MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0279136 1

ORNL/TM-10591

ornl

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

A Comparison of Algorithms for the Analysis of Images Generated by Two-Dimensional Electrophoresis

R. C. Mann
D. J. Riese
B. K. Mansfield
J. K. Selkirk

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
CIRCULATION SECTION
4500N ROOM 175
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this report, send in name with report and the library will arrange a loan.

OPERATED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A03; Microfiche A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-10591

Engineering Physics and Mathematics Division

A COMPARISON OF ALGORITHMS FOR THE ANALYSIS OF IMAGES GENERATED BY TWO-DIMENSIONAL ELECTROPHORESIS

R. C. Mann
D. J. Riese*
B. K. Mansfield+
J. K. Selkirk#

Date Published: November 1987

*ORAU Undergraduate Research Participant, Summer 1986, from Wabash College, Crawfordsville, Indiana.
+Biology Division.
#National Toxicology Program, NIEHS, Research Triangle Park, North Carolina.

Prepared for the
Office of Health and Environmental Research
U.S. Department of Energy

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
operated by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
Under Contract No. DE-AC05-84OR21400

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0279136 1

TABLE OF CONTENTS

	Page No.
ABSTRACT	v
1. INTRODUCTION	1
2. MATERIALS AND METHODS	3
2.1. THE COMPUTER SYSTEM	3
2.2. GEL IMAGE PROCESSING AND ANALYSIS PROCEDURES	3
2.3. TWO-DIMENSIONAL GEL ELECTROPHORESIS	5
2.4. EXPERIMENTS	6
3. RESULTS	7
4. DISCUSSION	11
5. REFERENCES	13

ABSTRACT

Three algorithms for quantitative 2d gel electrophoresis were compared when applied to a set of simulated images, and to gel images obtained with separated cytoplasmic proteins of FELC. The methods involved fitting of Gaussian surfaces to protein spots as well as quantitation that did not rely on any assumptions concerning the distribution of radioactivity within a protein spot. Image processing algorithms preceding the analysis were identical for the three methods. Results based on isolated spots show that for the majority of cases there is no statistical difference between the estimates produced by the three methods. This suggest that computationally expensive fitting of two-dimensional surfaces is not necessary in order to achieve satisfactory quantitation.

ABBREVIATIONS USED

CV - Coefficient of variation

FELC - Friend erythroleukemia cells

2d - two-dimensional

1. INTRODUCTION

The importance of computer analysis and pattern recognition methodology for quantitative two-dimensional electrophoresis is being recognized, as documented in several reports on analysis systems that appeared in the literature during the past few years [1,2,3,4,5,6,7,8,9,10]. However, no data are available from systematic studies that would allow a critical assessment of the relative advantages and disadvantages as well as potential intrinsic limitations of the different methods described so far. The purpose of this article is to present results that we obtained from experiments designed to compare the performance of different methods for gel image analysis.

The goal of quantitative gel image analysis is to reduce the pictorial data to a list of protein spots, where each spot is defined by a triplet, quintuplet or a different but usually small amount of numbers, e.g. location (x,y), spread in x and y, integrated density, and sometimes boundary coordinates. The analysis is preceded by a processing step, generally a bandpass operation that eliminates high and very low frequencies, the latter corresponding to a locally varying image background that would make simple thresholding inadequate for image segmentation. Some systems use a local thresholding technique for background elimination [2,4], others a non-linear min/max filter [1,9] first described in the context of "mathematical morphology" [11] as "image opening".

Although the algorithms for spot detection and quantification that have been described differ in detail, partially because of different hardware configurations, we can categorize these methods in two sets: parametric methods that operate on the basis of some model for the spots [1,3,5,6], and non-parametric methods that do not rely on any assumptions as to the shape of the spots [4,9,10]. We will subsequently refer to these sets as P and NP respectively.

Methods in P usually assume Gaussian spot profiles. This can be justified considering the diffusion processes involved in electrophoresis combined with the sieving effect produced by the concentration gradient in the polyacrylamide gel medium. Each spot is characterized numerically by at most six parameters: location, spread in x and y, maximal intensity, and angle of principal component if the Gaussian is not assumed to be separable. A synthetic image produced by the superposition of all Gaussians is fitted optimally, usually in the least squares sense, to the bandpass-filtered image.

