CLIN. CHEM. 27/11, 1807-1820 (1981)

The TYCHO System for Computer Analysis of Two-Dimensional Gel Electrophoresis Patterns

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

We describe here a computer system for the analysis of high-resolution two-dimensional gel-electrophoresis patterns, with some initial applications. The system (called TYCHO) comprises programs for image acquisition, background subtraction and smoothing, spot detection, gaussian spot modeling, and pattern matching and comparison. It is based on a conventional minicomputer, but makes extensive use of a high-speed array processor in the image-processing and -modeling steps. Used in concert with the ISO-DALT two-dimensional electrophoresis system (*Anal. Biochem.* **85:** 331–354, 1978), TYCHO allows quantitative measurement of hundreds of proteins in complex biological samples, and constitutes the initial data-reduction system required for work towards a Human Protein Index.

Additional Keyphrases: *ISO-DALT system · computerized data acquisition and handling · electrophoresis, polyacrylamide gel*

The complexity of human cells is reflected in the number of different proteins involved in normal cell function. Any given nucleated cell type is generally estimated to contain 3000 to 8000 different proteins, while the total of all the different proteins required by all human cell types during development and at maturity is estimated to be between 30 000 and 50 000. Of these substantial numbers, only a few hundred human proteins (1 to 3% of the total) have been characterized in any reasonable detail, a figure that must evoke modesty in any contemporary biologist or physician. Nearly all of the proteins that have been extensively studied have been found to have important functions, and it seems reasonable to assume that those as yet undescribed have equally important functions that remain to be discovered. With so much unknown, it appears that any comprehensive understanding of how cells work—or fail to work—is currently beyond reasonable expectation.

Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory, Argonne, IL 60439.

Received July 21, 1981; accepted Aug. 25, 1981.

The principal factor limiting knowledge of the variety of proteins found in human cells is the resolution of previously available analytical methods. Classical methods of protein fractionation can generally resolve perhaps 10 proteins, and such methods must therefore be applied in series to effect the isolation of even the more plentiful proteins from a cell homogenate or extract.

The introduction of electrophoretic separations in acrylamide gels (*1*, *2*) substantially improved the possible resolution. These techniques could distinguish 20 to 100 proteins in a mixture, but were still far short of the resolution appropriate to the analysis of cells. Combining in series two quite different such methods, each having high individual resolving power (>100 proteins), O'Farrell (*3*) and others (*4–6*) developed two-dimensional gel electrophoretic techniques that could resolve about 10 000 proteins. In general, these methods involve isoelectric focusing (IEF; *7*) in urea and non-ionic detergents in the first dimension, and sodium dodecyl sulfate (SDS) electrophoresis (*8*, *9*) in the second, yielding a two-dimensional map in which proteins are separated on the basis of chemical properties (pI) in one direction and on the basis of size (SDS-M_r) in the other. Both individual techniques dissociate protein subunits and to some extent denature the polypeptide chains; this effectively minimizes the size of the molecules to be separated and maximizes resolution. The emergence of this technology, and its development into a reproducible, calibrated, routine analytical tool (*10–13*), now allow us to separate and make visible thousands of cellular and body fluid proteins (*14–20*) and thereby become familiar with a greatly increased proportion of the working parts of human cells. Systematization of this knowledge in a "Human Protein Index" (*13*, *21*) now seems to be an attainable goal, given some way of dealing with the large amount of data involved.

In this paper we describe a computer system that is designed to handle data from the two-dimensional gel technique, and report some general results obtained with it. The system in its present form is called TYCHO after Tycho Brahe, the Danish astronomer who painstakingly collected data on the positions of stars and planets in the sky. As did its namesake, TYCHO produces quantitative data from images. In this case, the resulting data are in the form of lists of protein spots, their positions, and their abundances in particular two-dimensional

separations. TYCHO includes programs for generating and editing such lists and for matching and scaling lists obtained from different analyses. A system (called KEPLER) for constructing, maintaining, and studying a data base derived from many analyses and many types of experiments will be described in a later paper. Together, TYCHO and KEPLER constitute the basic information-handling system required for the Human Protein Index.

There are various approaches to the analysis of two-dimensional gel images; the choice among them has generally depended on the computer and display hardware available, the quality of the patterns to be analyzed, and the mathematical or programming strategy preferred. Garrels (*22*) has described a system based on use of a desk-top computer with which it is possible to resolve spot profiles into gaussian peaks in one dimension. A somewhat similar approach, involving a more elaborate gaussian decomposition and a large computer, has been described by Lutin et al. (*23*). Both of these systems are fairly automatic, but each falls short of the complete two-dimensional least-squares gaussian fitting approach desired for optimal quantitation of overlapping spots.

Segmentation analysis, which makes no supposition about ideal spot shape, has been made use of in several sophisticated systems (*24–28*, *13*). With this approach there is difficulty in detecting shoulders and in properly assigning density in regions of overlapping spots. Nevertheless, the technique can be fast and has some advantages in treating spots of irregular shape. Several groups have written computer programs aimed at positional analysis of two-dimensional patterns without provision for quantitation of integrated densities (*29*, *30*). Even though little hardware is required in this approach, the lack of quantitation seems a major drawback and severely limits the usefulness of such systems.

