Applied and Theoretical Electrophoresis (1991), 2, 141–149

Automatic detection of noisy spots in two-dimensional Southern blots International Electrophoresis Society Meeting, Washington DC, March 16–19, 1991

Peter F. Lemkin[1] & Peter K. Rogan[2]

[1]*NCI-DCBDC, IPS/LMMB/FCRDC, Frederick, Maryland 21702;* [2]*Department of Pediatrics, Division of Genetics, Hershey Medical Center, Hershey, Pennsylvania 17033, USA*

One of the problems of automatically quantitating 2D DNA gels, is clearly detecting spots visualized by Southern hybridization blots using DNA probes (Rogan *et al.* in this conference). Spots appear noisy due to multiple transfer steps. If one applies standard 2D PAGE protein gel spot segmentation methods, spots fragment due to highly textured image noise and a weak radioautograph signal and thus are poorly detected. We have observed that these spots are all of a minimum size. Therefore an image processing filter which both takes minimum spot size into account and has immunity to image texture-type noise should be able to reliably detect this class of spots. The 'Busse' Laplacian filter used in the GELLAB-II system, is a modification of the standard (1 −2 1) digital approximation of the Laplacian. In the Busse Laplacian, the sampling interval is n pixels ($n > 1$) instead of 1. In addition, 3 × 3 averaged 'super pixel' values are used instead of single pixels for each element of the Laplacian convolution. This gives the needed noise immunity by filtering out the high spatial frequency image noise while preserving the low spatial frequency character of the spots. We have used this filter successfully on 2D DNA Southern blot image data.

Keywords: two-dimensional gel electrophoresis; Southern blots; image processing; GELLAB.

Introduction

While the separation of genomic DNA by agarose gel electrophoresis can resolve a broad range of restriction fragment sizes, the gel band widths of individual fragments can be quite pronounced. This can limit the resolution of similarly sized fragments in a digest. By sequentially digesting genomic DNA with two different restriction enzymes, similarly sized fragments in the first digest can be distinguished on the presence or absence of internal restriction sites by redigestion with a second enzyme. This approach has been coupled to electrophoresis onto a two-dimensional (2D) agarose grid (Rogan *et al.*, 1991a; Mietz & Kuff, 1990). However, the second electrophoresis step somewhat dilutes homogeneous bands, resulting in broad spot patterns (Holmes & Stellwagon, 1990). One of the problems associated with automatically quantitating these spots is to clearly detect them.

Correspondence: P. Lemkin, NCI-FCRDC, Building 469, Room 150, Box B, Frederick, MD 21702, USA

Multiple fragments are visualized as spots in 2D DNA gels by Southern analysis with probes containing repetitive genomic sequences. Most of the genome can be sampled by sequential hybridization of different repetitive sequence probes to the same filter. The total database, therefore, can consist of a large number of unique spots. The ability to discriminate between proximate, nonidentical spots is a critical requirement of this form of genome analysis. This lets us construct a computer database consisting of spots from many probes reflecting the expression of these genes as a function of experimental conditions. This could also lead to an automatic system for 2D DNA gel analysis based to a large extent on the GELLAB-II system for 2D PAGE gels (Lemkin & Lester, 1989; Lemkin, 1989). Although these low resolution 2D DNA gels may not appear to warrant a sophisticated computer analysis, the construction of a reliable multiple gel database requires using an accurate spot segmentation tool. Studies of multiple gels under different experimental conditions with several probes generates a large enough set of spots to require the use of a spot management database (Rogan *et al.*, 1991).

This issue stimulated the development of a spot segmentation procedure which was sensitive enough to differentiate closely-spaced loci, but, also could recognize the diffuse spot shapes produced by the agarose gel electrophoresis/Southern transfer procedure.

Standard 2D protein electrophoretic spot segmentation methods do not work as well for 2D DNA gels because spots fragment due to highly textured image noise and weak radioautograph signals. We have observed that these spots are all of a minimum size. We present a robust image processing filter which has immunity to this type of image texture-type noise allowing the reliable detection of this class of spots. It is currently used in the GELLAB-II system and is a modification of the 'Busse' Laplacian filter used in the GELLAB-I system (Lipkin & Lemkin, 1980). The original Busse filter was developed by Hans Busse (Univ. Kiel) when working with the DEC10 based GELLAB-I system.

2D DNA gel spot quantitation

The potentially large amount of data to be analyzed motivated the development of automated computer analysis methods to find and quantitate spots in scanned 2D DNA gel images. The general problem of finding and quantitating spots in 2D gels has been solved a number of different ways over the years for 2D PAGE gels (Anderson *et al.*, 1981; Garrels, 1979;

Lipkin & Lemkin, 1980; Vincens *et al.*, 1986; Vo *et al.*, 1981 and others). The GELLAB-II system is an enhanced version of our earlier work (Lipkin & Lemkin, 1980).

As mentioned, one of the problems of this particular type of 2D gel is in clearly detecting spots in Southern blots using DNA probes. Spots appear noisy due to the multiple transfer steps and broad gel-band widths. Noise appears as superimposed small random texture elements within spots and small 'shot' noise superimposed over the entire gel. Spots are characterized as relatively large, fuzzy, blotchy noise regions as compared to spots in PAGE gels which are better defined. Since the Southern procedure produces a minimum size spot, this minimum size threshold makes it possible to discriminate valid signals from noise spots.

