# Contour Area Filtering of 2-Dimensional Electrophoresis Images

Ramakrishnan Kazhiyur-Mannar,[1] Dominic J Smiraglia,[2]
Christoph Plass,[3] Rephael Wenger[1]

## Abstract

We describe an algorithm, Contour Area Filtering, for separating background from foreground in gray scale images. The algorithm is based on the area contained within gray scale contour lines. It can be viewed as a form of local thresholding, or as a seed growing algorithm, or as a type of watershed segmentation. The most important feature of the algorithm is that it uses object area to determine the segmentation. Thus it is relatively impervious to brightness and contrast variations across an image or between different images.

Contour Area Filtering was designed specifically for image analysis of 2D electrophoresis gels, although it can be applied to other gray scale images. A typical gel image is an electrophoretogram or a phosphor image of 1000 to 2500 spots representing protein or DNA restriction fragments. The images are quantitative with spot intensities reflective of the number of proteins or the DNA fragment copy number. The background intensity can vary widely across the image caused both by variation in spot density and by the physical laboratory process of creating a gel. Analyzing and comparing gel images entails extracting and segmenting spots, registering images and matching spots, and measuring differences between spots.

We present experimental results which show that Contour Area Filtering is a quick, efficient method for separating electrophoresis gel background from foreground with extremely high accuracy.

## Introduction

Segmentation of biomedical images is plagued by image inhomogeneities. Image contrast and brightness often vary between different images and across a single image (Clarke, Velthuizen et al. 1995; Sugahara, Akiyoshi et al. 1998; Rapantzikos, Zervakis et al. 2003; Sebastian, Tek et al. 2003; Yao, Abolmaesumi et al. 2005). On the other hand, biomedical images often consist of objects whose sizes are consistent across an image

[1] Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, 43210

[2] Department of Cancer Genetics, and Comprehensive Cancer Center, Roswell Park Cancer Institute, Buffalo, NY 14263.

[3] Division of Human Cancer Genetics, Department of Molecular Virology, Immunology and Medical Genetics, and Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, 43210.

and between different images. These sizes are known a priori to the clinician or researcher.

Global thresholding is a simple technique for image segmentation in which all pixels with intensity above or below a given threshold are set to foreground and all other pixels are set to background. Global thresholding is almost always inadequate for biomedical images. Because of differences in contrast and brightness between biomedical images, different thresholds are required for each individual image. Because of differences across an image, different thresholds are required for different portions of the image. Local thresholding algorithms address the problem of variation within an image by setting thresholds within local windows (Sugahara, Akiyoshi et al. 1998; Rapantzikos, Zervakis et al. 2003). However, local thresholding requires some algorithm for setting the local threshold within a window and this algorithm itself depends upon user specified parameters.

The simplest form of segmentation is separating foreground from background pixels. We present an algorithm, Contour Area Filtering, to separate foreground from background which depends on the sizes of the objects. We identify isocontours which enclose regions of a specified area and select those regions as foreground. We do not actually construct the isocontours, only the set of pixels contained by the isocontour. Since algorithm parameters are based on area, not intensity, the algorithm is relatively impervious to variations in brightness and contrast between and across images and can detect even the lightest objects in an image.

Contour Area Filtering can be described as a form of local thresholding with thresholds determined dynamically by object size. It can also be described as a watershed style algorithm which allows small watersheds to merge until they have a desired size. Thus Contour Area Filtering avoids the oversegmentation problems which plague watershed algorithms. Finally, it can also be viewed as a seed or region growing and merging algorithm but with seeds automatically created at every local minimum. (See Bieniek and Moga (2000), Dawant and Zijdenbos (2000), Rogowska (2000), Fu et al. (2004), Clarke et al. (1995), Sebastian et al. (2003), and Yao et al. (2005), for descriptions and applications of watershed and seed growing algorithms.)

We developed Contour Area Filtering specifically to analyze Restriction Landmark Genomic Scanning (RLGS) gels. RLGS is a 2D gel electrophoresis technique developed by Hatada et. al. (1991) for detecting DNA molecular changes that occur near restriction enzyme sites. (See Figure 1.) Genomic DNA is digested by a "landmark" restriction enzyme (i.e., *Not*I or *Asc*I) and radioactive nucleotides are incorporated into the cleavage sites. The fragments are further digested by a second enzyme (i.e., *Eco*RV) and separated along the first dimension using agarose gel electrophoresis. A third enzyme (i.e., *Hinf*I) digests these fragments in gel followed by second dimension separation via polyacrylamide gel electrophoresis. Autoradiography or phosphor imaging is applied to the dried gel producing an RLGS image of approximately 2500 spots (Figure 2). Spots on this image correspond to radioactively labeled DNA restriction fragments.

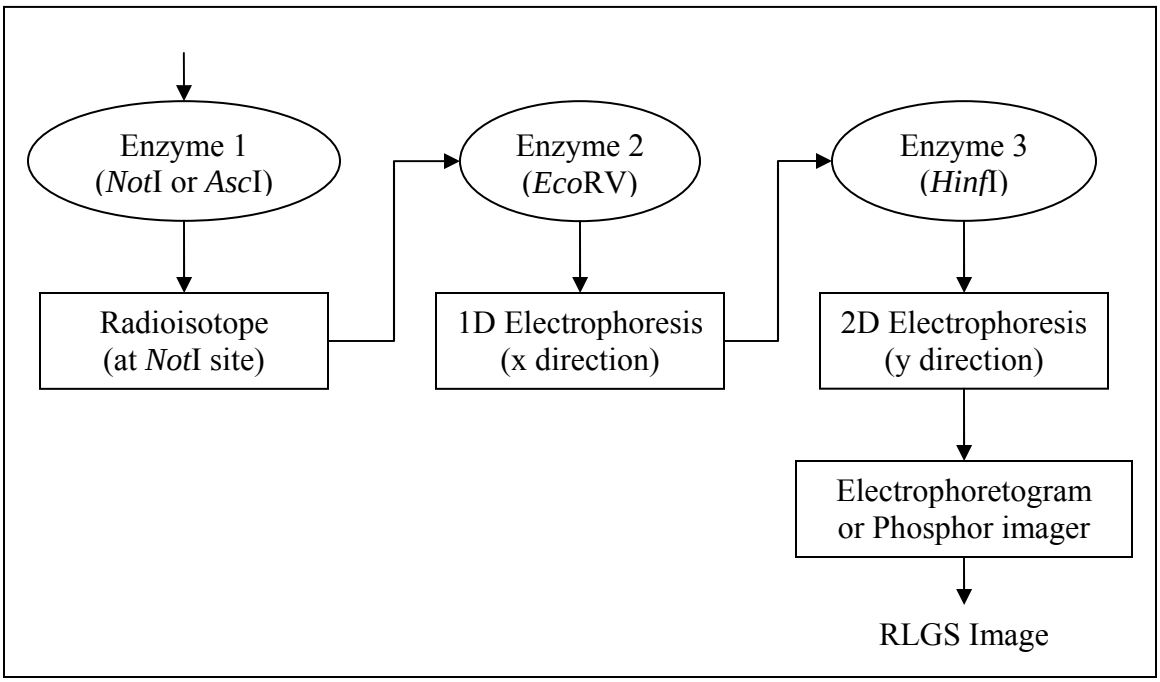



**Figure 1.** Restriction Landmark Genomic Scanning (RLGS).

The location of each spot in an RLGS image is determined by two DNA restriction fragments, the fragment after application of the second enzyme and a subset of that fragment produced by application of the third enzyme. By convention, RLGS gels are oriented so that DNA migration is first horizontal from right to left, and then vertical from top to bottom.

After application of the second enzyme, fragments for spots on the right have 4500-5800 base pairs while fragments for spots on the left have 500-1000 base pairs. Because there are many large fragments, spot density is high on the right side of the gels and gel analysis is difficult in this region. After application of the third enzyme, fragments for spots on the top have 1000-1700 base pairs while fragments for spots on the bottom have 100-200 base pairs.

Not all fragments appear on the gel. Fragments which are very large do not enter the gel at all while fragments which are very small migrate out of the gel. Fragments without a first enzyme cleavage site do not have an attached radioactive nucleotide and do not create spots on the image.

Fragments which are small (~800 base pairs) after the application of the second enzyme may not be split at all by the third enzyme. These fragments will congregate in the upper left of the gel, forming a dark curve in that region. Because spots along this curve are not separated in two dimensions, they are very dense and cannot be distinguished.

**Figure 2.** Human *Not*I-*Eco*RV-*Hinf*I master gel. DNA fragments migrate first horizontally from right to left and then vertically from top to bottom. Spot density is higher on the right side of the gel, causing the region to be darker. The dark curve on the upper left of the gel is caused by fragments which are not split by the third enzyme and cluster along this curve.

The majority of spots arise from diploid DNA fragments and appear as "normal" spot intensity while haploid DNA fragments create spots half as dark (Asakawa, Kuick et al. 1995) and may be due to heterozygosity or to allele specific methylation. (See Figure 3.) The 15-20 very large dark spots appearing in the gel are fragments from repetitive elements as well as ribosomal DNA fragments. Each ribosomal DNA fragment has approximately a 200 fold excess in fragment number over the diploid DNA fragments.

RLGS spots from normal, non-repetitive DNA fragments are oval in shape with dimensions approximately 0.5x0.4 cm on the autoradiographs. Spots often have a slight "tail" to the right, a result of the horizontal migration of the DNA from right to left in the electrophoresis separation along the first dimension. While the shape of individual, isolated RLGS spots is fairly uniform, the shape of clusters of overlapping spots can be quite arbitrary.