In 1977, we began an effort to analyze our gels by using a segmentation-type approach on the PDP 10-based ALICE image-analysis system at Argonne National Laboratory (*31*, *13*). Although quantitative data of some usefulness was obtained (*13*), it became clear that the special features of our two-dimensional gel image data demanded a different approach. Therefore we designed and assembled a dedicated computer system specifically for this effort in early 1978 (TYCHO I). High-resolution color television graphics capabilities were included, because it had become apparent that development and operation of the desired software would require extensive inspection of and interaction with high-quality image data. An array processor was later added, to handle the computational load involved in the image preparation and gaussian fitting algorithms we chose. TYCHO I was located within and operated as part of the Molecular Anatomy Program's ISO-DALT laboratory (*10*, *11*), so that interaction between the gel-running and gel-analysis efforts would be maximal. As expected, some problems in the computer analysis can be most easily solved by altering the gels, and vice versa. In particular it became apparent that gels of very high quality are required to obtain the information we desired; the ISO-DALT system, because of the improvements attendant on large-scale (10 000 gels per year) operation, can be made to achieve routinely the required resolution and reproducibility. As with the development of the ISO-DALT and the original centrifugal analyzer systems, we have been mindful in designing TYCHO to follow a path that would lead in five to eight years to a relatively inexpensive system for research and clinical use.

Materials and Methods

Isolation and Labeling of Lymphocytes

Lymphocytes were prepared by the Ficoll-Paque (Pharmacia Fine Chemicals, Piscataway, NJ 08854) centrifugation procedure, within 4 h of blood collection. The cells were washed twice and finally resuspended in RPMI 1640 medium minus methionine (Selectamine Kit; GIBCO Laboratories, Grand Island, NY 14072) to give a concentration of approximately 2×10^6 cells per milliliter. The complete medium contains, per liter (and in addition to the RPMI 1640 ingredients), 50 mL of fetal bovine serum, 40 μmol of mercaptoethanol, 10^6 USP units of penicillin, 100 mg of streptomycin, and 200 mg of gentamicin. [^{35}S]Methionine (Amersham, Arlington Heights, IL 60005) was added (25 μCi per 400-μL culture well) and the cells were incubated (37 °C, 5% CO_2 + 95% air, 18 h) with continuous gentle rocking. Cells were then harvested by a 1-s centrifugation (Microfuge B; Beckman Instruments, Palo Alto, CA 94304) in capillary-bottom Microfuge tubes (Walter Sarstedt, Inc., Princeton, NJ 08540), the labeling medium was quickly aspirated, and the cells were promptly lysed in 50 μL of a mixture containing, per liter, 9 mol of urea, 20 mL of the non-ionic detergent Nonidet P-40 (NP-40; Particle Data Laboratories, Elmhurst, IL 60126), 20 mL of mercaptoethanol, 20 mL of pH 3.5–10 Ampholines (LKB Instruments, Rockville, MD 20852), and 0.1 mmol of phenylmethylsulfonyl fluoride (as a protease inhibitor). Condensed DNA (nuclear remnants) was then removed by 30- to 90-s centrifugation in the Microfuge. Samples so prepared lack the troublesome viscosity encountered when nuclear DNA is allowed to swell in the sample solution. They can be used immediately or stored frozen at −80 °C for at least one

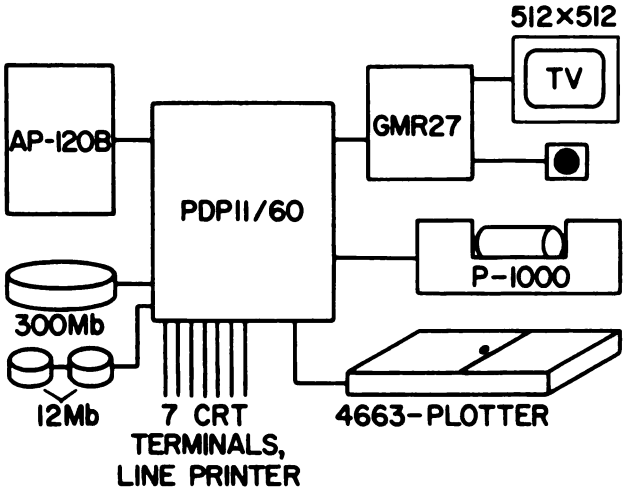

Fig. 1. Schematic diagram of the TYCHO I computer system hardware
Major components are as described in the text