As with parametric methods the implementation details for methods in NP differ considerably between the systems described in the literature. The common feature of NP methods is that time-consuming surface fitting is not performed. Overlap between spots is resolved by examining spatial derivatives of the image. Spots are characterized by their location and volume, and often boundary coordinates are stored in order to allow for display of a contour map depicting the result of computer analysis.

NP methods cannot achieve the amount of data reduction obtained with P methods without sacrificing the frequently important capability of displaying a synthetic gel image, based on the numerical information on the spots, that resembles the original image. Obviously, non-Gaussian spot profiles can be handled better by NP methods.

In this paper we describe the results of experiments designed to compare the accuracy and consistency of quantification obtained with 3 methods: a sub-optimal fit of Gaussian profiles with principal axes aligned to the image coordinate axes (method A), an optimum least squares fit of Gaussian profiles without restriction on the angle of principal axes (method B), and a typical non-parametric method (method C).

2. MATERIALS AND METHODS

2.1. THE COMPUTER SYSTEM

Our analysis algorithms were implemented on a configuration that consists of a PDP 11/70 (Digital Equipment Corp., Maynard, MA), operating under RSX-11M-PLUS with 1.5 Mbyte main memory, a 68 Mbyte RM03 disk, a 475 Mbyte RA81 disk, and a TE16 magnetic tape drive. It is linked via Ethernet to a microVAX II (DEC, Maynard, MA) computer with 9 Mbyte main memory, 210 Mbyte disk space, and a TK50 tape cartridge, operating under microVMS 4.3. Images are digitized with an Optronics P-1000 scanning densitometer (Optronics International Inc.,Chelmsford, MA). Some image processing functions, image storage during analysis, and image display are performed by a specialized pipelined processor, an IIS Model 75 (International Imaging Systems,Milpitas, CA), with the PDP 11/70 as host computer. The IIS Model 75 is currently equipped with 12 512x512x8 bit refresh memory channels. Images are displayed on a Mitsubishi color TV monitor (Mitsubishi Electronics America Inc.,Compton, CA), and can be copied onto film by a Color Graphic Recorder Model 3000 (Matrix Instruments Inc., Orangeburg, NY). The software is written in FORTRAN 77.

2.2. GEL IMAGE PROCESSING AND ANALYSIS PROCEDURES

Gel images in our laboratory are approximately 15 cm x 15 cm, and are digitized at 100 u intervals, resulting in digital images of 1536 x 1536 pixels. We refer to the isoelectric focussing dimension as x, and the molecular weight dimension as y. The steps involved in gel image processing were previously described in detail [6]. Briefly, images are smoothed with a Gaussian low pass filter, and locally varying background, including vertical and horizontal streaks, are removed, using a non-linear min/max filter [1,11]. Figure 1 illustrates the effect of these filters on a 100 x 100 pixel area of a typical gel image. This image processing procedure is common to the 3 analysis methods that we tested.