Figure 1a shows a typical 2D DNA gel scanned with a DataCopy CCD camera. Note the poor spot definition with respect to spot blotchiness and noisy background in this unprocessed gel image. Unlike 2D protein PAGE gels where polypeptide spots often appear more

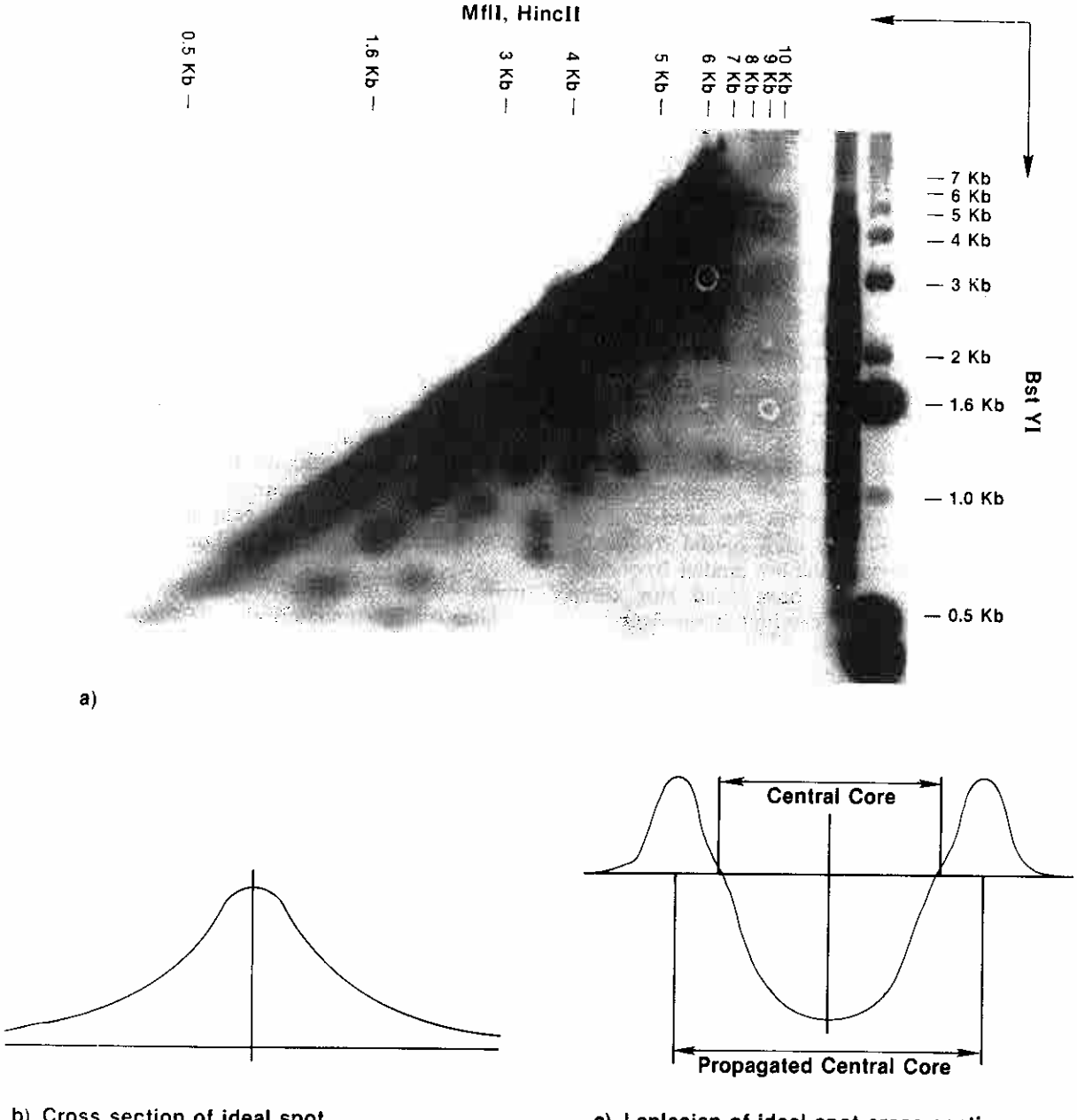


Figure 1 (a) a typical 2D DNA gel: *Saccharomyces cerevisiae* strain LP2619-1a (α, *ura3*, *GAL*$^+$) transformed with the *YSH-19* vector carrying the *dam* methylase gene (courtesy of S. Hattmann). Cells were propagated in glucose prior to galactose induction of the methylase gene. Genomic DNA was then isolated and digested with *MfII* and *HincII*. Gel slabs containing electrophoretic restriction fragments were redigested *in situ* with *BstYI* and embedded perpendicularly in a second agarose gel. Southern analysis was carried out using a cloned radiolabeled probe containing a solo long terminal repeat element, a δ sequence. The gel was scanned with a DataCopy 612F camera 55 mm lens at *f*4, with 186 microns/pixel. (b) Illustration of 1-D cross section on an 'ideal' spot as a monotonically increasing function of OD. (c) 2nd derivative of this monotonic function. The region with negative values is called the *central core*. The region on the outside of the central core (CC) is propagated until it reaches the extent of the positive peaks of the side lobes. This *propagated central core* (PCC) region, computed in two dimensions, is then effectively used as a mask for quantitating that spot. Noise present in real data makes finding the CC and PCC difficult if false peaks are present

tightly focused and have an easily recognized monotonically increasing cross-sectional density (Lipkin & Lemkin, 1980), these are embedded in noise and spots appear as texture conglomerates of small objects.