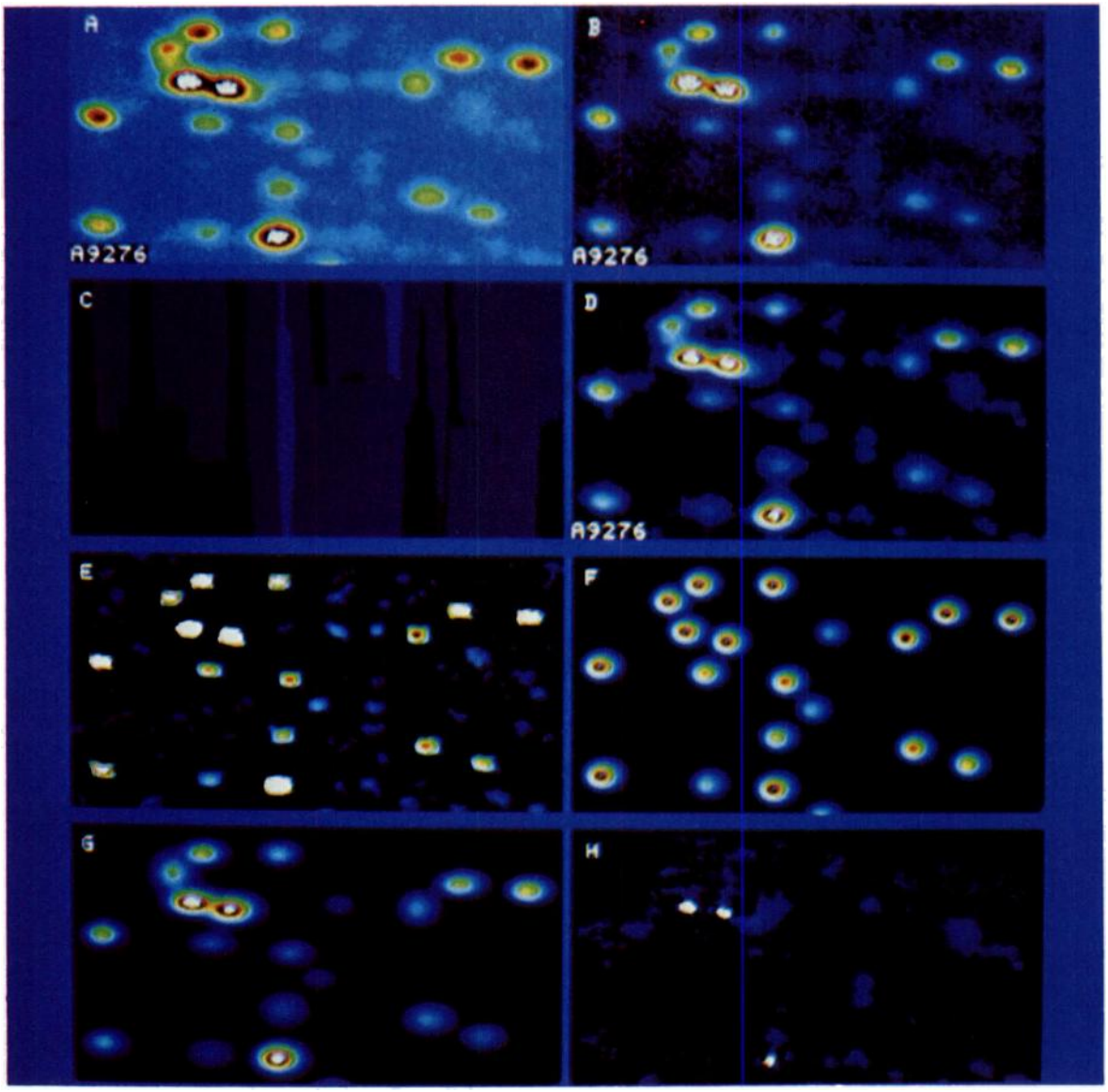

Fig. 2. Photograph of a multi-panel color-television display, showing steps in the image-analysis procedure

Image grey scale has been converted to a color scale ("pseudo-color") for enhanced visual discrimination. Each panel shows the same small section (220 × 120 pixels) of the image after various transformations; the whole image is approximately 90 times the size of this section. (*A*) original image data from the densitometer, (*B*) image after conversion to a scale linear in cpm and with film base density subtracted, (*C*) background and horizontal streaks determined from the converted image, (*D*) image from which background and streaks have been subtracted, (*E*) result of a spot-detecting convolution applied to the converted image, (*F*) synthetic image made from starting spotlist constructed by placing uniform size spots at the spot locations detected by convolution, (*G*) the final synthetic image made from the spotlist after fitting to the background-subtracted image, (*H*) the residual density remaining when the fitted spots of *G* are subtracted from the processed image of *D*. Very little density remains unaccounted for. The three bright regions are over-range (>255) density, which, since it is not accurately measured by the densitometer, is preserved as an unknown value (defined as equal to 255) throughout the processing. Because a threshold is used in the convolution/spot detection step to yield a reasonable number of spots in the starting spot list, some of the very faintest spots in the image are ignored in this example

year without appreciable degradation, as assessed by two-dimensional gel electrophoresis.

Growth and Labeling of Fibroblasts and Lymphoblastoid Cells

The lymphoblastoid cell line GM607 (derived from a normal individual, obtained from the Human Genetic Mutant Cell Repository, Camden, NJ 08103) was grown in complete RPMI 1640, and labeled and harvested as for lymphocytes. The fibroblast line 1493 (also originating from a normal individual, obtained from Meloy Laboratories, Inc., Springfield, VA 22151) was plated in complete Dulbecco's Modified Eagle's Medium but labeled by replacing the medium with RPMI 1640 minus methionine as used for lymphocytes. Fibroblasts were harvested by gently washing the plastic well-bottom with 60 μL of NP-40/urea solution (as for lymphocytes).

