Reprint from: Electrophoresis '79
Advanced Methods
Biochemical and Clinical Applications
Editor: Bertold J. Radola

ESTIMATION OF TWO-DIMENSIONAL ELECTROPHORETIC SPOT INTENSITIES AND POSITIONS BY MODELING

J. Taylor†, N. L. Anderson, B. P. Coulter†, A. E. Scandora, Jr.††, and N. G. Anderson

Molecular Anatomy Program, Division Of Biological and Medical Research, Argonne National Laboratory, Argonne, Illinois 60439, USA.

Introduction

The prototype ISO-DALT analytical systems (1-4) for two-dimensional electrophoretic analysis of proteins allow a large fraction of the estimated 30,000 to 50,000 human protein gene products (PGP's) to be separated, quantified, and identified by their positions on two-dimensional maps, and by techniques described by other papers in this symposium (see papers in this volume by N. L. Anderson, C. S. Giometti, S. L. Nance, S. L. Tollaksen, J. J. Edwards, B. J. Hickman, and K. E. Willard and their co-workers).

The necessity for computerized data reduction as described here arises from the fact that both the number of spots (PGP's) and the number of gels are too great to allow visual assimilation and intercomparison of the data. Comparison of two patterns by optical superimposition is relatively easy (5). However, visual intercomparison of many gels is extremely difficult when large numbers of spots are involved, and virtually impossible if quantitative results are required. The magnitude of the problem can be appreciated when one realizes that up to 2,000 spots may be seen on a single gel, and over 20,000 two-dimensional gels have been run thus far in this laboratory.

The reduction of such a large amount of information requires a system

*This work is supported by the U. S. Department of Energy under contract No. W-31-109-ENG-38.
†Applied Mathematics Division
††Science Applications, Inc.

of image scanners and computers that can process a large number of patterns and yield lists of abundances (integrated optical densities) and positions for each PGP on a gel. Such data can then be examined for differences between samples and for correlations of abundance of any spot with internal variables (stage of differentiation, for example) or external variables (radiation, viral infection, carcinogens, etc.). This approach is known as "list-based biology".

A prototype data reduction system for the processing of two-dimensional electrophoretic gels is being developed at this laboratory. The system includes a Digital Equipment Corporation PDP-11/60 minicomputer and a Floating Point Systems AP-120B array processor. The images are scanned with an Optronics P-1000 high speed drum densitometer, and are displayed using a Grinnell GMR-27 512 X 512 color image display system.

The software currently consists of several major components. The first part is the scanning and preliminary processing system. Images are scanned and stored on a large capacity disk storage module. Typical images have 100 micron resolution and consist of 1.5 to 2.5 X 10^6 8-bit numbers. The user visualizes the image on the display system, selects the region to be processed, and if appropriate, corrects the image for the photographic response curve. All images are scanned and processed in the optical density domain. The second major component of the gel analysis system performs image filtering and background subtraction operations on the data. Filtering is necessary to remove image noise such as that caused by photographic grain. Both linear filters and median filters can be used. The background function is then estimated using interpolation techniques and is subtracted from the entire image. The third part of the software system is the spot detector. This component automatically detects the individual spots and estimates their intensities and positions. The results are reviewed and can be interactively corrected using the display system and its trackball device. The fourth component of the analysis system is the spot modeling software. If all the PGP's were well separated then the preliminary measurements would be accurate enough to provide the required information. Unfortunately, spots are clustered in many areas of the gel. Segmentation by the generation of spot boundaries has been used by other workers to resolve this problem (6). Segmentation is inherently inaccurate in the overlapped areas and may seriously

underestimate the PGP abundances. Also, only lower limits can be assigned to the abundances for PGP's with overrange picture element (pixel) values (caused by exceeding the range of the photometric system). The spot modeling system fits two dimensional models for each PGP. In this way the contribution from each spot can be estimated, even when some of the PGP's are moderately overexposed, or overlapping other PGP's. The last part of the gel analysis software is the matching system, which associates identifiable PGP's with the corresponding spots on a master pattern. The pattern is stretched by the computer until it matches the standard as closely as possible, thus allowing tentative identifications to be made for many PGP's at one time.

It is the purpose of this paper to describe the spot modeling system. Its basic concepts and underlying assumptions will be discussed. Details of the algorithms are beyond the scope of this paper and will be presented elsewhere.