Materials and methods

If one applies standard 2D PAGE protein gel spot segmentation methods such as used in GELLAB-II, spots fragment due to the highly textured image noise and weak radioautograph signal and thus are poorly detected. Therefore an image processing filter which both takes minimum spot size into account and has immunity to image texture-type noise should be able to reliably detect this class of spots. We present three classes of image processing filters designed to increasingly take this type of noise into account: a simple Laplacian filter, a single-pixel Busse filter, and a super-pixel Busse filter.

Why the standard Laplacian filter failed

The GELLAB-II spot quantitation program sg2gii (Lemkin & Lipkin, 1981; Lemkin & Lipkin, 1983) is used for automatically segmenting (finding) and quantitating spots in 2D PAGE gels. The program finds spots and then quantitates them by integrating pixel optical density within the area of the spot. Under optimal conditions, the cross-sectional density of a spot appears as a monotonically increasing function as illustrated in Figure 1b. The Laplacian or second derivative of this function is shown in Figure 1c. We define the central region of negative values for digital approximations to both partial second derivatives $\Delta^2 x$ and $\Delta^2 y$, with respect to x and y directions, as the *central core* region (c.f. Figure 1c). This is the reason we want to use the Laplacian to help analyze the image.

The digital approximations are derived as follows. In one dimension the first difference, Δg_i, at point i for image pixel optical density value g_i is defined by

$$\Delta g_i = g_{i+1} - g_i$$

Then, the second difference $\Delta^2 g_i$ is defined by

$$\Delta^2 g_i = \Delta g_{i+1} - \Delta g_i$$

This simplifies to

$$\Delta^2 g_i = g_{i+2} - 2g_{i+1} + g_i$$

Then the partial second derivatives of $g_{x,y}$ are

$$\Delta^2{}_x g_{x,y} = g_{x+1,y} - 2g_{x,y} + g_{x-1,y}$$

and

$$\Delta^2{}_y g_{x,y} = g_{x,y+1} - 2g_{x,y} + g_{x,y-1}$$

During the segmentation process, the image is first smoothed using a Gaussian convolution filter of size K (Lipkin & Lemkin, 1980) and the resulting smoothed image is used when computing the digital approximation of the Laplacian of that image. We store the Laplacian direction and magnitude values in two additional images. All Laplacian direction image pixels $D_{x,y}$ are set to 1 if the Laplacian values are negative in both X and Y; otherwise they are set to 0. This directional image defines the initial central core regions of a gel. The central core regions are propagated to adjacent pixels until they reach the maximum value of the Laplacian magnitude image. These final regions are called the *propagated central core* region (c.f. Figure 1c) and, after some minor corrections, define the extent of the spot to be quantitated. Then at each (x, y) in the smoothed image, the direction $D_{x,y}$ is defined by

$$D_{x,y} = \begin{cases} 1 \text{ if } (\Delta^2{}_x(g_{x,y}) < 0) \text{ and } (\Delta^2{}_y(g_{x,y}) < 0), \\ 0 \text{ otherwise} \end{cases}$$

Single spots appear to fragment when the calculated Laplacian image is too noisy. This occurs when the *central core* regions are poorly defined. We have found that Southern analysis generates noisy spots that are characterized by blotchy texture. These blotchy spots tend to segment as sets of sub-spots. Laplacian functions are very sensitive to this type of textured noise prompting the development of this new Laplacian filter.

The standard GELLAB-II Laplacian digital filter consists of either a 3 × 3 or 5 × 5 partial second derivative convolution filters illustrated in Figures 2a and 2b. This convolution filter is applied in turn to each pixel in the smoothed image and consists of the sum of products of pixels from the convolved image with pixels from the filter.

The 'simple-pixel' Busse Laplacian filter

In the original Busse Laplacian, the sampling interval is N pixels where $(n > 1)$ instead of 1 and is illustrated on Figure 3a. This sampling of the smoothed image ignores local, high spatial-frequency noise in adjacent pixels and so is less likely to counter-indicate trends in the Laplacian. However, even with a larger sampling grid, this type of data is still sometimes too noisy to use for use with this type of filter.

The 'super-pixel' Busse filter

To further reduce spot fragmentation, we added even more smoothing to the Laplacian function. Instead of computing the Laplacian with values of single pixels, an averaged pixel value is used at each sample point in the grid. We call this averaged pixel, a 'super-pixel'. A 3 × 3 pixel averaging region seems to be adequate. This tends to greatly limit the high frequency noise of the grid pixels used in the calculation. Figure 3b illustrates the super-pixel filter.