Two-Dimensional Electrophoresis

We used the original method of O'Farrell (*3*) as modified in this laboratory (*10*, *11*, *32*). Isoelectric focusing in 9 mol/L urea, 20 mL/L NP-40 (the first dimension) was carried out for 14 000 V·h, with use of 18 mL of 3.5–10 and 2 mL of 2.5–4 LKB Ampholines per liter, in batches of 20 gels (the ISO apparatus). Second-dimension SDS gels were 10–20% acrylamide gradient slabs, run overnight (100 V) with 10 gels per DALT tank. The gels were fixed, stained, dried, and autoradiographed on Kodak XR-5 or XAR-2 film as previously described (*32*). All gels carried unlabeled creatine kinase (EC 2.7.3.2) charge

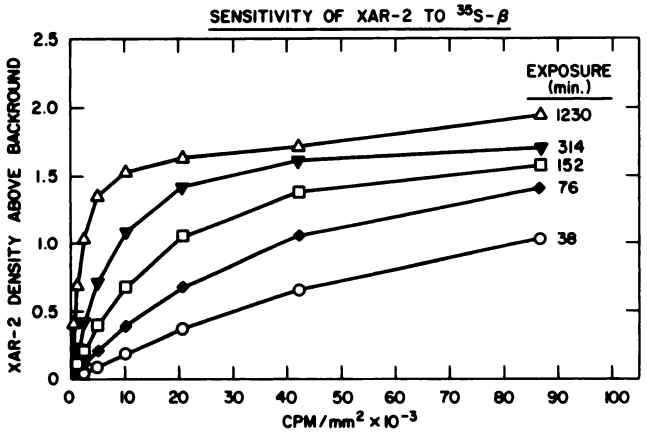

Fig. 3. Absorbance produced on Kodak XAR-2 film by exposure for various times to a stepwedge containing the β-emitting isotope ^{35}S

The maximum useful density is approximately 1.5 *A*; above this density the second (backside) emulsion is beginning to contribute, yielding an irregular curve (see 1230-min exposure)

standards (*12*), invisible by autoradiography, for checking the pH gradient on the stained gel.

Computer System

The present hardware system (diagrammed in Figure 1) is based on a PDP 11/60 with 256 k-bytes of memory and a hardware floating-point unit (Digital Equipment Corp., Maynard, MA 01754). Images (gel autoradiographs) are digitized by using an Optronics P-1000 rotating drum scanner (Optronics International Inc., Chelmsford, MA 01824), and a pair of 300 M-byte disc units are used for data storage (Plessey Peripheral Systems, Inc., Irvine, CA 92714). Images are displayed on a Grinnell GMR27 color video display system (512 × 512 nine-bit pixels) equipped with a trackball for controlling two cursors (Grinnell Systems Inc., San Jose, CA 95131). Most of the arithmetic computation is done in an AP-120B array processor having a 64 kilo-word main data memory and a 4 kilo-word program source memory (Floating Point Systems, Inc., Portland, OR 97223). The AP-120B is capable of executing 12 million floating point operations per second at 38-bit (8$^+$ decimal digits) precision. A large flatbed plotter (Tektronix 4663; Tektronix, Inc., Beaverton, OR 97077) is used for output of line graphic results. This hardware is now being upgraded for larger-scale studies by the addition of a Digital Equipment Co. VAX 11/780 computer, a second AP-120B array processor, a DeAnza IP8500 color CRT display (DeAnza Systems, Inc., San Jose, CA 95131), and an Optronics Gelscan 12 × 12 inch wet gel scanner.

Methods Used in the Analysis of Gel Images

Processing of two-dimensional gel images with the TYCHO system involves the five steps detailed below. Most of these can be run automatically, allowing the computer to process sets of images overnight. Most of the programs involved are written in either FORTRAN or "C," with frequently executed portions coded in MACRO (for the PDP-11) or APAL and VFC (for the array processor). The latter code is highly machine-specific, though the remainder of the system should be transportable to any machine using the RSX-11/M operating system.

Autoradiographic films are digitized on a 100 × 100 μm grid (10^4 μm^2 pixels) as eight-bit densities over either 0–2 or 0–3 *A* ranges, with use of the Optronics P-1000 scanner. The resulting image, usually 1530 × 1530 points, is stored on disk. Processing algorithms are executed by passing the image through the AP-120B array processor (usually in blocks of five to 50 lines) and storing the processed output on the same or another disk file. Figure 2*A* shows a small region of a digitized image in pseudo color (i.e., a color scale has been substituted for the autoradiographic grey scale). Throughout processing, any pixels detected as over range (i.e., defined as >255 in the original densitometered image) are preserved as 255 values.