The Model

It is assumed for this analysis that each PGP can be estimated by a two-dimensional distribution, the parameters of which may vary from spot to spot. Gaussian distributions were chosen for this work, but this choice is not essential to the method. Other distributions could have been used as long as their first and second derivatives exist over the image domain. Gaussian parameterization has been used in the analysis of gel fluorograms (7); in that work, the data means and widths are estimated in a single dimension, and the amplitudes are calculated by a linear regression technique. The present technique differs from that of Garrels et al., 1979, in that it is a true two-dimensional fitting system, in which all parameters are allowed to vary.

The modeled image is assumed to be composed of the sum of the individual distributions:

$$I(x,y) = \sum_{k=1}^{n} F_k(x,y).$$

Here $I(x,y)$ represents the actual image as a function of the pixel coordinates (x,y), F_k represents the two-dimensional distribution for PGP k, and n represents the number of PGP's found in the image by the spot detector. Note that this assumption of superposition would be inappropriate if the image were not being processed in the optical density domain. The method of fitting sums of Gaussian distributions to data is often encountered in the analysis of one dimensional spectra. Three dimensional examples can be found in the fitting of protein molecular models to electron density maps (8). Another assumption is that the individual distribution functions $F_k(x,y)$ can be factored into parts a_k, g_k and h_k, where g_k is a function of x only and h_k is a function of y only:

$$F_k(x,y) = a_k g_k(x) h_k(y).$$

This restriction is necessary to simplify the calculations. The parameter a_k is independent of x and y and represents the amplitude for spot k. For Gaussian distributions the functions $g_k(x)$ and $h_k(y)$ are defined as follows:

$$g_k(x) = e^{-r_k^2(x-x_k)^2}, \quad h_k(y) = e^{-q_k^2(y-y_k)^2}.$$

Thus PGP k is parameterized by the five numbers a_k, r_k, x_k, q_k, and y_k. Note that in the case of Gaussian distributions, the PGP's must have their major and minor axes parallel to the x and y coordinate axes. Initial studies have indicated that these assumptions are not overly restrictive. Autoradiograph patterns exhibit the best fits so far to the model, probably because small samples can produce reasonable patterns. Overloaded gels of all types tend to violate the assumption of superposition, have PGP's that are locally not aligned along the x and y axes of the scan, and are asymmetric.

Cluster Processing

Typical gel patterns may contain over 1000 PGP's that need to be processed. It is obvious that optimizing the parameters for all spots simultaneously is an overwhelming and unnecessary task. The optimization of parameters of spots in one region of the gel should have little effect on the parameters of other spots that are not close. Equally obvious is the fact that optimizing the parameters for each PGP singly will result in serious inaccuracies when clusters of overlapping spots are present, especially when the spot intensities are radically different. Thus it is necessary to have a procedure to decide which spots can be fitted alone, and which spots have to be optimized simultaneously with other PGP's. It should also be realized that memory space constraints in the hardware impose practical limits on the number of spots that can be optimized at one time. With the present system, 20 PGP's (100 parameters) per cluster is a realistic limit. The procedure for selecting which spots to group together must therefore limit the number of spots per complex, and it must do so in such a way as to minimize the effect of the spots that are left out.

The present work uses a grouping program based on a modified net building algorithm. The input to the program is a list of the PGP parameters, sorted by position. Only the current state of the parameterization is considered, not the actual image data. Hence the question of whether or not two spots are overlapping can be answered quickly, without resorting to region analysis on the image itself. One simply evaluates the overlap integrals

$$O_{jk} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_j(x,y) F_k(x,y) dxdy \ .$$

Spots j and k are considered to be connected if their overlap O_{jk} is greater than some threshold τ. The net building algorithm proceeds in the following manner:

> Any spot j, which is not already a member of another net, is selected to be a seed, and is placed at the head of a list which will eventually enumerate all PGP's which form this

complex. All spots k with $O_{jk} > \tau$ are then added to this list. The list is then examined, one element at a time, to find additional spots with an overlap greater than τ with at least one member of the list. These additional spots are appended to the list, and the procedure continues, until no more PGP's can be added.

The sorting of the parameter file by spot position reduces the search time (and main memory requirements) considerably. The algorithm is modified to induce a premature termination of the complex growing if too many spots are being added. Only the first p (usually 15) spots in the membership list are considered when searching for additional members. The spots in the list which are added after the first p elements are processed in the usual way during the fitting procedures, but their parameters will not be updated. They remain eligible for membership in other complexes. Spots from the first complex that are overlapped with these special spots can be included in other complexes in a similar manner. In this way very large groups of spots can be considered as smaller complexes, with boundary zones to minimize the effect of not fitting all the spots in the group simultaneously. The procedure is written in both PL/I and FORTRAN, and can process a file of 1000 spots in about 15 seconds on the PDP-11/60 computer.