This then gives additional noise immunity against spot fragmentation by filtering out the high spatial frequency image noise while preserving the low spatial frequency character of the spots.

Additional segmenter optimizations

Two other modifications were made to the GELLAB-II segmenter which speeds up spot identification and definition. One is to conditionally reject any potential spot if the central core area is less than some specified

a) 3x3 Laplacian Convolution filter:

$$\Delta^2 x = \begin{matrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{matrix}$$

$$\Delta^2 y = \begin{matrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{matrix}$$

b) 5x5 Laplacian Convolution filter:

$$\Delta^2 x = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & -4 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$\Delta^2 y = \begin{matrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{matrix}$$

Figure 2 Partial second derivative convolution filter computations for (**a**) 3 × 3 or (**b**) 5 × 5 pixel neighborhoods. Each pixel in a neighborhood of the central pixel is multipled by the matrix element and the results summed. The calculation is performed at each pixel in the smoothed gel image in order to compute $\Delta^2 x$, $\Delta^2 y$

lower limit. This is useful in a situation like 2D DNA gels where the spots are all expected to be quite large and approximately the same area. The other modification is to only attempt to propagate pixels on the outside of the central core region. Interior central core pixels can never propagate since their neighbors are all central core pixels. This avoids needless calculations on interior central core pixels which do not propagate in any case.

Results

The segmented images using the previous 2D PAGE gel segmentation algorithm *without* the Busse filter are shown in Figure 4. Spots appear fragmented, too small, and do not extend to what we observe to be the boundaries of the spots. The Laplacians used were the 3 × 3 and 5 × 5 illustrated in Figure 2. Slightly more accurate, fuller spot shape results were found using the 13 × 17 Gaussian smoothing filter rather than the 7 × 7, although fewer spots were found and light fuzzy spots were missed. In all cases, spots were poorly segemented, although the 5 × 5 Laplacian gave better results.

The results with the *single-pixel* Busse filter in Figure 5 are better than with the non-Busse Laplacians. However, the segmented image suffered to a lesser degree to fragmentation, insensitivity to light spots and their incomplete definition. Differences between the 7 × 7 and 13 × 17 Gaussian smoothing filters were not significant.

The new *super-pixel* Busse filter in Figure 6 adequately segments these spots over a wide range of mean densities and textures. Notice that the fragmentation has disappeared and all spot regions are detected including light regions missed in Figure 5. Although the spot regions are not as round as they appear on the film, they adequately represent the spots. It is relatively robust with respect to grid size since varying grid size n from 5 to 9 does not make much difference (not shown). The number of light spots detected and the extent of those detected decreased slightly with n set to 5. However, using a 13 × 17 Gaussian smoothing filter reduced the number of very light spots – similar to what was observed for the other two types of Laplacians. The image segmented using the optimal filter is shown in Figure 6a and 6b, and uses the Busse super-pixel filter with a grid spacing of 7 and smoothed with the 7 × 7 Gaussian filter.

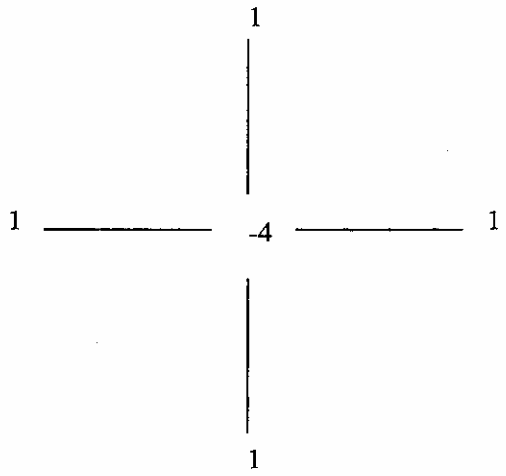
Discussion

This filter successfully segments spots from 2D DNA Southern blot image data which because of high image noise do not work well with standard 2D PAGE gel quantitation techniques. Standard techniques use Laplacian filters with a grid spacing of 1 pixel.