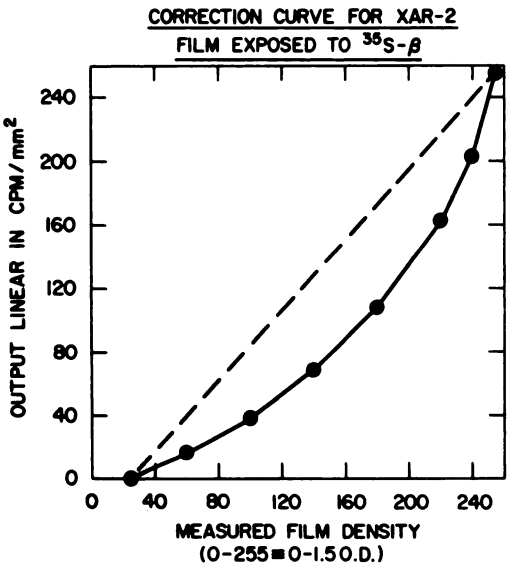

Fig. 4. Correction function for converting measured XAR-2 film density to a scale linear in counts per minute

The film's base-density (grey-scale value: 25) is subtracted at this stage. Sensitivity is significantly diminished in the low-density region (grey-scale values: 25–120) by this correction, leading to suppression of some faint spots

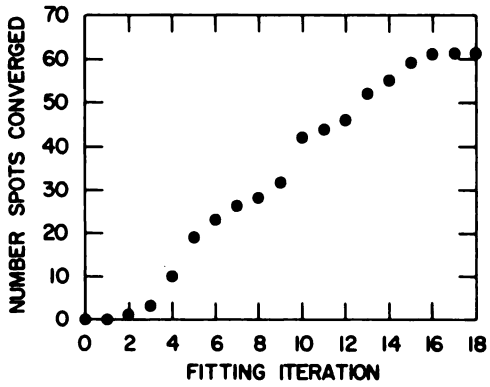

Fig. 5. Number of "converged" spots in a small area of an image as a function of fitting iteration.

Spots are considered converged if they change by less than 1% from one iteration to the next

I) Conversion of radiograph density to a linear measure of radioactivity. The exposure of X-ray film to β-particles can be calibrated by use of step-tablets in which each step has, for instance, twice the radioactivity (cpm) per square millimeter of the preceding step. When film is exposed to such a step tablet for various intervals, sensitometric curves such as those in Figure 3 are obtained. These curves are neither linear nor logarithmic over the entire usable absorbance range (0–1.50 *A* for XR-5), and hence an empirical lookup-table correction function (Figure 4) is used to translate the absorbance image from the densitometer into a corrected image linear in radioactivity (a function of the amount of label-containing protein) per area (Figure 2*B*).

An unfortunate feature of this curve is the 50% loss in sensitivity at the low-density end. To preserve information on faint spots it may ultimately be necessary to increase the image depth (dynamic range) to 10–12 bits at this point and throughout subsequent processing. Double-emulsion (rapid-process) X-ray films such as the XR-5 and XAR-2 used here suffer from the defect that only the "front" emulsion (the one touching the gel) is exposed according to a reasonable sensitometric curve; the back emulsion is protected from β-particle exposure by the film base. Thus only about half of the total achievable film absorbance (0–1.5 out of 0–3) is usable with these films. Single thick-emulsion films such as Kodak SB will probably prove superior for direct autoradiography, because they can be used over a range of approximately 0–3 *A*, even though they are generally slower and require processing in a special long-cycle film processor.

II) Smoothing, background, and streak subtraction. The general approach we have taken to the problem of background and streak subtraction is to find the minimum element in some selected region surrounding each pixel, for each pixel of the image. Convenient regions to choose are vertical or horizontal line segments, "+" or "×" shapes, or filled squares. Here we employ vertical and horizontal line segments because these allow detection and subtraction not only of general background, but of streaks as well. Streaks are subtracted because, even though they represent actual protein present in the gel, they are often not reliably attributable to any particular spot. For detection of the vertical component of background, a vertical kernel (45 pixels high and one wide, centered on the object pixel) is passed over the median-filtered (3 × 3 element) image, and the minimum value in the kernel is recorded for each pixel. This procedure tends to erode regions of high background by a distance equal to the kernel arm length (22 pixels), so the output must be corrected in a second pass in which the maximum is taken over the same kernel. Thus re-expanded, the vertical background from Figure 2*B* is shown in Figure 2*C*. This background is subtracted from the original density-corrected image, and a horizontal background calculated by a similar procedure (this time using as the kernel a horizontal line segment 50 pixels long and one pixel high). After re-expansion, this horizontal background is subtracted from the image. After filtering with a 3 × 3 element median filter, a final image appears (Figure 2*D*), essentially a collection of spots on a uniform zero background, suitable for use in a mathematical fitting procedure.

III) Spot detection. By convolving the original corrected image (Figure 2*B*) with a "+"-shaped 21 × 15 (x and y, respectively) element kernel consisting of centered cosine curves (14- and 10-pixel periods, respectively), the "sharpened" image of Figure 2*E* is obtained. Results are similar if the procedure is applied to uncorrected image data. Such a convolution produces a sharp spike wherever the image has a peak or shoulder approximately the shape of the kernel's central peak. Here, spots can easily be detected as local maxima (center element equals maximum in a 7 × 7 element "+"-shaped kernel). The set of local maxima whose pixel values in the original image are above some threshold constitutes the starting spotlist. With our present hardware the processing steps described to this point require approximately 15 min for a 1500 × 1500 point image. The speed may be substantially improved by using sampled image data (every other point in x and y, for instance) for background calculation.