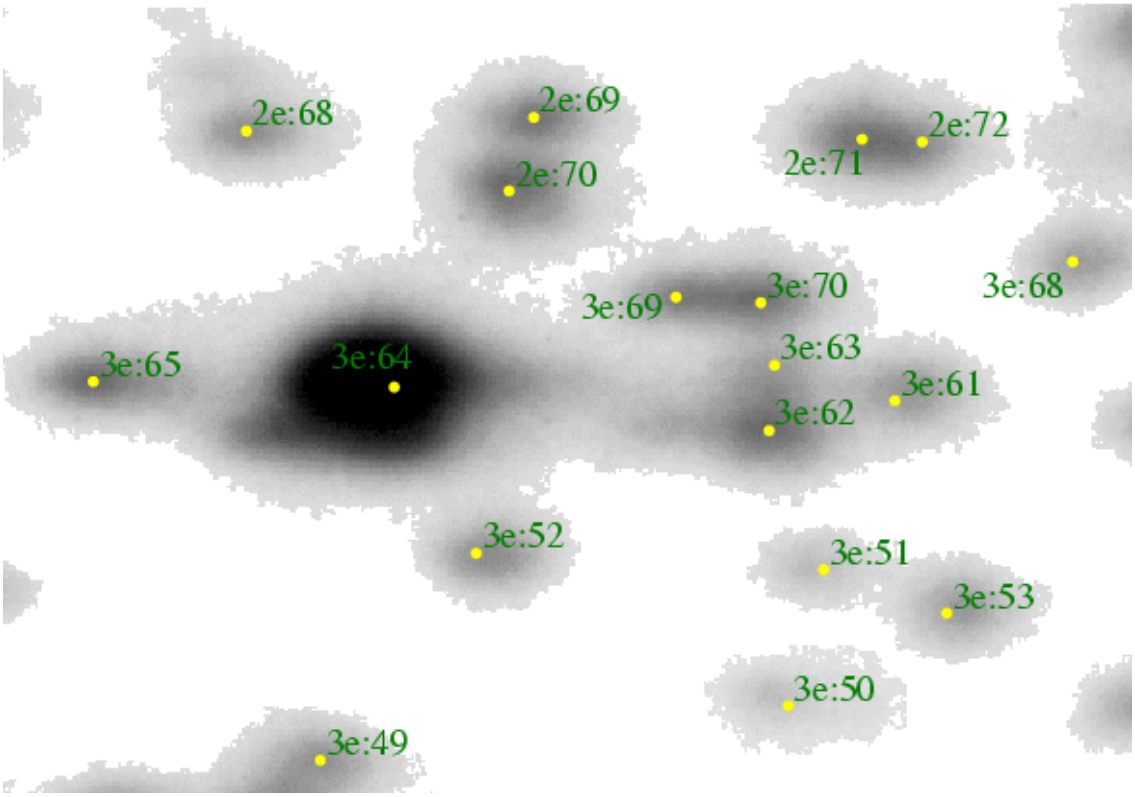



**Figure 3.** Examples of various intensity spots. RLGS spot intensity is measured from 0 to 1 with 0 as white and 1 as black. Spot 3e:64 has maximum intensity 1 and may represent ribosomal DNA or repetitive elements. Spots 2e:69, 2e:70, 3e:52, 3e:53, 3e:62, 3e:65, 3e:69 and 3e:70 have intensity from 0.4 to 0.55 and probably represent diploid DNA fragments. Spots 3e:50 and 3e:51 have intensity 0.13 and 0.19, respectively, and represent haploid DNA fragments.

RLGS is used to detect DNA methylation changes between two genomes or between normal and tumor DNA. (For review, see Smiraglia and Plass (2002).) Methylation sensitive landmark enzymes (i.e., *Not*I or *Asc*I) cut the DNA only at unmethylated sites but not at methylated sequences. RLGS images are created using DNA from tumorous and normal DNA which are then compared. Missing, added, or amplified spots indicate DNA methylation changes or DNA copy number changes respectively. These changes are commonly found in cancer and are potential biomarkers of the tumor. This methodology has made significant contributions to our understanding of the importance of aberrant DNA methylation in cancer biology both through estimation of the extent to which it occurs (Costello, Frühwald et al. 2000; Smiraglia, Rush et al. 2001) and the identification of novel tumor suppressor genes such as SOCS1(Yoshikawa, Matsubara et al. 2001), BMP3B (Dai, Lakshmanan et al. 2001) and SLC5A8(Li, Myeroff et al. 2003).

Two-dimensional gel electrophoresis is a standard technique in protein analysis. Extensive research and software has been developed for automatic analysis of protein images (http://expasy.ch/melanie; http://gelmatching.inf.fu-berlin.de; http://www.bio-rad.com; http://www.amershambiosciences.com; http://www.phoretix.com; Appel, Plagi

et al. 1997; Appel, Vargas et al. 1997). Commercial software includes ImageMaster, Melanie, PDQuest, and Phoretix among others (http://expasy.ch/melanie; http://www.amershambiosciences.com; http://www.phoretix.com; http://www.bio-rad.com;). These packages provide excellent user interface, statistical and database tools for assisting in gel analysis. They have more difficulty in fully automating the detection, identification and comparison of spots on the gel images.

Two software packages have been developed specifically for the analysis of RGLS gels, RAT (RLGS Analysis Tool) by Sugahara et. al. (1998) and DNAInsight by Takahashi et. al. (1997, 1998, 1999 & 2001). To the best of our knowledge, neither package is used by any RLGS laboratory in the United States and neither package is under further development.

Analysis of RLGS gels from genomic DNA poses certain challenges compared with analysis of protein gels. RLGS spots are often lighter and smaller than protein spots. As a result, identifying individual RLGS spots is more difficult than identifying individual protein spots. It is often the lightest RLGS spots, their existence or lack thereof, which is of most interest in detecting DNA methylation. Moreover, samples of tumor tissue almost always contain some normal tissue, creating faint images of spots from the normal DNA which are affected, or methylated, in the tumor DNA. In addition, RLGS gels typically contain over 2500 spots (although this is enzyme dependent) which is the high range for gel analysis.

Numerous algorithms and techniques are used for filtering background pixels and identifying spots in protein and RLGS gels. Sternberg (1983) applies 3D gray scale morphological operators to separate foreground from background. Takahashi et. al. (1997 & 1998) use the same morphological operators and then apply a "ring operator" to identify individual spots and their centers. They subtract these spots from the image and reapply their "ring operator" to further identify hidden spots. Sugahara et. al. (1998) use local thresholds to remove background pixels from the image. The software package Melanie by Appel et. al. (1997) uses thresholding of the second derivative of the gray scale intensities to identify foreground pixels. ImageMaster (http://www.amershambiosciences.com) compares pixel intensities to intensities on the boundary of a surrounding window to identify foreground pixels.

The filtering methods described above have two major drawbacks. First they all require the setting of some sensitivity threshold related to the gray scale intensity of the spots. Users must often adjust these parameters for individual gels. Second, because of this sensitivity thresholding, the algorithms may fail to detect the lightest spots. Many of the algorithms implicitly smooth gray scale intensities, causing these lightest spots to "wash out" with the background. In addition, full intensity saturated spots have to be handled specially by some of the algorithms.

As previously noted, Contour Area Filtering does not rely upon any gray scale sensitivity threshold. In fact, there is no sensitivity parameter as input to our algorithm. Instead, the primary input to our algorithm is the maximum area of a cluster of overlapping spots.

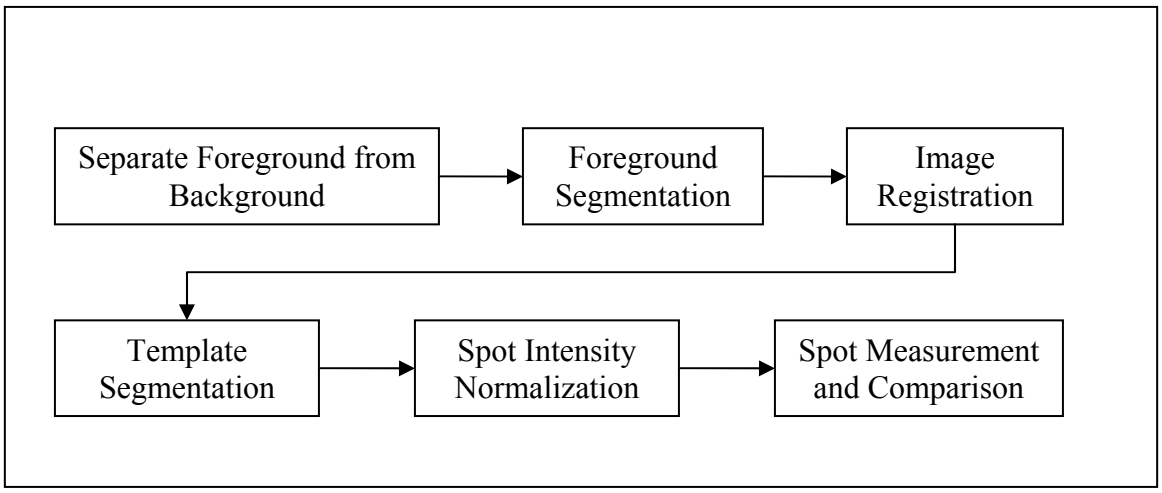



**Figure 4.** RLGS image analysis pipeline.

This metric depends upon spot density and is much more robust across gels than spot intensity. Our algorithm can detect even the faintest spots as long as they are not in regions of high spot density. Faint spots which are adjacent to large clusters of darker spots can be missed by our algorithm but such hidden spots pose a problem for almost all algorithms.

All biomedical images contain noise and RLGS images are no exception. We remove noise in post-processing steps after Contour Area Filtering by first applying an opening operator (erosion followed by dilation) to the foreground and then removing any connected components with area under some user specified threshold. (See Figure 9.) We use 4-connectivity where pixels are connected to pixels directly above, below or right or left of them. Note that opening and removing small components are post-processing steps applied after Contour Area Filtering. Contour Area Filtering is applied directly to the unfiltered data and does not perform any implicit smoothing of the data. Thus it can detect even the faintest spots.

The operations of opening and removing small components are extremely well suited for eliminating noise from RLGS images, but other post or preprocessing steps may be more suitable for other images. For instance, streaking is a common problem in protein gels which is not found in RLGS gels. Streak removal using gray scale opening as described in Sternberg (1983) may be an appropriate preprocessing step for such protein gels.