The total image is processed from top to bottom, one line at a time. As each complex is encountered, a region of main memory is dynamically allocated to contain its data structures, and the complex is marked as active. Processing of this complex continues as new image lines are read in. Details of the actual processing will be discussed below. There may be several complexes active at any one time. After the entire region of a complex has been processed, the parameters are updated, convergence is checked, and the memory is deallocated. The cycling of the grouping program and the processing program continues until all complexes are converged. Complexes are considered converged if all of their member spots are converged. Conversely, if a complex is not converged, then all of its spots are marked not converged, regardless of the changes of their parameters on the previous cycle.

The Fitting Kernel

The parameters are optimized in a least squares sense using a linear differential-correction technique (9). The method assumes that the image I(x,y) can be most closely approximated by the function

$F(x,y,c_1, \ldots ,c_m)=\sum_{k=1}^{n} F_k(x,y)$. The c 's represent the spot parameters for all the spots in the complex, and are to be determined by the optimization process. The estimates for the c 's at any given stage in the fitting process are denoted by c_1', c_2', ... , c_m'. It is assumed that the estimated parameters are close to the actual, or converged, parameters and that

$$F(x,y,c_1,\ldots,c_m) \tilde{} F(x,y,c_1',\ldots,c_m') + \sum_{i=1}^{m} \frac{\partial F}{\partial c_i} (c_i-c_i') \ .$$

The partial derivative is evaluated at $c_i=c_i'$. The residual

$r(x,y)=F(x,y,c_1, \ldots ,c_m)-I(x,y)$ can then be approximated in terms of the current parameterization:

$$r(x,y) \tilde{} [F(x,y,c_1',\ldots,c_m')-I(x,y)] + \sum_{i=1}^{m} \frac{\partial F}{\partial c_i} \delta c_i \ ,$$

where δc_i is defined as c_i-c_i'. These are the values that need to be approximated to determine the new parameters. The calculated residual R(x,y) (based on the current parameterization) is defined to be

$$R(x,y) = F(x,y,c_1',\ldots,c_m') - I(x,y) \ .$$

The square of the residual is integrated over the region of the complex (denoted by Ω) to form the function Q, which is to be minimized:

$$Q = \iint_{\Omega} r^2(x,y)dxdy \tilde{} \iint_{\Omega} \left[R(x,y) + \sum_{i=1}^{m} \frac{\partial F}{\partial c_i} \delta c_i\right]^2 dxdy \ .$$

Q is considered to be a function of the δc_i's. The equations

$\frac{\partial Q}{\partial \delta c_k} = 0$ hold at the minimum of Q for k from 1 to m.

Evaluating the k-th partial derivative, we get

$$0 = \frac{\partial Q}{\partial \delta c_k} \approx 2 \iint_{\Omega} \left[R(x,y) + \sum_{i=1}^{m} \frac{\partial F}{\partial c_i} \delta c_i \right] \frac{\partial F}{\partial c_k} \, dxdy \ .$$

The above formula represents a set of simultaneous equations which can be put into the matrix form

$B \Delta = D$, where B is an m by m matrix with

$$B_{ij} = \iint_{\Omega} \frac{\partial F}{\partial c_i} \frac{\partial F}{\partial c_j} \, dxdy \ .$$

The column vector Δ has elements δc_i, and the vector D has elements

$$D_i = \iint_{\Omega} [I(x,y) - F(x,y)] \frac{\partial F}{\partial c_i} \, dxdy \ .$$

The matrix equation is easily solved for Δ, which is used to update the parameters. Note that in the matrix equation above, only the vector D is a function of the image data. This fact eases the memory allocation requirements since the m X m matrix does not have to be calculated until after the line processing of image data is finished.

The region Ω over which the integrations are performed is set to be that area for which $F(x,y,c_1', \ldots, c_m') > \varepsilon$. The value of ε is small and is chosen to be consistent with the overlap threshold τ used in the grouping. In practice, the region is usually taken to be $-\infty$ to ∞ for both x and y for the B matrix, as these integrals are close to those over the finite region and are much easier to compute. However, there often arise two conditions which require that certain subregions be omitted from the integrations. The first condition is the problem of a

spot occurring on the edge of the scanned image, and the second condition is the existence of an area where the pixel values are overrange. Both conditions result in incomplete data for a spot from the integration. In the case of the B matrix, the integrations are performed over the omitted regions and subtracted from the integrals from $-\infty$ to ∞.