IV) Gaussian fitting. Ideally, densitometry of spots on two-dimensional gels with very low sample loads would yield two-dimensional gaussian peaks (*22, 33*) having slightly different half-widths in the two dimensions. We have made use of this fact to allow mathematical modeling of the spots. Each spot in the starting spotlist has five parameters: x and y position, a default size specified by two half-widths ($\sigma_x = 3$, $\sigma_y = 2.5$ pixels), and an amplitude equal to the original image value at the spot center. This list yields a starting synthetic image as in Figure 2*F*. The spots are then fitted to the processed image (Figure 2*D*) by a least-squares method (*33*), the mutually altering effects of inter-penetrating neighbors being taken into account. This iterative process can be carried to convergence for 80 to 100% of the spots on a given autoradiograph (Figure 5). Any errors as to the number of spots in a group can be corrected manually if necessary before the final rounds of fitting. The system specifically excludes over-range regions from the fitting, because the density in these areas is not accurately known.

The processing time required for gaussian fitting depends greatly on the number of spots and the size and number of over-range regions; however, a time of 5 min per cycle for an image with 500–700 spots is typical. Depending on the circumstances, 10 to 30 cycles of fitting may be required (50 to 150 min).

A finished synthetic representation appears as in Figure 2*G*. If this is subtracted from the processed image, very little density remains unaccounted for (Figure 2*H*). The fitted spotlist, containing from 200 to 2000 spots, is quite an accurate

Fig. 6. Superimposed plots of the positions of spots in two gels after stretching and matching
Ellipses mark positions of spots from one gel, *pluses* mark spots in the other. Registration is very close throughout the pattern, fewer than 2% of the spots remain unmatched