We applied our algorithm to three different types of RLGS gels and demonstrate that it performs exceptionally well, correctly identifying most of the background pixels and only rarely misidentifying spot pixels as background pixels. We would have liked to compare our results with results from DNAInsight by Takahashi et. al. (1998) but were unable to obtain a copy of their software. Instead we compared results from our algorithm with results from ImageMaster (http://www.amershambiosciences.com). We also report the results of applying our algorithm to two benchmark protein gels used in Raman et. al.

(2002) and Rosengren et al. (2003). We did not compare our results on these protein gels with results from other software, although reports on the application of PDQuest, Progenesis, Z3 and Melanie to these benchmark protein gels are included in Raman et al. (2002) and Rosengren et al. (2003).

Our isocontour filtering algorithm is one component of an automated RLGS gel analysis system under development (Figure 4). The final step in this system is spot measurement and comparison. Spots are compared by their intensity, both the maximum "normalized" intensity of the spot and the integral of the "normalized" intensity of all points in a spot. The latter is called the spot "volume".

Construction and use of isocontours is a standard technique in image processing and visualization. Variations in background and image intensity, the difficulty of selecting appropriate isocontours and lack of control over object shape has led to them being supplanted by active or deformable contours (also called snakes) in medical image processing. Contour Area Filtering addresses the first two of these problems by using the enclosed area in choosing isocontours. Contour Area Filtering is substantially different than all other published algorithms for protein or RLGS gel segmentation, none of which use isocontours. We have been unable to find any published image processing algorithm, biomedical or otherwise, which uses area to control isocontour selection.

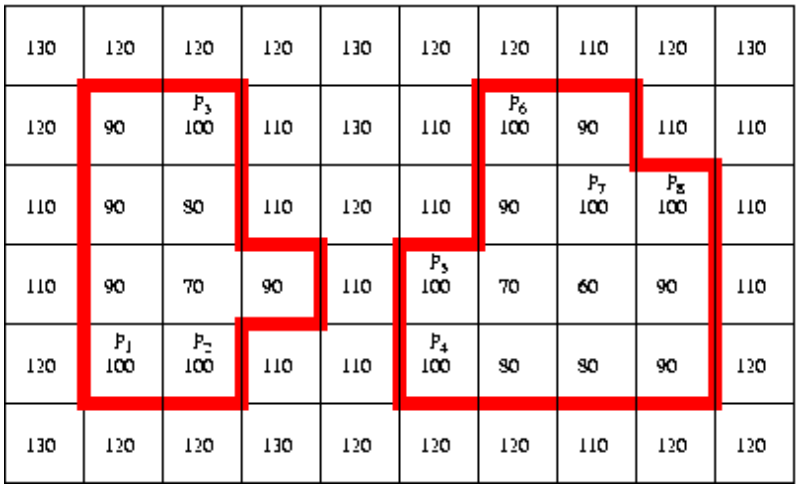



**Figure 5.** Maximal connected components of pixels with intensity gray scale value 100. Range of gray scale values is 0 (black) to 255 (white). The connected component on the left has nine pixels. The connected component on the right has thirteen pixels. The contour area of pixels $p_1$, $p_2$ and $p_3$ is nine. The contour area of pixels $p_4$, $p_5$, $p_6$, $p_7$ and $p_8$ is thirteen.

## Materials and Methods

*RLGS gels*

RLGS gels for both mouse and human genomic DNAs were run as previously described in Okazaki et al. (1995) and modified as described in Smiraglia et al. (1999).

The autoradiograms are scanned at 300 dots per inch and stored as a tiff image of 5100 x 4200 pixels with 8 bits per pixel representing a gray scale in the range 0 to 255.

*Contour Area Filtering Algorithm*

The convention in RLGS is to display spots as dark while the background is light or white. Thus, our presentation of the Contour Area Filtering algorithm will assume that foreground pixels are dark while the background pixels are light. Of course, for the many medical images with the reverse representation we need to simply switch the roles of the dark and light pixels.

The Contour Area Filtering algorithm is shown in Figures 6, 7, 8 and 9. The main procedure, Contour_Area_Filter, takes a single parameter M which bounds the size of the largest connected component of foreground pixels. This parameter requires a priori knowledge of the size of the objects in the image.

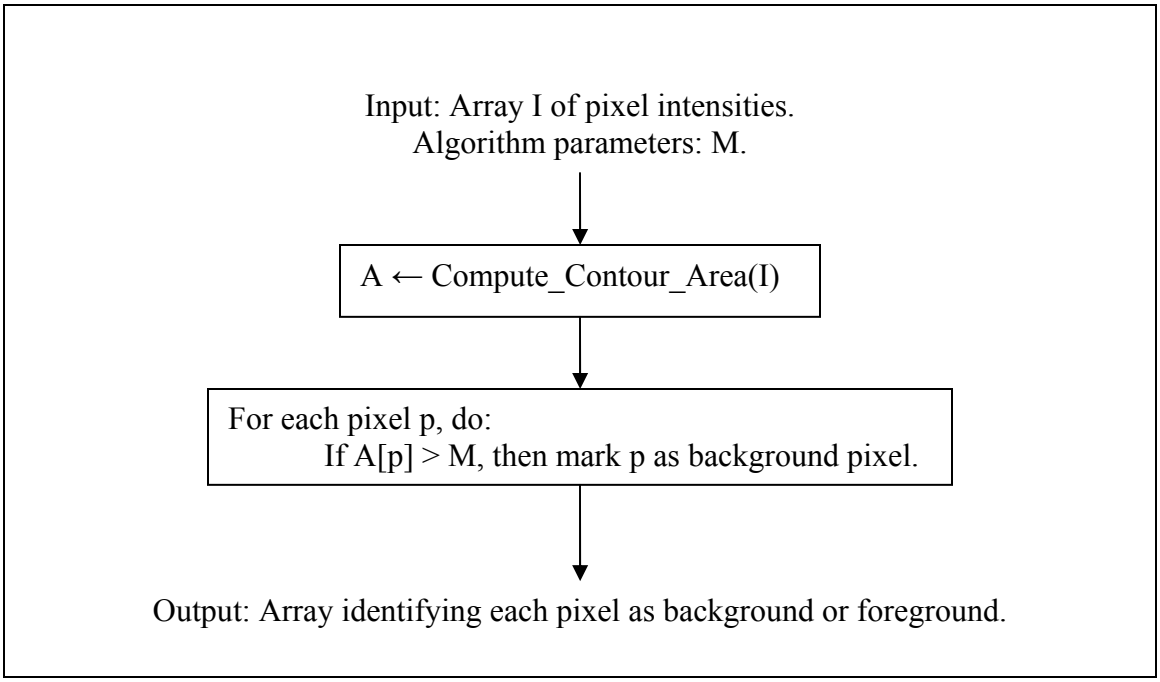



**Figure 6.** Contour Area Filter algorithm.

If a pixel $p$ with intensity $\gamma$ is foreground, then all the adjacent pixels with intensity darker than $\gamma$ should also be foreground. All their adjacent pixels with intensity equal to or darker than $\gamma$ should also be foreground. Consider the maximal connected component containing $p$ and pixels with intensity equal to or darker than $\gamma$. (Use 4-connectivity, connecting a pixel to the pixels to the left, right, above and below.) If $p$ is foreground, then all the pixels in this component are also very likely foreground. If this component is very large, then it is a good indication that pixel $p$ is not foreground.

If we replace the pixels by a continuous scalar field, then we can replace the maximal connected component by the area enclosed by an isocontour through $p$. An *isocontour* is a curve consisting of points with the same scalar value. If the area contained by the isocontour through $p$ is large, then we mark $p$ as background. This area can be thought of as part of the "watershed" containing $p$ (where intensity represents depth) and is used in many similar "watershed" based algorithms (Bieniek and Moga 2000; Fu, Hojjat et al. 2004).

For each pixel $p$ with intensity $\gamma$, let $A_p$ be the number of pixels in the maximal connected component containing $p$ and pixels with intensity equal to or darker than $I_p$ (Figure 5). We mark all pixels with $A_p$ greater than the input parameter $M$ as background. (See Figure 6).