Test Results

The fitting system has been tested on several images with up to 800 spots being processed per pass. Spots are generally converged in 7 or 8 cycles. The left frame of Figure 1 shows a detailed view of one of the more difficult spot complexes to fit. There are three strongly overlapping spots in the foreground, and the center one is overrange. The parameterization of this region is shown in the right frame. Note that the overrange region of the image is predicted by the model.

Future Enhancements

The initial tests have been very encouraging. Future work will be concentrated in three areas. The first area is the increased utilization of the array processor, as the current implementation uses the array processor only for the equation solving. The second major area is the calculation of error estimates for each of the parameters. The third area is the inclusion of different distributions to handle highly loaded gels and patterns with prominent streaks. The inclusion of these features should allow the routine and rapid processing of gel patterns on a production basis.

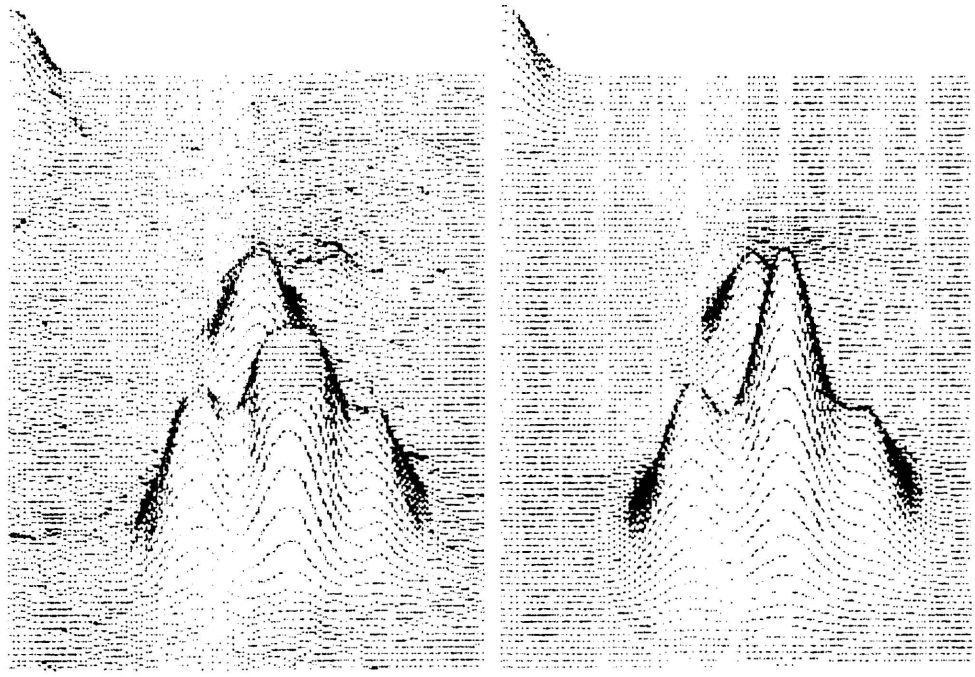

Figure 1. A complex of overlapping and overrange PGP's. The image is shown on the left and its model on the right.

Acknowledgment

The authors would like to express appreciation to J. I. Garrels of the Cold Spring Harbor Laboratory for the use of his profile plotting program used to produce Figure 1.

References

1. O'Farrell, P. H.: J. Biol. Chem. 250, 4007-4021 (1975).

2. Anderson, N. G. Anderson, N. L.: Anal. Biochem. 85, 331-340 (1978).

3. Anderson, N. L. Anderson, N. G.: Anal. Biochem. 85, 341-354 (1978).

4. Anderson, N. G., Anderson, N. L.: Behring Inst. Mitt. 63, 169-210 (1979).

5. Anderson, N. G., Anderson, N. L., Tollaksen, S. L.: Clinical Chemistry 25, 1199-1210 (1979).

6. Bossinger, J., Miller, M. J., Vo, K.-P., Geiduschek, E. P., Xuong, N.-H.: J. Biol. Chem. 254, 7986-7988 (1979).

7. Garrels, J. I.: J. Biol. Chem. 254, 7961-7977 (1979).

8. Diamond, R.: Acta Cryst. A27,436-452 (1971).

9. McCalla, T. R.: Introduction to Numerical Methods and FORTRAN Programming, John Wiley & Sons. New York, 1967.