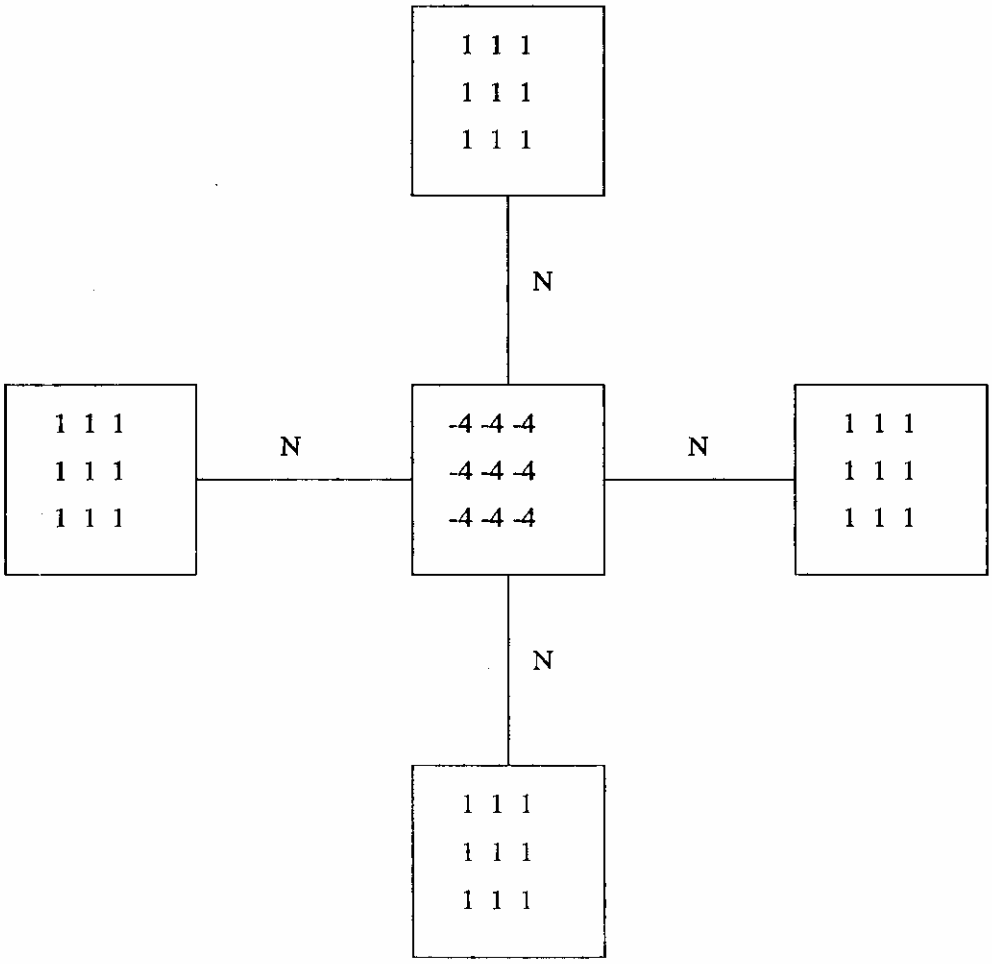
The selection of the Laplacian filter seems to be more important than that of the initial Gaussian smoothing for this highly texture type of data. Because of the better noise immunity, the super-pixel Busse Laplacian helps avoid sensitivity to noise and lets the segmenter find the correct extent of fuzzy spots correctly. It does this without merging almost any adjacent spots. The Laplacian sampling grid size n should be tuned to the average spot size, although this seems to not be that critical. Smoothing with the initial Gaussian filter tends to eliminate some of the light fuzzy spots so care should be taken to avoid this. Since additional smoothing is done using the super-pixels, the initial smoothing is not so critical. Another implementation of this filter could be to average the Gaussian smoothed data in one smoothing filter step which then would make the super-pixel filter equivalent to the simple-pixel Busse filter.

Obviously, in any distribution of spots there may exist some highly noisy light overlapping spots which might not be adequately detected although this filter greatly minimizes this problem. As with 2D PAGE systems, it should be possible to merge different exposures of the same gel to resolve very light spots in the presence of very dark spots – although we have not done so with this data.

Being able to reliably segment spots in 2D Southern gels is an important step in constructing multiple-probe computer databases. The ability to discriminate between proximate, nonidentical spots is an important component of our on-going effort to develop new techniques for genome analysis.



a) 3x3 Simple-pixel Busse Laplacian convolution filter.



b) 3x3 Super-pixel Busse Laplacian convolution filter.

Figure 3 Busse Laplacian second derivative convolution filters for a sampling grid spacing of n, (**a**) *simple-pixel* convolution filter, (**b**) *super-pixel* convolution filter for 3 × 3 super-pixel. The spacing is taken between the centers of the super-pixels