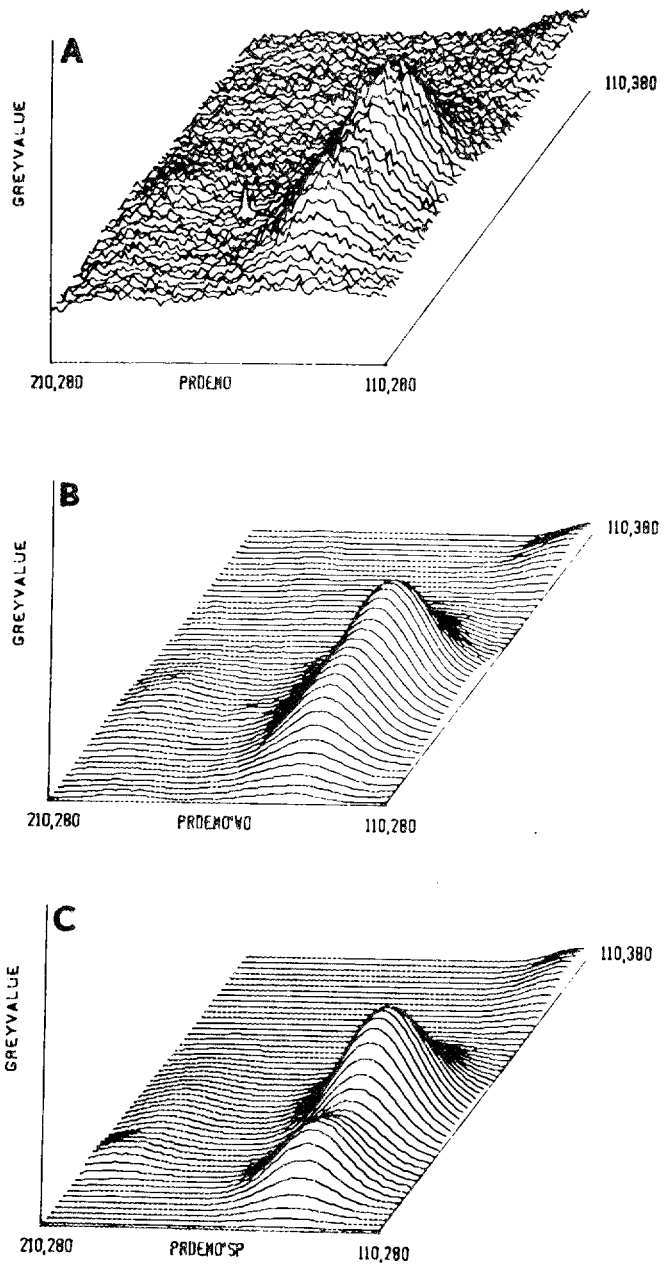

Figure 1. Panel A shows an isometric plot of a 100x100 pixel area (i.e. $1cm^2$) of a typical, unprocessed gel image digitized with a scanning densitometer. Panel Bd epicts the same image section after smoothing and elimination of spatially varying background. The fit generated by image analysis method A (see text), is shown in panel C. The scale on the z-axis labeled "greyvalue" is the same in all 3 panels.

Method A was described in detail in [6]. Spot estimation is based on the assumption that the spots can be adequately described by separable Gaussians. It follows that images can be analyzed in the 2 dimensions separately. A set of rules is used to detect a spot or overlapping spots by examining the derivative of the greyvalue function. The fit of the separable Gaussian model to the image is sub-optimal because not all parameters are allowed to vary simultaneously. In order to increase the speed of analysis, the simplifying assumption is made that the spot finding process results in acceptable estimates of spot location and size. The solution of linear equations then gives estimates of spot intensities [6].

Method B consists of an optimal fit, in the least squares sense, of a two-dimensional Gaussian profile whose major axes are not constrained to be parallel to the image coordinate axes. All parameters are allowed to vary simultaneously. For the purpose of this study, the initial estimates required by method B were the results obtained with method A. The non-linear least squares fit was computed using a CORLIB subroutine (LMDER1) that is based on the Levenberg-Marquardt algorithm [12].

Method C estimates spot volumes by adding pixel values that are above a selected threshold within a rectangular area around an isolated spot. The location of this area, i.e. the location of its center, was set equal to the coordinates of the spot center determined by method A. The size of the area was determined as $4(s_x \times s_y)$, where s_x and s_y are the standard deviations of the Gaussian fitted by method A.

Only spots that were well defined, clearly detectable, and well isolated were included in the comparative analysis of the 3 methods.

2.3. TWO-DIMENSIONAL GEL ELECTROPHORESIS

The gel image data presented here were obtained with cytoplasmic proteins from Friend erythroleukemia cells (FELC). FELC, clone 19-9 [13], were labeled with a ^{14}C amino acid mixture from day 4 to day 5

after plating at 2x105 cells/ml in Dulbeccos modified Eagle medium supplemented with 10% FCS. Two-dimensional polyacrylamide gel electrophoresis was performed as a modification of the O'Farrell method [14]. Twelve 1.4 mm i.d.x 130 mm long tube gels prepared according to Anderson [15] were loaded with 60-150 ug FELC cytoplasmic proteins and isoelectrically focused for 690 Volthours. Twelve second dimension SDS 10-20% linear gradient slab gels 16 x 18cm x 1.4mm were electrophoresed simultaneously at 175 ma h/slab. The second dimension slab gels were fixed overnight. After a 10 min. distilled water rinse, slab gels were prepared for fluorography by agitation in 4 volumes of 97% Amplify (Amersham)/3% glycerol (v/v) for 1 h and dried at 80° C under vacuum for 1h followed by an additional 2h under vacuum on filter paper. Kodak type SB X-ray films were preflashed and exposed to the dried gels according to the method of Laskey and Mills [16]. Optical density to dpm conversion calibration strips impregnated with the same fluor accompanied each group of films for fluorography.

2.4. EXPERIMENTS

We performed 2 experiments to compare the performance of the 3 methods in terms of accuracy and variability of spot volume estimates.

