Tumor profile.

(b) The RLGS landmark spots identified automatically from Tumor profile (magnified).

Figure 2: The locations of the spots, identified from Tumor profile by applying the fully-automated algorithm, are shown. The spots identified in the first stage of the recognition (through STEP4 to 6) are shown by open diamonds, and the spots identified in the second stage of the recognition (through STEP8 to 9) are shown by open square. The number of spots finally survived in the last step, i.e. successfully fitted by Gaussian-type functions, was 2742.

Table 1: Number of spots identified by means of the fully-automated algorithm.

	Auto-detected [1]	Auto-detected [2]	Finally survived [3]	True-negative	False-positive
Normal	1142	745	1879	8	4
Tumor	1842	950	2742	7	5

[1] The number of spots, detected then survived through STEP4 to STEP6.

[2] The number of spots, detected then survived through STEP8 to STEP9.

[3] The number of spots successfully fitted by Gaussian-type functions after the final stage of the algorithm.

Table 2: Computation time [1] (in sec) required to process digitized RLGS profile.

	image size	S1	S2	S3	S4	S5+S6+S7	S8	S9	S10	total
Normal	878 × 867	0.5	1.9	0.3	19.9	5.4	10.7	2.8	185.2	226.7
Tumor	1245 × 832	0.7	2.6	0.4	37.1	6.7	15.2	3.6	200.9	267.2

[1] User-time required to process each step in the algorithm, which was measured on the DOS/V computer running with Linux: MMX Pentium/233MHz with 64MB SDRAM.

4.2 Spot Matching

As the next experiment, we have performed the identification of the equivalent pairs of landmark spots from the two RLGS patterns. Here, the spot locations and intensities were exactly the same as obtained in the above experiment without any modification. Table 3 summarizes the number of the matched spots and the computation time required. By applying our fully-automated algorithm into the RLGS pattern self-matching, all of the landmark spots were correctly mapped onto themselves, except for only one unmatched spot.

When our algorithm is applied to compare the Normal and Tumor profiles, more than a thousand equivalent spot pairs were identified, and almost identical number of equivalent spots were found even when the reference (Normal/Tumor) and the object (Tumor/Normal) patterns were exchanged. To clarify the quality of such spot pattern comparison, the matched spots on each RLGS profile are depicted in Fig. 3. In this figure, the landmark spots are shown overlaid on the background normalized profile image , together with the RNG constructed only from the matched spots.