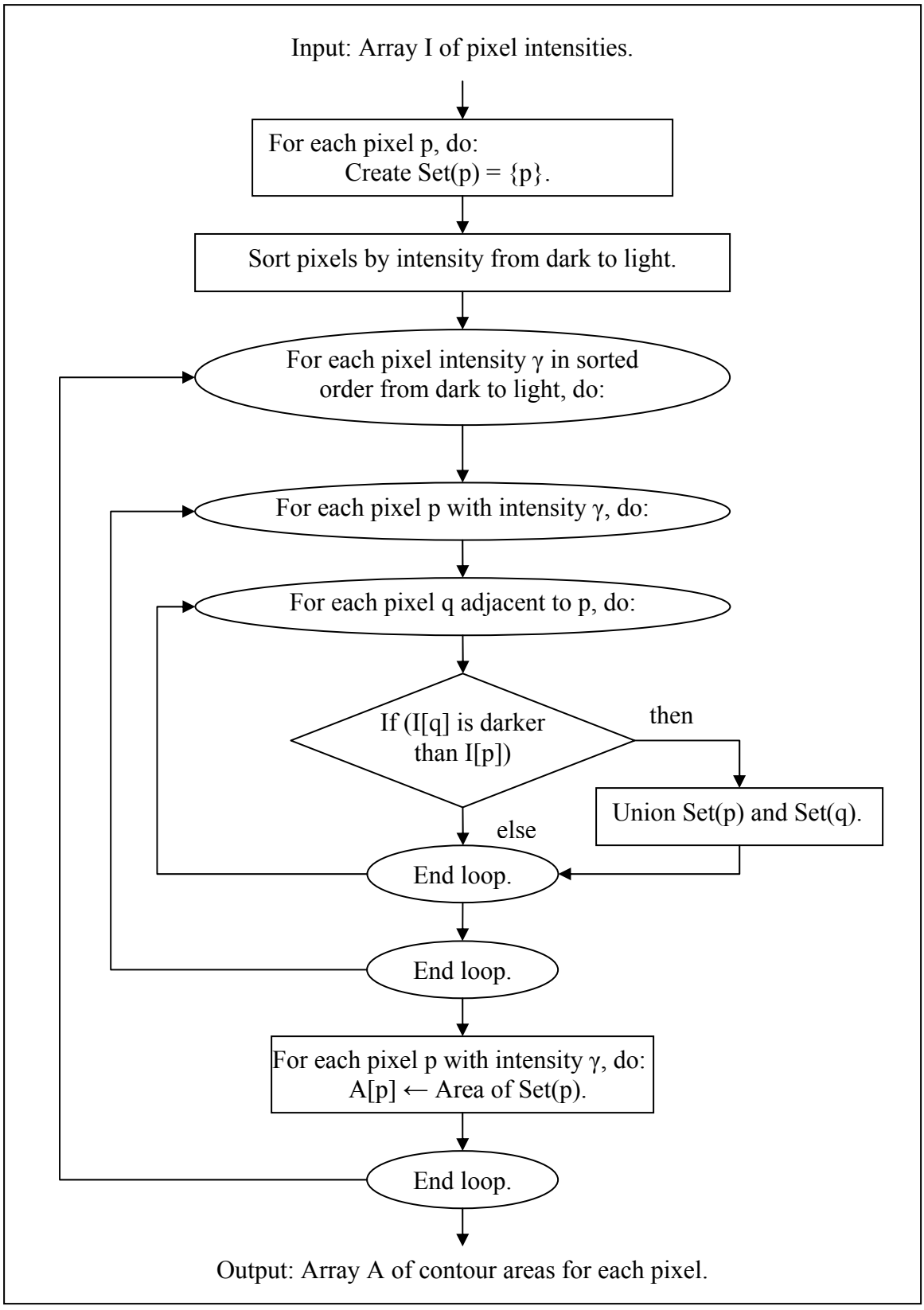



**Figure 7.** Compute Contour Area algorithm.

```
Create_Set(p)
/* p is a pixel */
    1. U[p] ← p;
    2. A[p] ← 1;

Union(p, q)
/* p and q are pixels */
    1. s ← FindSet(p);
    2. t ← FindSet(q);
    3. U[t] ← p;
    4. A[p] ← A[p] + A[t];

FindSet(q)
    1. if (q ≠ U[q]) then
    2.       U[q]  ← FindSet(U[q]);
    3. return(U[q]);
```

**Figure 8.** Subroutines for Compute Contour Area. Array U represents the set containing each pixel. Array A stores the size of the set containing p. Note that once the contour area of p is stored in A in Compute_Contour_Area, array A is never modified.

Of course, if we compute $A_p$ for each pixel separately, the algorithm would be far too slow. Instead we compute $A_p$ in a single pass by slowly growing components starting at their most intense pixels. For each pixel, create a set containing only that pixel. Sort the pixels by intensity from dark to light. Sorting the pixels by intensity also sorts the set of pixel intensities. For each intensity γ in order from dark to light, make two passes over the set of pixels with intensity γ. First, for each pixel $p$ with intensity γ, union the set containing $p$ and sets containing pixels adjacent to $p$ (left, right, top, bottom) with intensities greater than or equal to γ. This forms maximal connected components of pixels with intensity at least γ. Next, for each pixel $p$ with intensity γ, store the size of the set containing $p$. This size is $A_p$. (See Figure 7).

We use a slight modification of the standard union-find data structure to represent the sets of pixels (Cormen, Leiserson et al. 2001). The data structure is represented in an array, U, of pointers, one for each pixel. A set is represented by a tree of pointers, pointing back to the root. FindSet($q$) returns the element at the root of the tree containing $q$ by following pointers U[$q$] back to the root. (See Figure 8). It also performs "path compression" by resetting the pointers along the way to point to the root. To form the union of two trees rooted at $s$ and $t$, we simply set U[$t$] equal to $s$.

The algorithm requires the input array, I, of pixel intensities and two other arrays U and A. Array U contains the pointers used to find the root of the tree containing a given pixel. Array A initially stores the size of the set containing each pixel. However, as each pixel $p$ is processed in Compute_Contour_Area, the contour area of the isocontour through $p$ is stored in A[$p$]. Element A[$p$] is never modified after this.

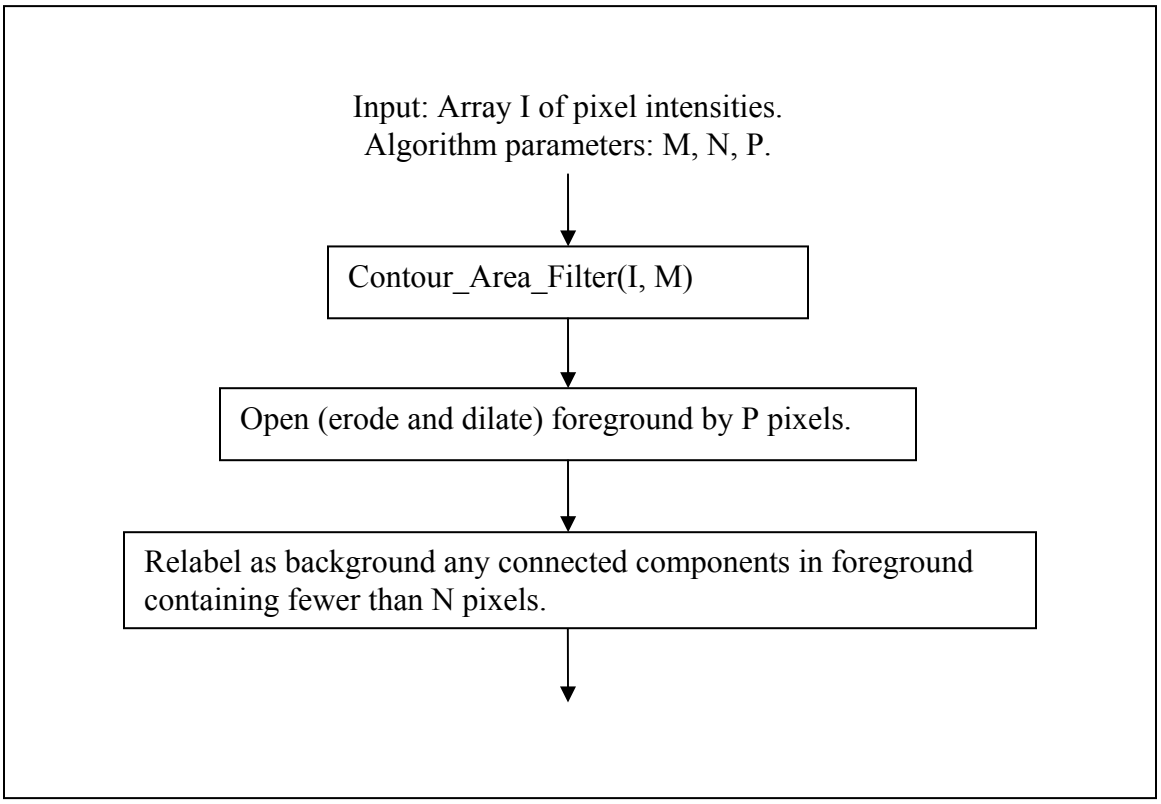
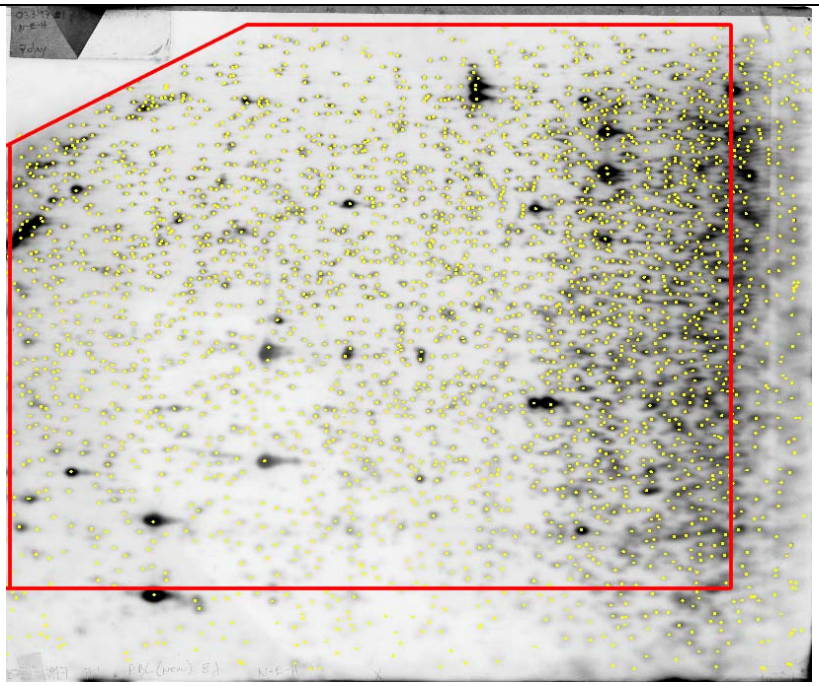