The first experiment (EX1) used a computer-generated image containing 5 isolated Gaussian spots of different sizes and intensities. Ten images were generated from this synthetic image by adding independent samples of Gaussian noise produced by a random number generator. A mean value of 30 and a standard deviation of 5 greyvalues were chosen for the noise process to simulate realistic conditions (Kodak SB X-ray film, preflashed). Maximum spot intensities were 50, 100, 140, 170, and 200 greyvalues. All images were analyzed by methods A, B, and C.

The second experiment (EX2) used gel images resulting from the separation of cytoplasmic proteins of FELC. Electrophoresis was performed as described in the previous section. Three independent gels were produced from one protein sample. The resulting 3 images were

analyzed using methods A, B, and C.

3. RESULTS

The outcome of EX1 is summarized in Table 1. A one-way ANOVA of the spot volume estimates obtained with the 3 methods showed that for 4 of the 5 spots the results were significantly different. Further analysis of the data using Scheffe's test [17] allowed to isolate differences and revealed that methods A and B gave results different than method C. The spot for which no difference was detected (number 2) had the lowest signal to noise ratio which resulted in high coefficients of variation of the estimates from methods A and B. With the very small and bright spot number 5, the surface fitting methods A and B gave less accurate estimates than method C. This is due to the fact that small errors in the size estimates have a major impact on the volume estimate for this kind of spot. Another source of error in general as well as for this spot in particular is the fact that in our system, estimates of size and location are always rounded to the nearest integer. Moreover, some practical considerations regarding the choice of parameters for subroutine LMDER1 [12], which is central to method B, result in less than optimal fitting. Most important among these is the selection of the parameter on the basis of which the decision is made whether the sequence of iterates has converged. We chose to end the iteration when the relative change in successive iterates was less than $r=5x10^{-3}$. This choice was based on recognizing the tradeoff between smaller values of r and increased computation times, often several minutes per spot for $r < 10^{-5}$ using a microVAX II computer as described above. We attribute the higher relative error obtained with method B on spots 1 and 3 as compared to method A to this choice of r.

This phenomenon raises an interesting point that deserves further comment. The least squares estimator is optimal in the presence of Gaussian noise, i.e. we would expect the results from B to be better than those from A or C. However, after the non-linear bandpass filtering that precedes any of the 3 methods studied here, image

TABLE 1. Summary of results form EX1. True spot parameters (x,y,s_x,s_y,max): spot1 = (100,100,10,10,100), spot 2 = (150,100,5,10,50), spot 3 = (200,200,12,5,140)spot 4 = (300,300,3,8,200), spot 5 = (100,300,4,4,170). The true spots were corrupted with Gaussian noise (m=30,s=5). Discussion of results in the text.

spot #	max. intensity [greyvalue]	average relative error [%] A	B	C	coefficient of variation [%] A	B	C
1	100	2.1	5.9	10.0	4.6	0.4	0.6
2	50	9.3	3.9	12.6	6.7	10.5	1.6
3	140	4.4	6.1	7.8	3.0	0.1	1.0
4	200	3.1	0.0	7.4	5.2	0.5	1.7
5	170	21.8	15.1	4.8	8.7	0.6	1.5

	one-way ANOVA $F_{0.05(2)2,27}=4.24$	Scheffe's test $S_{cr0.05}=2.912$ A vs B	A vs C	B vs C
1	21.817	3.159	6.604	3.445
2	4.108	1.749	1.092	2.841
3	157.391	1.983	14.277	16.260
4	27.724	3.555	8.689	5.134
5	49.656	2.376	9.507	7.130

distortions, although small, are not Gaussian any more. Moreover, practical choices of parameters for the discrete implementation of the estimator must be made, which eventually leads to less than optimal results.

For the analysis of EX2 the 3 images produced from the same sample where brought into registration using a polynomial spatial coordinate transformation [6], and 35 well-defined, isolated spots were included in the comparison of the 3 methods. Table 2 depicts the results of EX2 as analyzed by ANOVAs and Scheffe's test performed with dpm estimates. Only the spots for which differences were detected are shown in the table. The one-way ANOVA pointed out differences between the estimates of methods A, B, and C for 5 out of the 35 spots, with Scheffe's test detecting differences between A and C, B and C, but no difference between A and B. A two-way ANOVA was performed to assess the influences of the 3 images and the 3 methods on the results. It can be seen from Table 2 that in all 5 cases the methods were the dominant factor contributing to a statistical difference, except for spot 61 for which the different images had a significant effect also.