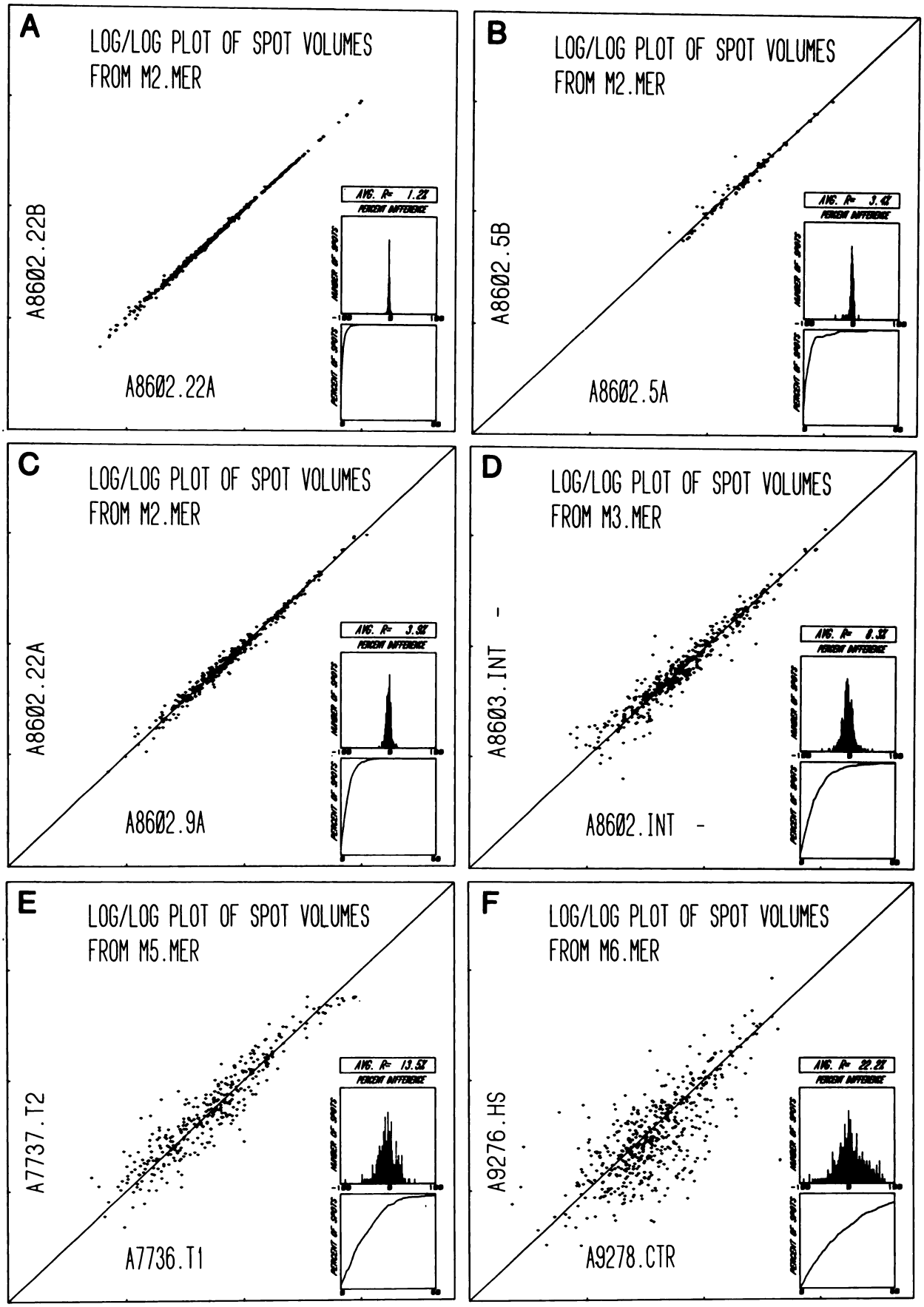

Fig. 7. Gel-to-gel comparisons

These are presented as $\log_{10}/\log_{10}$ plots of spot volumes (integrated densities). Each mark represents one spot measured on both gels. Tick marks on the axes of the large boxes indicate factors of 10 in volume. (*A*) two scans of the same autoradiograph, (*B*) two films exposed to the same gel for five days each, (*C*) two films exposed to the same gel for different periods (five days vs 22 days) after scaling of one to the other by using a quadratic function constrained to pass through the origin, (*D*) autoradiographs of two gels (same batch) of the same sample, (*E*) samples of two fibroblast cell lines obtained from identical twins (two gels from the same batch), (*F*) GM607 lymphoblastoid cells with and without a heat shock (10 min, at 45 °C; two gels from the same batch). Small boxes in the *lower right* of each panel present (*top to bottom*) the average *r* factor between the gels, a histogram showing the number of spots having a given signed percent difference (single-spot *r*-value), and the cumulative percentage of spots showing less than a certain (absolute value) percent difference

representation of the positions, shapes, and integrated densities of the spots on the original autoradiogram. In the remaining stages of data analysis only this parameterized list is used; the image itself, with its large storage requirement, can be dispensed with. Note that the default parameters are set to exclude very small spots, which almost invariably are caused by dust or physical imperfections on the film.

V) Pattern matching. To use the quantitative data obtained for comparative studies, one must establish a correct correspondence between spots on different gels. This is not a trivial problem, because the dynamic properties of isoelectric focusing pH gradients and the elastic nature of the acrylamide gel itself give rise to small, nonlinear distortions between one gel and the next. There appears to be no easy way of transforming the coordinates of one gel into those of another by means of a single polynomial transformation. However, the spot patterns are locally conserved, so it is possible to achieve near-congruence by applying sequentially (and cumulatively) a series of local stretches whose effect is made to decline exponentially from the center ("pivot point") about which the local stretch is computed.

We have implemented such a system (*34*) in which, given six to 10 initial identifications, two similar gels can be matched with high accuracy. The matching proceeds in four steps: first a fit over the whole gel, to establish overall registration, then a series of local deformations centered on the best-matched spots, a sequence of stretches centered on the nodes of a grid (usually 4 × 4) over the whole pattern, and finally another series of local stretches centered on the worst matches. At each stage, spots that are seen to correspond within certain distance and similarity (shape and abundance) limits are "matched" by making their spot numbers the same. At several points in the algorithm, the worst computer-made matches are broken to allow for readjustment. Figure 6 illustrates the capabilities of this system. In the match shown, the average residual final mismatch between 650 "matched" spots in the two gels is ~2.5 pixels (~0.25 mm), with no incorrect identifications found. The procedure requires about 2 min.

To organize data from two-dimensional gels into a coherent, useful form, it is necessary to adopt a set of "master numbers" for the protein spots seen in a given system, type of sample, or experiment. The matching program is then used to transfer the correct master spot numbers to the spots on each new gel. The data may then easily be merged into multigel files, with the abundance information for each protein properly collated. Obviously, new spots (i.e., new proteins) must be added to the master list as they are observed. This also is done by matching, except that here unmatched spots are added to the reference master list and given the next succeeding unused numbers. The master list so constructed is based on a prototype (usually a control) gel, but contains all the spots seen on any of the gels in the set to which it applies. In addition to providing an organizational framework for the abundance data, the master list can be used to standardize the fitting procedure itself by providing a "universal" starting spot list for a given preparation. To do this, we stretch the master spot list onto each new gel and then fit the stretched reference master pattern to the processed gel image. In this case each spot brings with it the correct master number at the start, and there is generally no need to rematch with the master list after fitting.

Except for the manual entry of six to 10 reference matches, all of the software described runs automatically. Provision is made to output visual data at every stage for purposes of "debugging" the procedures, monitoring production runs, and interacting with the final data. Extensive use is made of the array processor at all stages.

Table 1. Degrees of Difference between Some Pairs of Analyses

Comparison	*r*-factor, %
1. Duplicate scans, same autoradiograph	1–3
2. Duplicate autoradiographs, same gel (different films)	3–6
3. Varying exposures, same gel, scaled together:	
22 day vs 9 day	4
22 day vs 5 day	5
4. Same sample, duplicate gels (same batch of gels)	8–10
5. Samples from four different people, prepared same day, gels run in same batch	16–18
6. Lymphocytes + and − PMA	37[a]
7. Lymphocytes vs GM607	37[a]
8. Lymphocytes + PMA vs GM607	27[a]
9. Lymphocytes vs fibroblasts	41[a]
10. Two different fibroblast cell lines	11–15
11. GM607 + heat shock	22

[a] Based on comparison of spots measured in both gels; does not include proteins undetectable on one gel. PMA, phorbol myristate acetate; GM607, lymphoblastoid cell line.

"Scaling" of Spot Lists

For inter-gel comparisons, corrections must be made for the effect of differences in total detected protein that are the result of differences in amount of sample loaded or duration of autoradiographic exposure. This is done by determining the best polynomial function (as a function of integrated density) relating the two sets of spots (with furthest outliers removed), and correcting the object spot list according to this function.