**Figure 9.** Contour Area Filtering and noise removal.

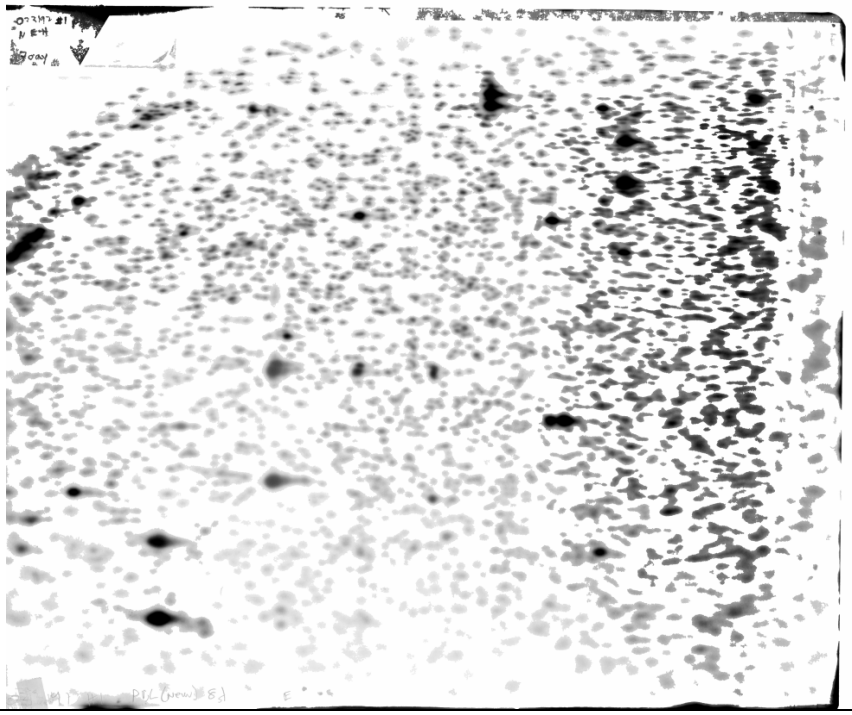
The algorithm runs in worst case O(n log(n)) time where n is the number of pixels. A modification which requires one more field per pixel can improve the running time to near, although not quite, linear. (See Cormen et al. (2001) for a full discussion of union-find algorithms and their implementations.) In practice, the algorithm seems to take linear time and the modification is unnecessary.

The running time and space for sorting pixels depends upon the type of sort used. Our images are 8-bit gray scale consisting of only 256 pixel intensities and so bucket sorting will sort the pixels in O(n) time using one array of size n. For larger sets of intensities, a more general O(n log n) sorting algorithm can be used.

Contour_Area_Filter is a very conservative procedure which cannot distinguish between noise and foreground. After applying Contour_Area_Filter, we use some standard morphological operators to remove some of the noise from the foreground (Figure 9). First, we apply the opening operator (erosion followed by dilation) to remove tenuous connections between pixels. Second, we remove any remaining "salt and pepper" noise by identifying very small foreground connected components and marking them as noise. Again, we use 4-connectivity where pixels are connected to pixels directly above, below or right or left of them. All reported results include the application of opening and of removal of small components.

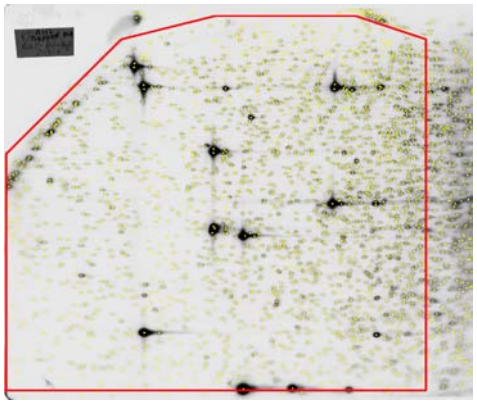


Human *Not*I-*Eco*RV-*Hinf*I master gel with spot centers marked. Bounded region contains annotated spots.

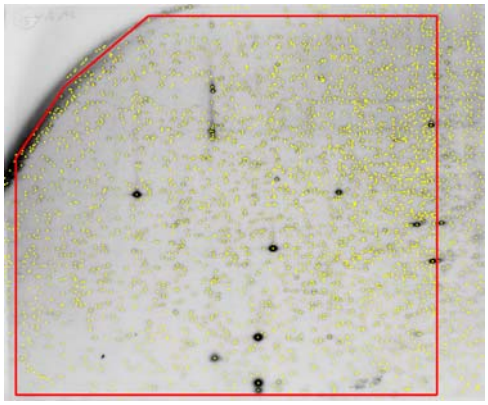


Contour area filtered *Not*I-*Eco*RV-*Hinf*I human master gel.

**Figure 10.** Human master *Not*I-*Eco*RV-*Hinf*I gel, hand annotated and contour area filtered.



Human *Asc*I-*Eco*RV-*Hin*fI master gel.



Mouse *Not*I-*Eco*RV-*Hinf*I master gel.

**Figure 11.** Human *Asc*I-*Eco*RV-*Hin*fI master and mouse *Not*I-*Eco*RV-*Hinf*I hand annotated master gels. Bounded regions contain annotated spots.

*Test Data: RLGS Gels*

For test images we used master RLGS images created using enzyme combinations *Not*I-*Eco*RV-*Hin*fI and *Asc*I-*Eco*RV-*Hin*fI on human DNA and *Not*I-*Eco*RV-*Hin*fI on mouse DNA. (See Figures 10 and 11.) The human DNA is from the peripheral blood lymphocytes (PBLs) of a single healthy female donor. A mouse master gel was created from a combination of DNAs from mouse strains FVB, C57/BL6J, and 129/SV. A mouse FVB gel was created from DNA from mouse strain FVB. The master gels are used as a reference for all other gels with matching enzyme and genome in our laboratories and have been extensively analyzed. We digitized autoradiograms of the gels at 300 dots per inch, creating tiff images of 5100 x 4200 pixels with 8 bits per pixel representing a gray scale in the range 0 to 255. We implemented and tested Contour_Area_Filter on a 2.8 GHz personal computer with 2 Gigabytes of RAM running under the Linux operating system. Our algorithm (including opening and removal of small components) runs in approximately 10 seconds on gels with dimension 5100 x 4200 pixels.

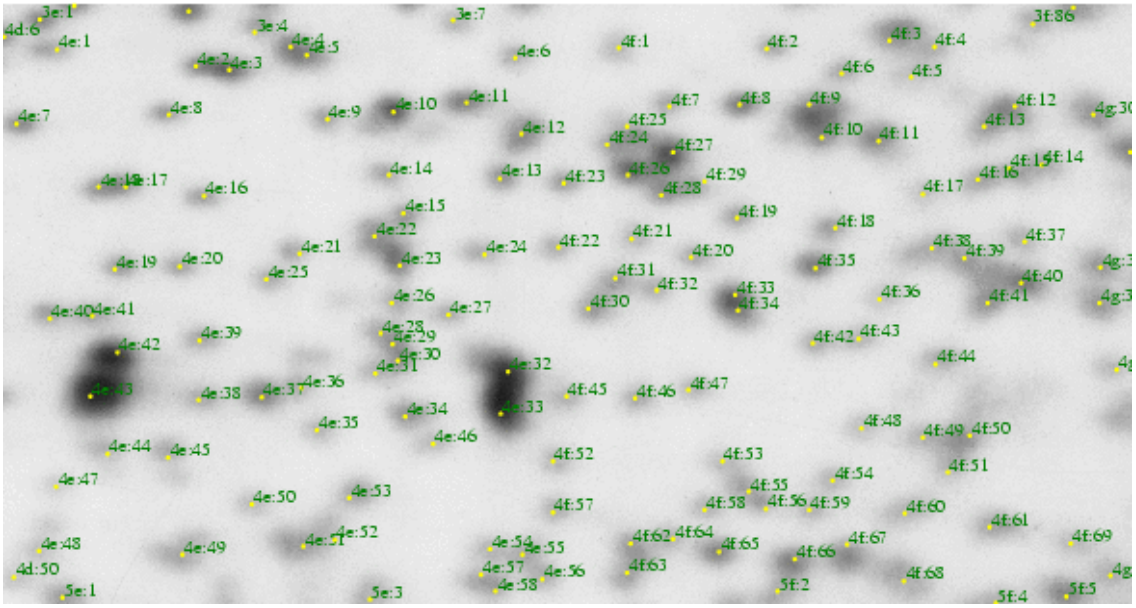


**Figure 12.** Hand annotation of regions 4E and 4F of human master *Not*I-*Eco*RV-*Hinf*I. The gel is partitioned into forty-eight rectangular regions, consisting of six rows and eight columns. Rows are labeled with numbers 1 through 6, while columns are labeled with letters 'a' through 'h'. Spots within each region of the partitioning are numbered starting at 1. Spot labels are the region label followed by the spot number within the region.

Typical spots on a 300 dpi RLGS image have 2000 pixels. The smallest we found has 500 pixels while the largest has 30,000 pixels. We applied Contour_Area_Filter using a threshold of 60,000 for the maximum contour area, which is twice the area of the largest spot. Spots above this size were marked as background. In post-processing, we used a pixel size of two for opening (eroding and then dilating the foreground by 2 pixels) and removed any small components with size less than 300 pixels. 300 pixels is a reliable lower bound on the area of the smallest spot. The pixel size of two for opening was based on examination of numerous RLGS images produced in our laboratories and experimentation with different opening sizes. The three settings are used for Contour Area Filtering all 300 dpi RLGS images in our laboratories and in practice are never changed.

Our laboratories use an annotated image of each master gel, with an identifier marking each spot as described in Costello et al. (2000) for the human *Not*I-*Eco*RV-*Hinf*I master image (Figure 12). Spots on the boundaries, particularly the right boundary, are not labeled. Labeled spots are in the bounded region marked on each gel. (See Figures 10 and 11.) We compared the spots identified on the annotated master image with the foreground pixels generated by our algorithm within the bounded regions. Any spot in the annotated image whose pixels were not identified as foreground by our algorithm were marked as a missed spot. Any significant set of foreground pixels which were not part of an annotated spot were marked as an added spot.