Independent of the choice of analysis method, the results of quantitation are sensitive to the choice of parameters for the procedures involved in gel image processing prior to analysis. We have observed this with the choice of window size for the non-linear filter that we use to eliminate streaks from the gel images. The window should be chosen larger than the maximum extension of the biggest spot in the image. A range of window sizes fits this rule of thumb, due to diffculties in defining spot sizes. Therefore, we used method A to analyze 7 gel images after background removal with 4 different window sizes: 55, 75, 95, 105 pixels. With 20 well-defined, isolated spots, we observed that the CVs of spot estimates ranged from 0% to 50%. This filter should be kept constant for a series of images that are to be compared to each other.

TABLE 2. Results for 5 out of 35 spots for which differences were detected in EX2. Discussion of results in the text.

spot #	one-way ANOVA $F_{0.05(2)2,6}=7.26$	Scheffe's test $S_{cr0.05}=3.81$ A vs B	A vs C	B vs C
12	17.61	0.00	5.14	5.14
55	28.57	1.11	5.92	7.03
61	15.40	0.04	4.83	4.78
98	12.05	0.39	4.43	4.05
177	12.37	1.96	2.98	4.94

	two-way ANOVA $F_{0.05(2)2,4}=10.6$	
	image	method
12	1.70	21.72
55	0.80	26.62
61	14.59	85.15
98	0.53	10.14
177	2.02	15.57

4. DISCUSSION

Automated analysis of gel images is an important part of the quantitative study of proteins that have been separated by two-dimensional gel electrophoresis. In order to establish the sensitivity of assays relying on this separation technique it is crucial to know how accurately and reliably information can be extracted from gel images. Despite the multitude of methods for gel image analysis that have been described in the literature, no comparative data have been available so far that would allow to assess the relative advantages and disadvantages of the different approaches. Among the problems that make it difficult to compare several analysis systems currently in operation when applied to same set of gel images, are different gel sizes and the fact that the analysis systems are carefully tuned to the kind of gel images that are produced in the respective electrophoresis laboratories, and the logistics of such a comparative study. Therefore, we decided to implement on our general purpose image analysis system in addition to the method already in place (method A) two other methods (B and C) to be used in an experimental mode for the purpose of method comparison.

On the basis of our data for isolated spots, and the gel images produced in our laboratory, we can make the following conclusions: (1) The use of a computationally expensive fully two-dimensional fit of Gaussian profiles to model spots is not justified. Although CVs for the estimates of synthetic spots were generally lower with method B than with method A, there was no statistically detectable difference between the results from A and B with real gel images. (2) Although method C, which does not involve computationally expensive surface fitting, generated results that are statistically different than the results from A or B, the errors and CVs achieved with this method do not increase catastrophically in comparison with A and B. For a very small and bright spot this method even gave smaller errors than the surface fitting procedures. (3) A significant difference between the estimates from B and C was detectable in only a small fraction of spots (5/35) in

the real gel images that were analyzed. This suggests that in order to achieve satisfactory quantitation, fitting of two-dimensional Gaussian profiles to spots is not necessary.

Many of the methods, in NP as well as P, described in the literature are quite capable of separating overlapping spots. Therefore, we do not anticipate a significant change in these results in the case of overlapping spots, although we have not rigorously tested methods B and C on our system with non-isolated spots.

Although the development of algorithms for gel image analysis (and image analysis in general) should be addressed largely independent on the available hardware, a concluding comment on using general purpose parallel computers for gel analysis is in order because the recent increase in the availability of advanced general purpose concurrent computers at very attractive cost/performance ratios will have an impact on further development of new methods and will influence the implementation of existing gel analysis methods. As pointed out in [9], many of the operations involved in gel image analysis, including image comparison, can be performed in spatially separated image regions at the same time. Recent developments in rapid electronic autofluorography [18] make it possible to generate images from electrophoresis gels in a matter of minutes versus hours or days when exposing gels to film. Therefore, the speed of analysis becomes an important factor in selecting algorithms and hardware. We have started using an NCUBE general-purpose 64-processor hypercube computer (NCUBE Corp., Beaverton, Oregon) for the non-linear image processing involved in our method. Without major code modifications or optimizations we observed increases in processing speeds up to 2 orders of magnitude with the NCUBE machine compared to the pipelined image processor/minicomputer host hardware, and a speedup factor of 1.6 compared to a Cray X-MP 12 computer. Multi-processor boards are now available at very moderate cost for personal computers. Thus, the prospects are promising for affordable computer-assisted rapid quantitative two-dimensional gel electrophoresis for an increasing number of laboratories.