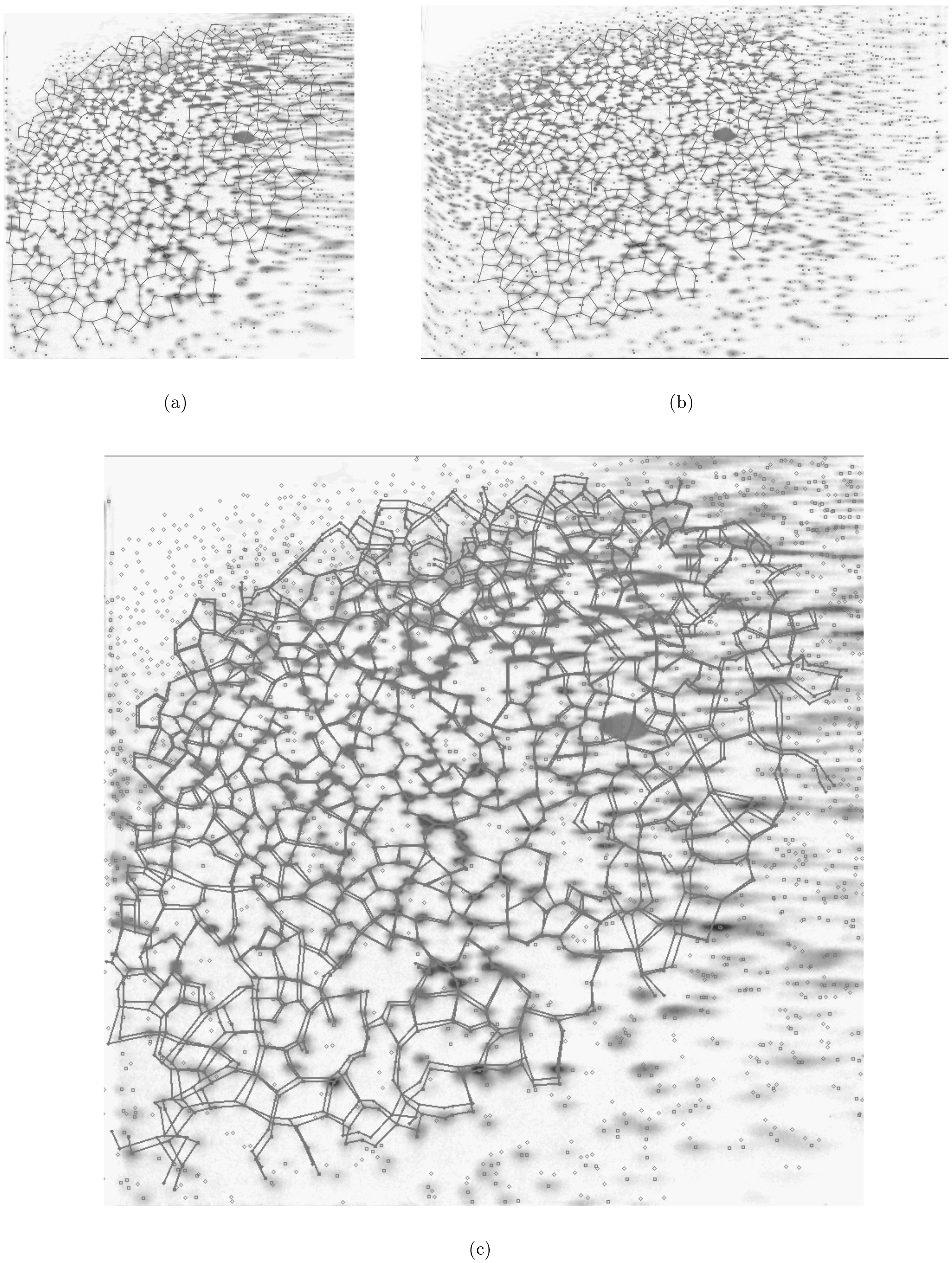
Table 3: Number of the identified equivalent pairs of landmark spots [1] and the computation time [2] required.

reference	object	Number of equiv. spots	initial pair search (sec)	spot matching (sec)
Normal	Normal	1878	51.7	4.4
Tumor	Tumor	2741	349.5	6.4
Normal	Tumor	1066	127.5	3.2
Tumor	Normal	1068	127.5	3.2

[1] Number of the spots found on Normal and Tumor profiles were 1879 and 2742, respectively.

[2] User time, which was measured on the DOS/V computer running with Linux: MMX Pentium/233MHz with 64MB SDRAM. The computational time required to construct DN of Normal and Tumor were 1.0 sec and 1.5 sec on the same computer, respectively.

As can be seen in Fig. 3a and Fig. 3b, both of matched spots in Tumor and Normal profiles are well distributed over the almost entire area in the profile. Because we adopted the breadth-first graph search algorithm to spot pattern comparison, it is possible that the graph search is terminated at the mis-matched node. Once such termination occurs, the matched spots does not spread over anymore.



(a)

(b)

(c)

Figure 3: The matched spots, identified by applying the fully-automated algorithms, are shown overlaid on the background normalized profile image. The RNG (relative neighborhood graph) constructed from the matched spots is depicted for convenience. **(a)**: the matched spots in Normal profile (reference). **(b)**: the matched spots in Tumor profile (object). The matched spots are shown as filled squares, though the un-matched spots are shown as open squares in (a) and (b). **(c)**: the landmark spots and the RNG constructed from the matched spots in Tumor profile are superimposed on (a). The superimposed spots of Tumor profile are depicted as the diamonds, while the spots of Normal profile are drawn with square.