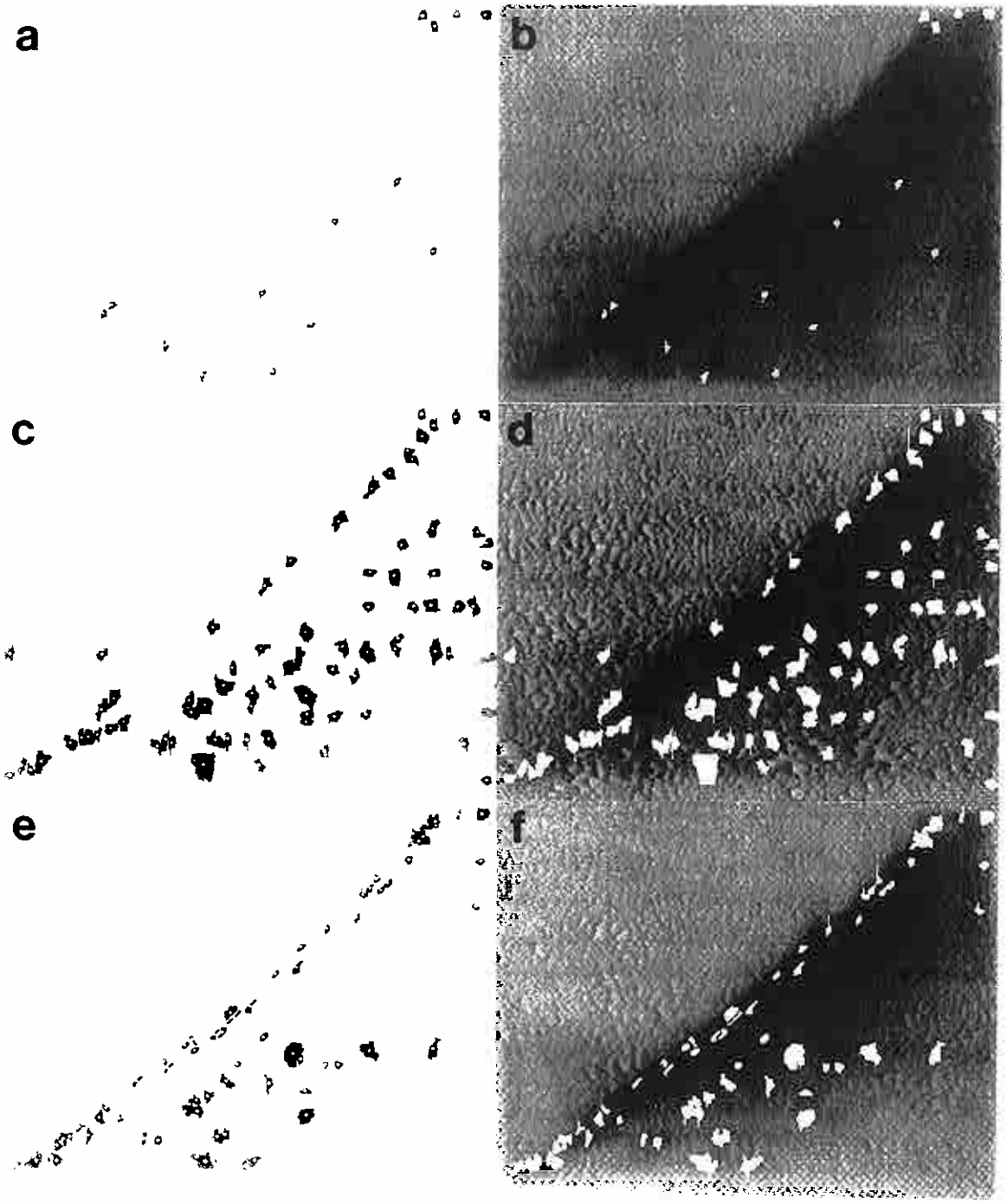


Figure 4 shows the gel in Figure 1 segmented with the standard 2D PAGE gel segmentation algorithm. We illustrate gel segmentation errors by subtracting spots found by the segmenter (left images) from the *unprocessed* highly textured gel image (c.f. Figure 1) resulting in images containing spots which were missed (right image). Spots were rarely merged during segmentation, but when they were, they can be seen with a little effort by comparing a segmented image with the original gel in Figure 1. (**a**) Spots found using 7 × 7 Gaussian smoothing with 3 × 3 Laplacian, (**b**) the original gel less the segmented spots in (**a**) showing those spots which were missed, (**c**) and (**d**) 7 × 7 Gaussian smoothing with 5 × 5 Laplacian, (**e**) and (**f**) 13 × 17 Gaussian smoothing with 5 × 5 Laplacian. Minimum area sizing was 75 square pixels for (**c**) and (**d**), the rest were 35 in order to identify spots which were poorly segmented. The centroids of segmented spots are indicated by a white dot in (**a**), (**c**) and (**e**). These figures, as well as Figures 5 and 6, were prepared using a laser printer. This process occasionally added some ridge and valley artifacts not visible in the photograph on the gel in Figure 1