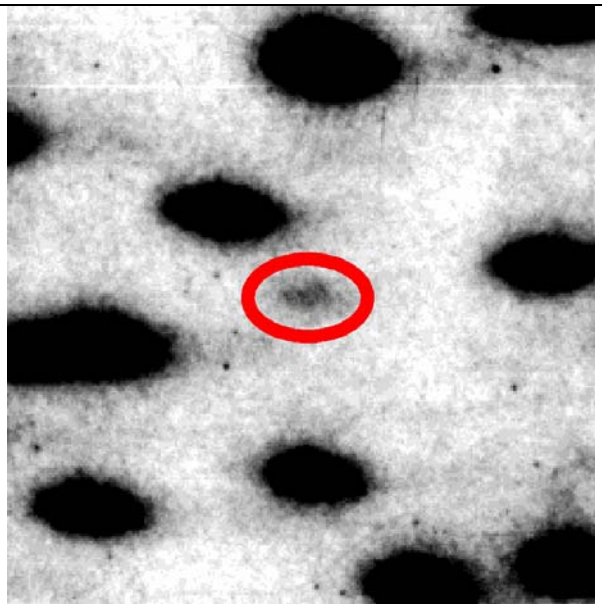


**Figure 13.** Faint spot added by Contour Area Filtering.

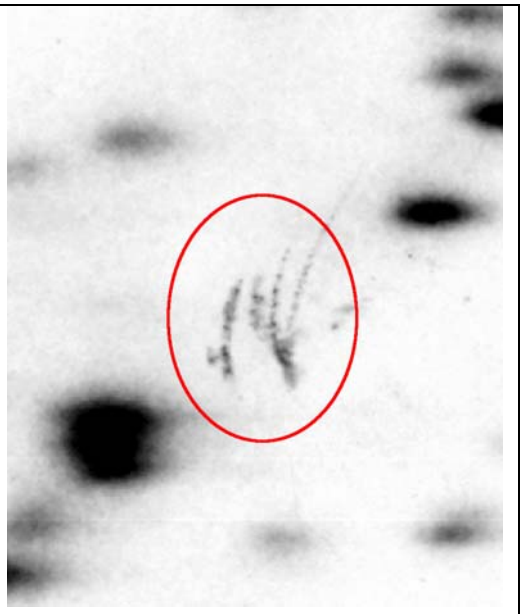
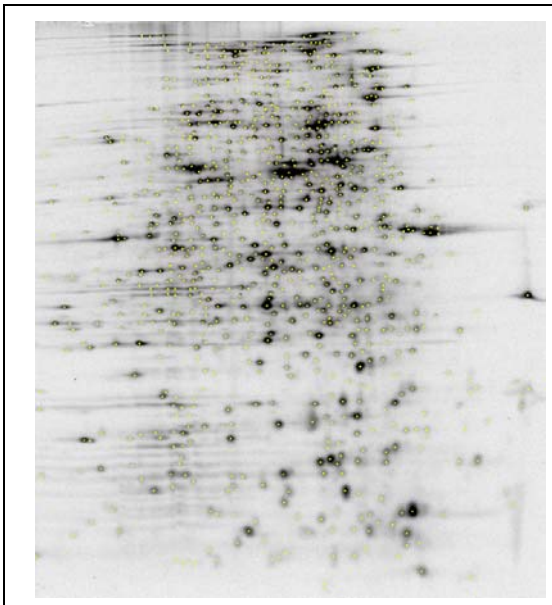


**Figure 14.** Noise identified as spot by Contour Area Filtering.

In order to better understand the nature of the errors made by the algorithm we broke the added spot errors into three distinct classifications: “Faint spots”, “Faint noise”, or “Dark noise”. Faint spots are pixels which have the appearance of a spot but are faint and more difficult to detect by hand (Figure 13). Some of these are faint only compared to surrounding spots. Others are so faint that they are only visible after applying contrast enhancement to the image. The “Faint spot” classification of added spots therefore does not necessarily represent errors on the part of the algorithm, but may also represent the advantage in detection capability of the algorithm over the inherently subjective analysis by hand. Added spots marked “Faint noise” are clusters of light pixels which are identified as spots which did not have the shape or appearance of a spot, even after contrast enhancement. Added spots marked “Dark noise” are clusters of dark pixels probably caused by imperfections or physical marks on the gel (Figure 14). Both “Faint noise” and “Dark noise” represent errors in the algorithm where marks that are clearly not true spots are added as spots.

The missed spots are broken down into two categories: “Faint” and “Distinct”. The “Faint” classification is defined in the manner described above. The “Distinct” classification represents spots that are clearly present in the image but not identified by the algorithm. These two classes represent true errors of lack of identification by the algorithm.



Annotated Gel a.

Contour area filtered Gel a.

**Figure 15.** Benchmark protein gel used by Rosengren et. al. (2003).

*Test data: Protein Gels*

We also applied Contour Area Filtering to two benchmark protein gel images described in Raman et al. (2002). (See http://www.umbc.edu/proteome.) Both are 8 bit gray scale images. Dimensions of Gel A are 666 x 727 pixels and of Gel B are 993 x 1087 pixels. Because the protein gel image resolutions differed from the RLGS image resolutions and from each other and because protein spots are different from RLGS spots, we adjusted the algorithm parameters. The average protein spot in Gel A has 150 pixels, the smallest has 20 pixels and the largest has 1000 pixels. The average spot in gel B has 300 pixels, the smallest has 30 pixels and the largest has 1100 pixels. For both protein gel images we used a threshold of 15,000 for the maximum contour area and a pixel size of two for opening. For Gel A we used a minimum center size of 20 pixels and for Gel B we used a minimum center size of 40 pixels. The parameters were chosen by experimenting on each gel.

Raman et. al. provide annotated versions of both images with boundaries drawn around each spot (Raman, Cheung et al. 2002). Unfortunately, it is not always possible to differentiate between spots in these annotated images. We used Raman et. al.'s annotation to generate our own annotated versions with a dot at each spot center (Figure 15).

All gels and annotated gels used in these experiments can be found at http://www.cse.ohio-state.edu/graphics/conime.

*ImageMaster*

ImageMaster is a commercial 2D protein gel analysis package from Nonlinear Systems (http://www.nonlinear.com). It is similar to the Phoretix package described in Mahon and Dupree (2001) and Rosengren et al. (2003). We applied ImageMaster to three of the master RLGS gels for comparison with Contour Area Filtering. Our configuration of the ImageMaster system could not handle 5100 x 4200 images so we reduced their dimensions to 2550 x 2100.

We note that the ImageMaster system is designed and used for protein, not RLGS gels. Protein spots tend to be darker and more well defined than RLGS spots and typical protein gels often have fewer spots than RLGS gels. The aim of protein gel analysis is often quantification as opposed to detection of added or missing spots. Thus correctly detecting faint spots is both less difficult and less important in protein gels.

ImageMaster identifies foreground pixels by comparing the average intensity of $k$ pixels in a neighborhood of a pixel $p$ with the average intensity of a $4k$ pixels on the boundary of a window around $p$. It uses three significant parameters: sensitivity, window size and noise. Pixel $p$ is classified as foreground if $(I_p - I_s)/I_p > s/10000$, where $I_p$ is the average intensity in the neighborhood of $p$, value $I_s$ is the average intensity of the $4k$ pixels on the boundary of the window around $p$, and s is the sensitivity parameter. Higher values of $s$ detect more spots but give more false positives. The noise parameter is the number $k$ of pixels used for the neighborhood of $p$. It reduces the effect of high frequency noise on the filtering. The window size determines the size of the spots detected. Smaller window sizes detect smaller spots, but fail to detect large saturated spots. We used sensitivity 9500, window size 15x15, and noise 7 on our 2550 x 2100 images.

| Gel | Enzymes | Annot # spots | CAF # added | CAF # missed | IM # added | IM # missed |
|---|---|---|---|---|---|---|
| Human DNA | *Not*I-*Eco*RV-*Hin*fI | 2425 | 46 | 34 | 104 | 51 |
| Human DNA | *Asc*I-*Eco*RV-*Hin*fI | 2277 | 63 | 64 | 84 | 164 |
| Mouse master | *Not*I-*Eco*RV-*Hin*fI | 3228 | 47 | 139 | 15 | 719 |
| Mouse fvb | *Not*I-*Eco*RV-*Hin*fI | 2590 | 38 | 5 | *** | *** |
| Protein Gel A | | 922 | 91 | 100 | *** | *** |
| Protein Gel B | | 1350 | 117 | 91 | *** | *** |

**Table 1.** Results from filtering DNA and protein images using Contour_Area_Filter (CAF) and ImageMaster (IM): Number of labeled spots on hand annotated master images, number of spots added or missed by Contour_Area_Filter (including post-processing,) and number of spots added or missed by ImageMaster.

## Results

*Contour area filter algorithm accuracy assessment*

As previously noted, all reported results include the application of opening and removal of small components for noise reduction. Analysis of the four RLGS images, human *Not*I-*Eco*RV-*Hin*fI and *Asc*I-*Eco*RV-*Hin*fI and mouse master and FVB *Not*I-*Eco*RV-*Hin*fI, resulted in excellent correlations between the spots annotated by hand and those identified by the Contour Area Filtering. Of the 2500 to 3300 spots identified on the annotated human *Not*I, human *Asc*I the master mouse *Not*I and the FVB mouse *Not*I images, approximately 96-99% were correctly identified. In addition, the algorithm added small numbers of spots not seen in the annotated images. We marked a set of foreground pixels produced by Contour Area Filtering as an added spot if its pixels did not lie in any of the annotated spots on the master image. Similarly, we marked an annotated spot as missed if the pixels of that spot were not identified as foreground by our algorithm. Table 1 shows the breakdown of added and missed spots for each image. Since spots on the master images were marked only at their centers as judged by human analysis of the gels, some degree of subjectivity is *necessarily* a part of the determination of the extent of a spot.

Table 2 contains the breakdown of added spot errors into “Faint spots”, “Faint noise” and “Dark Noise” for the first three images. “Faint spots” dominated the added spot errors on the human *Not*I image while “Faint noise” dominated the errors on the human *Asc*I and mouse *Not*I images. The gels contain very few imperfections of physical marks and so few of the errors are “Dark noise”.

| genome | enzymes | faint spots | faint noise | dark noise |
|---|---|---|---|---|
| human | *Not*I-*Eco*RV-*Hin*fI | 33 | 3 | 13 |
| human | *Asc*I-*Eco*RV-*Hin*fI | 13 | 43 | 8 |
| mouse | *Not*I-*Eco*RV-*Hin*fI | 9 | 47 | 0 |

**Table 2.** Breakdown of spots added by Contour_Area_Filter which are not identified in annotated gels. Number of faint added spots, number of added spots caused by faint noise on the gel, and number of added spots caused by dark noise on the gel.