5. REFERENCES

[1] Anderson, N. L., J. Taylor, A. E. Scandora, B. P. Coulter, N. G. Anderson, Clin. Chem., 27, 1807-1820 (1981).

[2] Bossinger, J., M. J. Miller, K. P. Vo, E. P. Geiduschek, N. H. Xuong, J. Biol. Chem., 254, 7986-7998 (1979).

[3] Garrels, J. I., J. Biol. Chem., 254, 7961-7979 (1979).

[4] Lemkin, P. F., L. E. Lipkin, Comp. Biomed. Res., 14, 272-297, 335-380, 407-446 (1981).

[5] Lutin, W. A., C. F. Kyle, J. A. Freeman in: N. Catsimpoolas (Ed.), Electrophoresis '78 Elsevier/North Holland, Amsterdam-New York, pp. 93-106 (1979).

[6] Mann, R. C., B. K. Mansfield, J.K. Selkirk in: E. S. Gelsema, L. N. Kanal (Eds.), Pattern Recognition in Practice II, Elsevier/North Holland, Amsterdam, pp. 301-311 (1986).

[7] Pardowitz, I., H. G. Zimmer, V. Neuhoff, Clin. Chem., 30, 1985-1988 (1984).

[8] Ridder, G., E. VonBargen, D. Burgard, H. Pickrum, E. Williams, Clin. Chem., 30, 1919-1924 (1984).

[9] Skolnick, M. M., S. R. Sternberg, J. V. Neel, Clin. Chem., 28, 969-978 (1982).

[10] Vo, K. P., M. J. Miller, E. P. Geiduschek, C. Nielsen, A. Olson, N. H. Xuong, Anal. Biochem., 112, 258-271 (1981).

[11] Serra, J., Image Analysis and Mathematical Morphology, Academic Press, London-New York, pp. 425-478 (1982).

[12] More, J. J., Proceedings of the 1977 Dundee Conference on Numerical Analysis.

[13] Friend, E., W. Scher, J. G. Holland, T. Sato, Proc. Natl. Acad. Sci. USA 68, 378-382 (1971).

[14] O'Farrell, P. H., J. Biol. Chem., 250, 4007-4021 (1975).

[15] Anderson, N. G. and N. L. Anderson, Anal. Biochem., 85, 331-340 (1978).

[16] Laskey, R. A. and A. D. Mills, Eur. J. Biochem., 56, 335-341 (1975).

[17] Zar, J. H., Biostatistical Analysis, Prentice Hall, Englewood Cliffs, NJ (1974).

[18] Davidson, J. B. in: V. Neuhoff (Ed.), Electrophoresis '84, Verlag Chemie, pp. 235-251 (1984).

INTERNAL DISTRIBUTION

1. V. B. Baylor
2. J. J. Dorning (Consultant)
3. G. H. Golub (Consultant)
4. R. M. Haralick (Consultant)
5. J. P. Jones
6. F. C. Maienschein
7-36. R. C. Mann
37-46. B. F. Mansfield
47. F. G. Pin
48. H. Postma
49. C. R. Richmond
50. D. Steiner (Consultant)
51. C. R. Weisbin
52. A. Zucker
53-54. EPMD Reports Office
55. Central Research Library
56. Document Ref. Section
57. Y-12 Technical Library
58-59. Laboratory Records
60. Laboratory Records - RC
61. ORNL Patent Office

EXTERNAL DISTRIBUTION

62. Office of Assistant Manager for Energy Research and Development, DOE, Oak Ridge Operations Office, P. O. Box E, Oak Ridge, TN 37831.

63- 72. D. J. Riese, 367 Cedar Street, Box 605, New Haven, CT 06510.

73- 82. J. K. Selkirk, Branch Chief, National Toxicology Program, NIEHS, Research Triangle Park, NC 27709.

83-113. Technical Information Center, P. O. Box 62, Oak Ridge, TN 37831.

NOTICE

When you no longer need this report, please return it to R. C. Mann, Engineering Physics and Mathematics Division, Bldg. 6025, MS-364, Oak Ridge National Laboratory, P. O. Box X, Oak Ridge, TN. 37831-6364.

THANK YOU