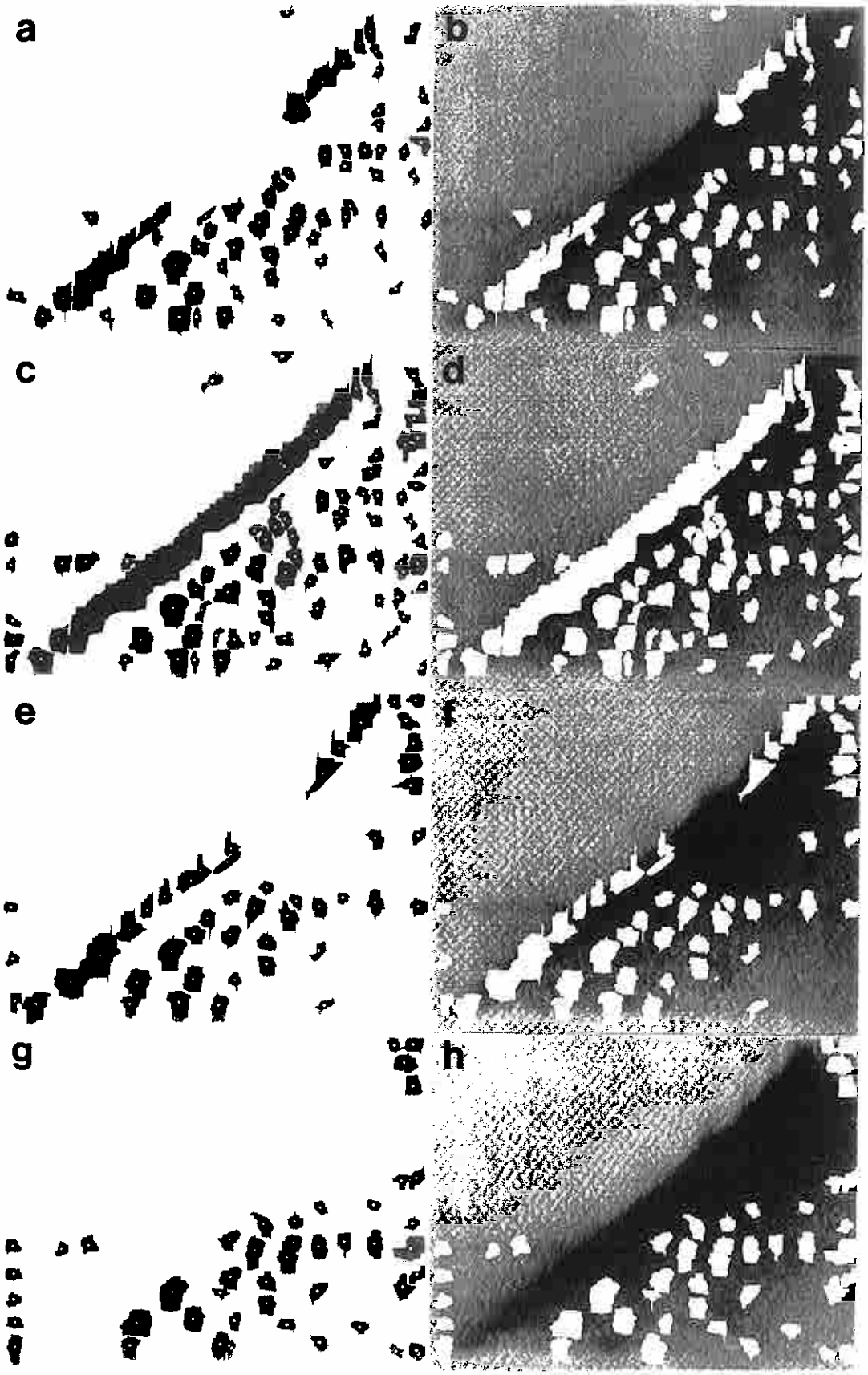


Figure 5 shows the same gel segmented with the original *single pixel* Busse filter using pixel values from a sampling grid of size $n \times n$. Minimum area was 150 square pixels. (**a**) Spots found using 7 × 7 Gaussian smoothing with 7 × 7 grid Busse Laplacian, (**b**) the original gel less the segmented spots in (**a**) showing those spots which were missed. (**c**) and (**d**) 7 × 7 Gaussian smoothing with 9 × 9 grid Busse Laplacian, (**e**) and (**f**) 13 × 17 Gaussian smoothing with 7×7 grid Busse Laplacian, (**g**) and (**h**) 13 × 17 Gaussian smoothing with 9 × 9 grid Busse Laplacian. The centroids of segmented spots are indicated by a white dot in (**a**), (**c**), (**e**) and (**g**)