Missed spots are described in Table 3. Almost no "Faint spots" were missed for the human *Not*I image. "Faint spots" were 68% and 51% of the human AscI and mouse *Not*I missed errors, respectively. Nearly all of the missed spot errors occurred in the areas of the highest spot density. Such high density regions occur near the right edge and the upper left corner of each image. They can also occur in the neighborhood of largely enhanced spots. Each gel contains about a dozen such large spots generated by the repetitive ribosomal DNAs (rDNAs).

For Raman et. al.'s two benchmark protein gels (Raman, Cheung et al. 2002; http://www.umbc.edu/proteome), Contour Area Filtering found 89% of the 922 spots in Gel A and 93% of the 1350 spots in Gel B. It erroneously reported 10% additional spots in Gel A and 9% additional spots in Gel B. Problems were in high density spot areas and with noise near the boundaries of the gels.

*Comparison to ImageMaster*

The results of the ImageMaster analysis are presented in Table 1. The ImageMaster program does not distinguish separating foreground from background and segmenting foreground into individual spots in reporting its segmentation results. As in the comparison of Contour Area Filtering and the annotated images, we are not interested in the segmentation of the foreground into individual spots, only whether the spots are included in the image foreground. A spot was counted as appearing in a filtered image if its center lay in the foreground area of that image. ImageMaster did quite well compared to Contour Area Filtering of the human, *Not*I-*Eco*RV-*Hin*fI image. However, ImageMaster failed to correctly report some of the large saturated spots in this image which is a glaring error since those spots are so prominent. With larger window sizes, ImageMaster reported those spots, but then missed many of the smaller ones.

For the human, *Asc*I-*Eco*RV-*Hin*fI and the mouse, *Not*I-*Eco*RV-*Hin*fI images, ImageMaster missed considerably more spots compared to Contour Area Filtering. The missed spots were concentrated in the lower left region of the gels where spots were extremely faint. A higher sensitivity number should have been used for the mouse images, detecting more spots at the expense of false positives. On the other hand,

| genome | enzymes | faint | distinct |
|---|---|---|---|
| human | NotI-EcoRV-HinfI | 2 | 32 |
| human | *Asc*I-*Eco*RV-*Hin*fI | 43 | 21 |
| mouse | NotI-EcoRV-HinfI | 76 | 71 |

**Table 3.** Breakdown of spots missed by Contour_Area_Filter which are identified in annotated gels. Number of missed faint spots and number of missed distinct spots.

ImageMaster was already reporting more false positives than Contour Area Filtering for the other images and a higher sensitivity would simply have increased that number. This illustrates the need to modify the ImageMaster parameters based on the gel or even specific regions within the gel. We did try setting sensitivity to 9999, the maximum, for each gel, but this always produced tremendous numbers of spurious spots.

Table 4 in Rosengren et. al. (2003) contains spot detection results for Melanie, PDQuest, Progenesis and Z3 on protein Gels A and B but those results include the effects of spot segmentation. Nevertheless, our algorithm and error rates seem competitive with the best of the others. In addition, our algorithm can quickly detect spots on images with a much higher resolution than the images provided for Gels A and B. It is this higher resolution which allows us to report much better results on the RLGS gels.

Rosengren et. al. report 3-15 minutes running times for PDQuest on a 750 MHz processor and 15-180 minutes for Progenesis on a 2.0 GHz processor. Our algorithms running time of 10 seconds on images which are 10 to 30 times the size of Gels A and B is clearly superior.

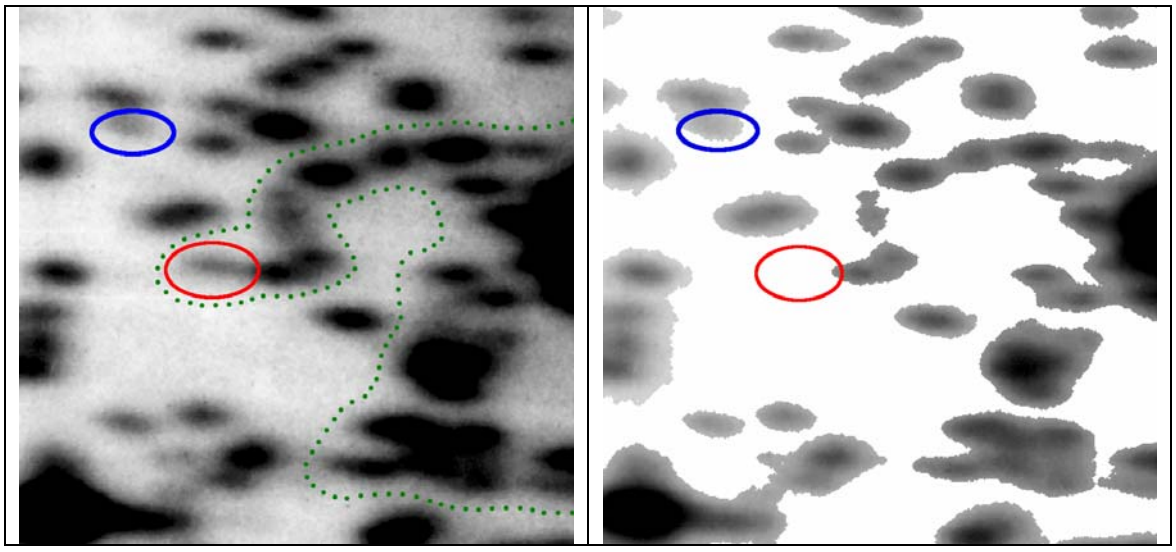


**Figure 16.** Circle near image center contains spot missing from contour area filtered image. Spot is part of a large cluster extending to the right and upward surrounded by dotted curve. Because the spot is so faint compared to other spots in the cluster, Contour Area Filtering marks it as background in the process of breaking up the cluster. Other spots with similar intensity such as the spot circled in upper left are detected by the algorithm because they are not part of a large dark cluster.

## Discussion

*Separating Foreground from Background Pixels*

Our algorithm reports correctly the foreground for 96-99% of the spots on 300 DPI RLGS images and 89-93% of the spots on lower resolution protein images, with errors concentrated in regions of high spot density. Background intensity varies greatly over these images and some of the spots are extremely faint. Good algorithm parameters depend upon spot size and density, not spot intensity, and thus do not need to be modified for each gel.

In all cases Contour Area Filtering agreed with the hand annotated gel more than ImageMaster software. As importantly, good filtering parameters in Contour Area Filtering depend upon the spot size and density, not upon the spot intensities. Since spot size and density are consistent between RLGS images, we don't modify the parameters for each gel. Good sensitivity values in ImageMaster depend greatly on spot intensity and are much more gel dependent. As previously noted, the ImageMaster system is designed and used for protein, not RLGS gels, where detection of faint spots is less difficult and less important. ImageMaster's intensity based algorithm has difficulty detecting faint spots because their intensities differ so slightly from the background.

The major weakness of our algorithm is in areas of high spot density where fainter spots may be obscured by stronger ones. High density areas are problematic for the algorithm since it uses the maximum contour area as a parameter. One can view this as the maximum size of a cluster of overlapping spots. If such a cluster has more pixels than

this maximum, the algorithm will remove fainter spots until it breaks the cluster apart (Figure 16). The distinct spots missed by Contour Area Filtering (Table 3) can all be attributed to this problem.

Areas of high spot density are either on the upper-left or the right side of the gels. In other areas, our algorithm gives 99% accuracy. Postprocessing or hybrid algorithms could perhaps be used to find faint spots in areas of high density.

Both our algorithm and ImageMaster missed a proportionately larger number of spots on the mouse master gel than on the other gels. These misses most likely stem from two factors. First, the mouse master gel is slightly anomalous since it uses a combination of DNA from three mouse strains, FVB, C57/BL6J, and 129/SV. Spots corresponding to DNA fragments which appeared in only one or two strains (strain specific polymorphisms) had significantly lower intensity than spots which appeared in all three strains making them more difficult to detect. Second, the density of spots on this combination mouse gel is higher than the other gels. The master mouse gel contains over 3200 identified spots, compared with under 2500 identified spots in the mouse FVB gel. Contour Area Filtering missed very few spots on that gel.

In the breakdown of Contour Area Filtering errors (Tables 2 and 3), Contour Area Filtering added 33 faint spots in the human NotI-EcoRV-HinfI annotated gel but only 13 and 9 such spots to the human *Asc*I-*Eco*RV-*Hin*fI and mouse *Not*I-*Eco*RV-*Hin*fI annotated gels. These represent errors in manual annotation of the gels rather than errors of Contour Area Filtering and illustrate a significant advantage over manual annotation and analysis. We hypothesize that the annotators of the human *Asc*I-*Eco*RV-*Hin*fI and mouse *Not*I-*Eco*RV-*Hin*fI gels were a bit more aggressive at identifying spots than the annotator of the human *Not*I-*Eco*RV-*Hin*fI gel. We also note that in the Contour Area Filtering miss errors, only two human *Not*I-*Eco*RV-*Hin*fI spots were classified as faint while 43 and 76 human *Asc*I-*Eco*RV-*Hin*fI spots and mouse *Not*I-*Eco*RV-*Hin*fI spots were classified as faint. This again supports the hypothesis that the annotators of the human *Asc*I-*Eco*RV-*Hin*fI and mouse *Not*I-*Eco*RV-*Hin*fI gels were more aggressive.

The number of added spots errors caused by dark noise are very small in the human gels and zero in the mouse gel. These errors are caused by physical imperfections in the gels, often caused by accidents in handling the physical gel and are immediately obvious as errors. The Master gels were chosen as representative of a set of gels exactly because they had excellent quality. Many gels have poorer quality and more added spot errors caused by physical imperfections.