Despite such possibility, this result indicates that the breadth-first graph search is robust enough to traverse the whole spot pattern. This is partly because of the adoption of the DN, which potentially has the redundant paths in itself, and partly because our new spot recognition algorithm can find the hidden spots, which cannot be recognized by the previous algorithm. The hidden spots drastically decrease the possibility of encountering the mis-matched nodes during the traverse.

The pattern matching of the two structured graphs, each of which has more than a thousand nodes, was finished in only a few seconds. The most time-consuming process was the identification of the initial equivalent pair of spots from both profiles, because we adopted heuristic search of the landmark spots. In spite of the simple search algorithm, the initial equivalent spot pair search was carried out within several minutes even with the inexpensive DOS/V computer, as shown in Table3.

As mentioned above, our fully-automated spot matching algorithm works very fine to compare the nonlinearly distorted RLGS profiles and to identify the corresponding spots accurately and rapidly, in concert with out new spot recognition algorithm.

5 Summary

We have developed the fully-automated algorithms for spot recognition from RLGS profiles and for pairwise spot matching. By applying these algorithms to the processing of the two-dimensional electrophoretograms of genomic DNA, molecular changes occurring at or near to the restriction landmarks, such as deletions, additions, translocations or DNA methylations, can be detected effectively. During the processing of the RLGS profiles, the boring and annoying interactions with computer, such as revising mis-recognized spots or spot landmarking prior to the RLGS profile comparison, are anymore unnecessary.

We have implemented such algorithms on the DOS/V computer running with Linux and demonstrated in this paper that the larger RLGS profiles which hold several thousands of landmark spots can be processed effectively even with the inexpensive computers. Actually, it takes less than ten minutes to identify all spots on a RLGS profile and to compare them with those on a reference RLGS profile, while it expected to take over ten minutes to process with our previous image processing system [23] and to take about two to three hours by visual inspection.

The fully-automated algorithms for spot recognition and spot matching should be applicable to the similar problems, such as the processing of the protein two-dimensional gel electrophoretograms.

6 Acknowledgement

We are grateful to Prof. Ueda at Keio University for providing several RLGS autoradiogram images, and to Dr. Hirota at NBT Inc. for his technical supports. Finally, we acknowledge Dr. Hayashizaki at The Institute of Physical and Chemical Research (RIKEN) for his kind suggestions on the first stage of this work.

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from The Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Asakawa, J., Kuich, R., Neel, J.V., Kodaira, M. Satoh, C. Hanash, S.M., Genetic variation detected by quantitative analysis of end-labeled genomic DNA fragments, *Proc. Natl. Acad. Sci. USA*, 91:9052–9056, 1994.

[2] Ballabio, A., The rise and fall of positional cloning?, *Nature Genet.*, 3:277–279, 1993.

[3] Barnard, S.T., Thompson, W.B., Disparity Analysis of Images, *IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-2*, 4:333–340, 1980.

[4] Bird, A.P., CpG island as gene markers in the vertebrate nucleus, *Trends Genet.*, 3:342–347, 1987.

[5] Brilliant, M.H., Gondo, Y., Eicher, E.M., Direct molecular identification of the mouse pink-eyed unstable mutation by genome scanning, *Science*, 242:566–569, 1991.

[6] Collins, F.S., Positional cloning: let's not call it reverse anymore, *Nature Genet.*, 1:3–6, 1992.

[7] Collins, F.S., Positional cloning moves from perditional to traditional, *Nature Genet.*, 9:347–350, 1995.

[8] Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W.K., Koop, B., Hood, L., et al., DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing, *Science*, 260:1649–1652, 1993.

[9] Fischler, M.A., Elschlager, R.A., The representation and matching of pictorial structures, *IEEE Trans. Computer, C-22*, 1:67–92, 1973.

[10] Garrals, J.I., Two-dimensional Gel Electrophoresis and Computer Analysis of Proteins Synthesized by Clonal Cell Lines," *J. Biological Chemistry*, 254:7961–7977, 1979.