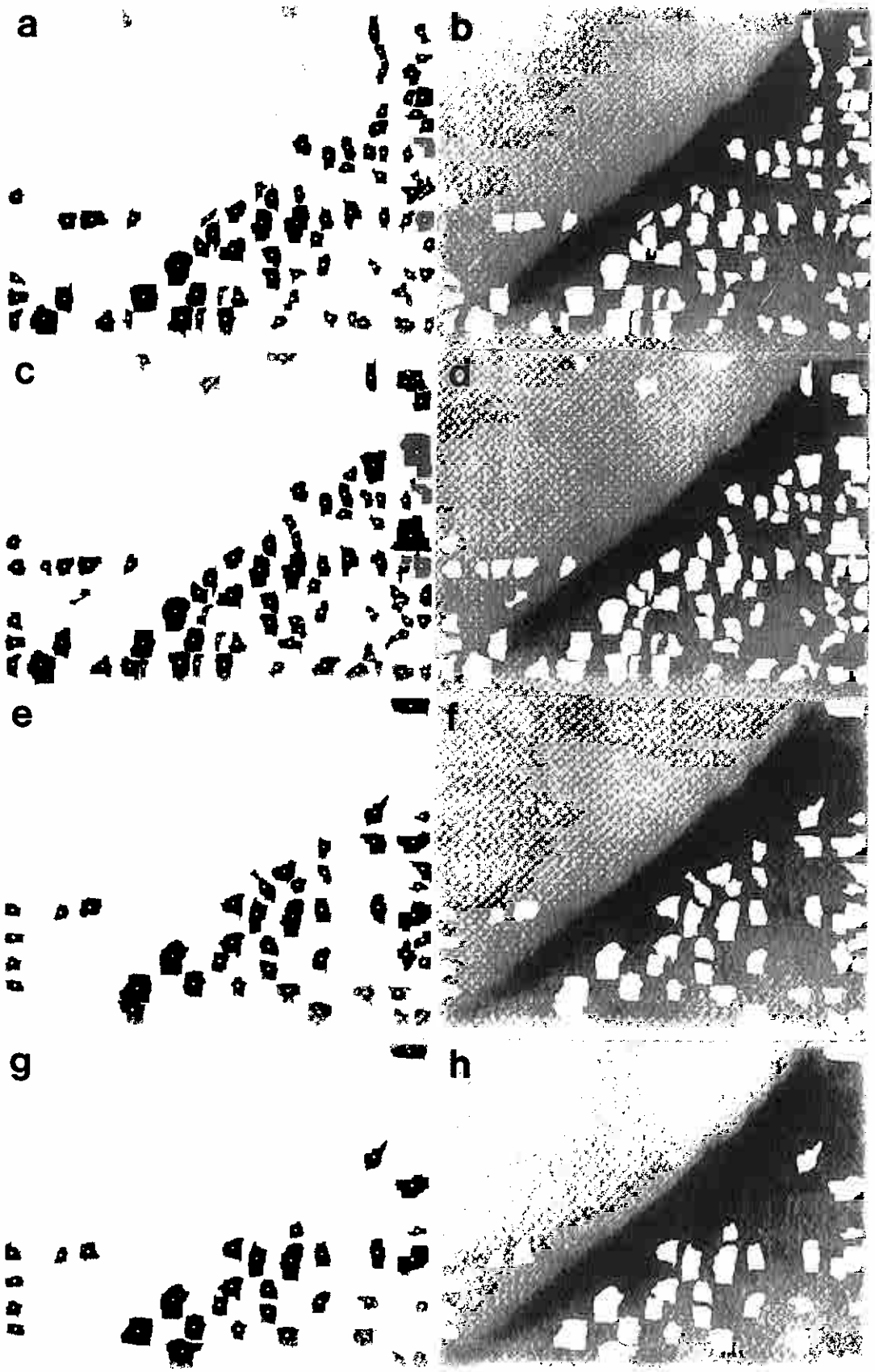


Figure 6 shows some gels segmented with the new *super-pixel* Busse filter using pixel values from a sampling grid of size $n \times n$. Minimum area was 150 square pixels. (**a**) Spots found using 7 × 7 Gaussian smoothing with 7 × 7 grid Busse Laplacian, (**b**) the original gel less the segmented spots in (**a**) showing those spots which were missed. (**c**) and (**d**) 7 × 7 Gaussian smoothing with 9 × 9 grid Busse Laplacian, (**e**) and (**f**) 13 × 17 Gaussian smoothing with 7 × 7 grid Busse Laplacian, (**g**) and (**h**) 13 × 17 Gaussian smoothing with 9 × 9 grid Busse Laplacian. The centroids of segmented spots are indicated by a white dot in (**a**), (**c**), (**e**) and (**g**)

References

Anderson, N.L., Taylor, J., Scandora, A.E., Coulter, B.P. & Anderson, N.L. (1981). The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.*, **27**, 1807–1821.

Garrels, J.I. (1979). Two dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.*, **254**, 7961–7977.

Holmes, D.L. & Stellwagon, N.C. (1990). The electric field dependence of DNA mobilities in agarose gels: A re-investigation. *Electrophoresis*, **11**, 5–15.

Lipkin, L.E. & Lemkin, P.F. (1980). Database techniques or multiple PAGE (2D gel) analysis. *Clin. Chem.*, **26**, 1403–1413.

Lemkin, P. & Lipkin, L. (1981). GELLAB: A computer system for 2D gel electrophoresis analysis. I. Segmentation and preliminaries. *Comp. Biomed. Res.*, **14**, 272–297.

Lemkin, P.F. & Lipkin, L.E. (1983). Database Techniques for 2D Electrophoretic Gel Analysis. In: Geisow, M., Barrett, A. (eds), *Computing in Biological Science*. Elsevier/North-Holland, 181–226.

Lemkin, P.F. & Lester, E.P. (1989). Database and search techniques for 2D gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis*, **10**, 122–140.

Lemkin, P.F. (1989). GELLAB-II, A workstation based 2D electrophoresis gel analysis system. In: Endler, T., Hanash, S. (eds). *Two-Dimensional Electrophoresis*. VCH Press: W. Germany, pp. 53–57.

Mietz, J. & Kuff, E. (1990). Tissue and strain specific pattern of endogenous proviral hypomethylation analyzed by two dimensional gel electrophoresis. *Proc. Natl. Acad. Sci. USA*, **87**, 2269–2273.

Rogan, P.K., Klar, A., Singh, J., Strathern, J.N. & Lemkin P. (1991). Identification of cell-type specific chromosomal loci by *in vivo* methylation. *Abstract, Int. Meet. Electrophoresis*, Soc. Wash. DC, Mar 19–21, 1991b.

Rogan, P.K., Lemkin, P.L., Klar, A., Singh, J., Strathern, J.N. (1991a). Two-dimensional agarose gel electrophoresis of restriction-digested genomic DNA. *Methods*, in press.

Vincens, Paris, N., Pujol, J-L., Gaboriaud, C., Rabilloud, T., Pennetier, J-L., Matherat, P. & Tarroux, P. (1986). HERMeS: A second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis*, **7**, 357–367.

Vo, K-P., Miller, M.J., Geiduschek, E.P., Nielsen, C. & Xuong, N.H. (1981). Computer analysis of two-dimensional gels. *Anal. Biochem.*, **112**, 258–271.