Changes in the Contour Area Filter parameter M will lead to changes in the classification of some light pixels. In addition, light pixels near isolated spots are more likely to be classified as foreground than light pixels near spot clusters. Since we segment all pixels in foreground into individual spots, the total number of pixels assigned to a spot is sensitive to parameter changes and nearby spots. However, spot comparison and measurement is done based on spot's maximum intensity or spot volume. Light pixels make little contribution to maximum intensity or spot volume, so the difficulty in

classification of light pixels will not significantly affect the comparison or measurement of spots.

RLGS gels have three properties which make them particularly suited for Contour Area Filtering. First, all pixels of an RLGS spot are darker than the local background intensity. Contour Area Filtering will not work well for non-homogeneous objects which have some parts lighter and some parts darker than the local background. Second, reliable upper and lower bounds can be placed on the size of RLGS spots. Without these bounds, Contour Area Filtering will oversegment large objects or treat small objects as noise. Third, the background does not vary significantly around a single RLGS spot. Thus the contour lines around that spot can be used to segment the spot.

*Segmentation of Foreground into Individual Spots*

In addition to using Contour Area Filtering for separating foreground and background, we also use the algorithm for segmenting foreground pixels into individual spots. Each RLGS spot has a "center", usually consisting of a relatively small number of high intensity pixels. If the area of a contour is approximately this size, then this contour surrounds a center. We call Contour Area Filtering with the parameter M set to the maximum spot center size. We use 100 pixels for the maximum spot center size on 5100x4200 RLGS images.

Some spots, called saturated spots, contain a large set of pixels which are all very dark. These pixels should be grouped together as a single spot center even thought this spot center size is much larger than average. If all the pixels in a contour have approximately the same dark intensity, then that contour forms a single center, even though the contour area may be very large. We modified Contour Area Filtering to identify such contours and avoid oversegmentation of saturated spots.

Segmentation into individual spots is an extremely challenging problem. Spots may overlap in ways which make it difficult or even impossible to tell by visual inspection whether they are two or one. In fact, biomedical researchers segment RLGS gels by comparing multiple gels or comparing gels with a segmented "master" gel. Contour Area Filtering makes many more errors when applied to foreground segmentation than when used to separate foreground from background. Foreground segmentation errors are almost always undersegmentation errors. Thus, Contour Area Filtering can be used as an initial foreground segmentation step, but further splitting of spots is required.

We experimented with an expectation maximization (EM) algorithm for improving the foreground segmentation produced by Contour Area Filtering, but with limited success. Instead we use a previously segmented "template" to add or split spots in a target gel. The template gel will usually, although not necessarily, be a master gel which has been meticulously segmented by biomedical researchers by hand. This master gel is repeatedly used to segment all other gels from the same genome and enzyme combination. A complete description of the algorithm and experimental results will be reported elsewhere.

Once all spot centers have been properly identified, spot boundaries must be constructed for each spot. For isolated spots, we simply use the boundary of the connected foreground component containing the spot center. This boundary may not have the oval shape characterizing most spots, and is sensitive to changes in the Contour Area Filter parameter M. As previously noted, spot comparison and measurement is done based on spot's maximum intensity or spot volume. Changes in the boundary shape or size of isolated spots have little effect on these measurements.

For clusters of overlapping spots, multiple spots contribute to pixel intensities in the region of overlap. Thus it is extremely difficult to determine appropriate spot boundaries or to accurately calculate spot intensities. Algorithms which assign each pixel to a single spot can greatly underestimate the spot volume of overlapping spots. The protein gel analysis software Phoretix (http://www.phoretix.com) models individual spots as a Gaussian distribution of intensities and then measures the maximum and volume of this distribution. This approach addresses the problem of overlapping spots but introduces its own inaccuracies in the simplified modeling. Accurate estimation of RLGS spot boundaries and spot intensities for overlapping spots is one subject of our ongoing research.

## Acknowledgements

This research was supported by NIH grant 3RE01DE13123-02S1. We also thank Dr. Haifeng Wu for use of the Nonlinear Dynamics ImageMaster software in his lab. We thank Dr Li Yu, The Ohio State University, for providing the mouse Master image for this work.

## References


Appel, R. D., P. M. Plagi, et al. (1997). "Melanie II -- a third-generation software package for analysis of two dimensional electrophoresis images: II. Features and user interface." Electrophoresis **18**: 2724-2734.

Appel, R. D., J. R. Vargas, et al. (1997). "Melanie II -- a third generation software package for analysis of two dimensional electrophoresis images: II. Algorithms." Electrophoresis **18**: 2735-2748.

Asakawa, J., R. Kuick, et al. (1995). "Quantitative and qualitative genetic variation in two-dimensional DNA gels of human lymphocytoid cell lines." Electrophoresis **16**(2): 241-52.

Bieniek, A. and A. Moga (2000). "An efficient watershed algorithm based on connected components." Pattern Recognition **33**: 907-916.

Clarke, L. P., R. P. Velthuizen, et al. (1995). "MRI segmentation: Methods and applications." Magnetic Resonance Imaging **13**(3): 343-368.

Cormen, T. H., C. E. Leiserson, et al. (2001). Introduction to Algorithms, MIT Press.

Costello, J. F., M. C. Frühwald, et al. (2000). "Aberrant CpG island methylation has non-random and tumor type specific patterns." Nature Genetics **25**: 132-138.

Costello, J. F., M. C. Fruhwald, et al. (2000). "Aberrant CpG-island methylation has non-random and tumour-type-specific patterns." Nat Genet **24**(2): 132-8.

Dai, Z., R. R. Lakshmanan, et al. (2001). "Global methylation profiling of lung cancer identifies novel methylated genes." Neoplasia **3**(4): 314-23.

Dawant, B. M. and A. P. Zijdenbos (2000). Image Segmentation. Medical Imaging: Medical Image Processing and Analysis. M. Sonka and J. M. Fitzpatrick, The International Society for Optical Engineering. **2:** 71-128.

Fu, G., S. A. Hojjat, et al. (2004). "Integrating watersheds and critical point analysis for object detection in discrete 2D images." Medical Image Analysis **8**: 177-185.

Hatada, I., Y. Hayashizaki, et al. (1991). "A genomic scanning method for higher organisms using restriction sites as landmarks." Proc Natl Acad Sci U S A **88**(21): 9523-7.

http://expasy.ch/melanie.

http://gelmatching.inf.fu-berlin.de.

http://www.amershambiosciences.com.

http://www.bio-rad.com.

http://www.nonlinear.com.

http://www.phoretix.com.

Li, H., L. Myeroff, et al. (2003). "SLC5A8, a sodium transporter, is a tumor suppressor gene silenced by methylation in human colon aberrant crypt foci and cancers." Proc Natl Acad Sci U S A **100**: 8412-8417.

Mahon, P. and P. Dupree (2001). "Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full." Electrophoresis **22**: 2075-2085.

Okazaki, Y., H. Okuizumi, et al. (1995). "An expanded system of restriction landmark genomic scanning (RLGS Ver. 1.8)." Electrophoresis **16**(2): 197-202.

Raman, B., A. Cheung, et al. (2002). "Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie." Electrophoresis **23**: 2194-2202.

Rapantzikos, K., M. Zervakis, et al. (2003). "Detection and segmentation of drusen deposits on human retina: Potential in the diagnosis of age-related macular degeneration." Medical Image Analysis **7**: 95-108.

Rogowska, J. (2000). Overview and Fundamentals of Medical Image Segmentation. Handbook of Medical Imaging. I. N. Bankman, Academic Press**:** 87-106.

Rosengren, A. T., J. M. Salmi, et al. (2003). "Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels." Proteomics **3**: 1936-1946.

Sebastian, T. B., H. Tek, et al. (2003). "Segmentation of carpal bones from CT images using skeletally coupled deformable models." Medical Image Analysis **7**: 21-45.

Smiraglia, D. J., M. C. Frühwald, et al. (1999). "A New Tool for the Rapid Cloning of Amplified and Hypermethylated Human DNA Sequences from Restriction Landmark Genome Scanning Gels." Genomics **58**(3): 254-262.

Smiraglia, D. J. and C. Plass (2002). "The study of aberrant methylation in cancer via restriction landmark genomic scanning." Oncogene **21**: 5414-5426.

Smiraglia, D. J., L. J. Rush, et al. (2001). "Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies." Hum Mol Genet **10**(13): 1413-1419.

Sternberg, S. R. (1983). "Biomedical Image Processing." IEEE Computer: 22-34.

Sugahara, Y., S. Akiyoshi, et al. (1998). "An automatic image analysis system for RLGS films." Mamm Genome **9**(8): 643-51.

Takahashi, K., M. Nakazawa, et al. (1997). DNAinsight: An Image Processing System for 2-D Gel Electrophoresis of Genomic DNA. Genome Inform. Ser. Workshop Genome Inform.

Takahashi, K., M. Nakazawa, et al. (2001). DNAinsight: A Web Based Image Processing System for Large Scale RLGS Analysis. Genome Inform. Ser. Workshop Genome Inform.

Takahashi, K., M. Nakazawa, et al. (1998). Fully-Automated Spot Recognition and Matching Algorithms for 2-D Electrophoretogram of Genomic DNA. Genome Inform Ser Workshop Genome Inform.

Takahashi, K., M. Nakazawa, et al. (1999). Automated Processing of 2-D Gel Electrophoretograms of Genomic DNA for Hunting Pathogenic DNA Molecular Changes. Genome Inform. Ser. Workshop Genome Inform.

Yao, W., P. Abolmaesumi, et al. (2005). "An estimation/correction algorithm for detecting bone edges in CT images." IEEE Transactions on Medical Imaging **24**(8): 997-1010.

Yoshikawa, H., K. Matsubara, et al. (2001). "SOCS-1, a negative regulator of the JAK/STAT pathway, is silenced by methylation in human hepatocellular carcinoma and shows growth- suppression activity." Nat Genet **28**(1): 29-35.