[11] Hatada, I., Hayashizaki, Y., Hirotsune, S., Komatsubara, H., Mukai, T., A genomic scanning method for higher organisms using restriction sites as landmarks, *Proc. Natl. Acad. Sci. U.S.A.*, 88:9523–9527, 1991.

[12] Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Hatada, I., Shibata, H., Kawai, J., Hirose, K., Watanabe, S., Fushiki, S., Wada, S., et al., Restriction landmark genomic scanning method and its various applications, *Electrophoresis*, 14:251–258, 1993.

[13] Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Shibata, H., Akasako., A., Muramatsu, M., Kawai, J.,Hirasawa, T., Watanabe, S., Shiroishi, T., et al., A genetic linkage map of the mouse using Restriction Landmark Genomic Scanning (RLGS), *Genetics*, 138:1207–1238, 1994.

[14] Hirotsune, S., Hatada, I., Komatsubara, H., Nagai, H., Kuma, K., Kobayakawa, K., Kawara, T., Nakagawara, A, Fujii, K., Mukai, T., Hayashizaki, Y., The new approach for detection of amplification in cancer DNA using Restriction Landmark Genomic Scanning method, *Cancer Res.*, 52:3642–3647, 1992.

[15] Miwa, M., Yashima, K., Sekine, T., Sekiya, T., Demethylation of a repetitive DNA sequence in human cancers, *Electrophoresis*, 16:227–232, 1995.

[16] Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramirez-Solis, R., Webster, T.D., Ledbetter, D.H., Caskey, C.T., Alu polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA source, Proc. Natl. Acad. Sci. U.S.A., 86:6686–6690, 1989.

[17] Ranade, S., Rosenfeld, A., Point Pattern Matching by Relaxation, *Pattern Recognition*, 12:269–275, 1980.

[18] Sibson, R., Locally equiangular triangulations, *The Computer Journal*, 7:243–245, 1980

[19] Skolnick, M.M., An Approach to Completely Automatic Comparison of Two-Dimensional Electrophoresis Gels, *Clinical Chemistry*, 28:979–986, 1982.

[20] Southern, E.M., Detection of Specific Sequences among DNA Fragments Separated by Gel Electrophoresis, *J. Mol. Biol.*, 98:503–517, 1975.

[21] Southern, E.M., Case-Green, S.C., Elder, J.K., Johnson, M., Mir, K.U., Wang, L., Williams, J.C., Arrays of complementary oligonucleotides for analyzing the hybridization behavior of nucleic acids, *Nucl. Acids Res.*, 22:1368–1373, 1994.

[22] Sternberg, S.R., Biomedical Image Processing, *IEEE Computer*, January:22–34, 1983.

[23] Takahashi, K., Nakazawa, M., Watanabe, Y., DNAinsight: An Image Processing System for 2-D Gel Electrophoresis of Genomic DNA, *Genome Informatics*, 7:135–146, 1997.

[24] Toussaint, G.T., The relative neighborhood graph of finite planar set, *Pattern Recognition*, 12:261–268, 1980.

[25] Toussaint, G.T., Pattern Recognition and Geometrical Complexity, *Proc. Fifth Int'l Conf. Pattern Recognition*, 1324–1347, 1980.

[26] To affinity ... and beyond!, *Nature Genet.*, 14:367–370, 1996.

[27] Watanabe, Y., Takahashi, K., Nakazawa, M., Automated Detection and Matching of Spots in Autoradiogram Images of Two-Dimensional Electrophoresis for High-speed Genome Scanning, *Proc. of International Conference on Image Processing, IEEE Signal Processing Society*, III:496–499, 1997.

[28] Yershov, G., Barskey, V., Belgovskiy, A., Kirillov, E. Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, A., Dubiley, S., Mirzabekov, A., DNA analysis and diagnostics on oligonucleotide microchips, *Proc. Natl. Acad. Sci. USA*, 93:4913–4918, 1996.