In the comparisons reported here, scaling was performed with a linear or quadratic function constrained to pass through the origin.

Measure of Difference between Spot Lists

For an overall index of the similarity or dissimilarity of two sets of matched spot data, we use a mean difference defined as:

$$r_i = \left(\frac{V_i^1 - V_i^2}{V_i^1 + V_i^2}\right) \times 100$$

where V_i^k is the abundance of the ith spot on gel k. The r-value is usually taken over some sets of spots s (usually all the spots detectable on both gels). Values for r range from 0% (exact similarity) to 100% (maximum possible dissimilarity).

Results

In most cases, each spot on two-dimensional gels represents a unique protein species. The quantitation of radioactive label incorporated in each spot therefore generally provides an independent and potentially meaningful measurement of the expression of some cellular gene. For the purposes of the present general overview, however, we will concern ourselves with the overall level of difference between spot lists as mea-

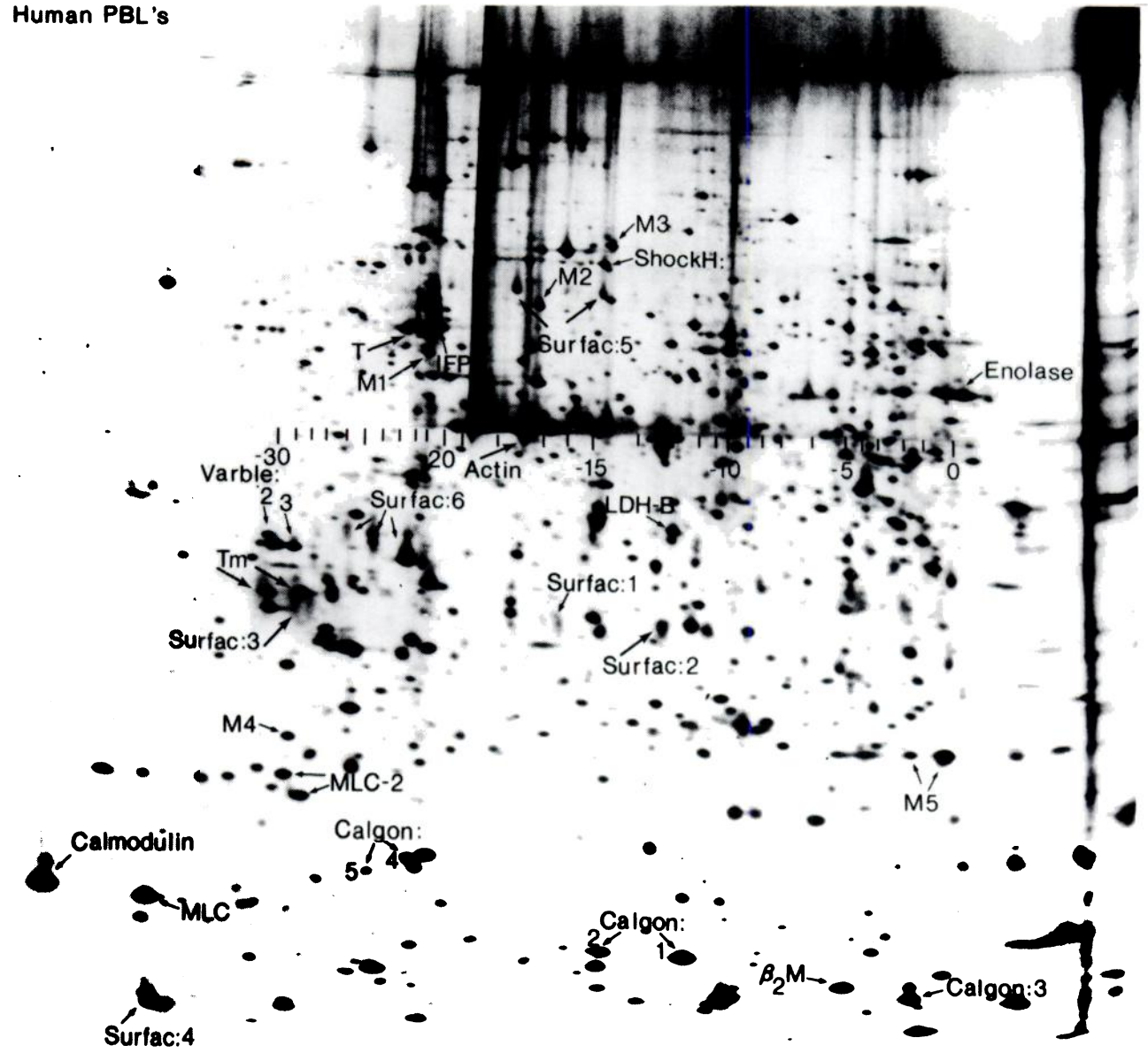

Fig. 8. Autoradiograph of a separation of human peripheral blood leukocyte proteins labeled with 35S-methionine

Several identified proteins are labeled including actin, tubulin (*T*), lymphocyte intermediate filament protein (*IFP*), β_2 microglobulin (β_2-*M*), calmodulin, enolase, lymphocyte tropomyosins (Tm), the LDH-B chain, and putative myosin light chains (*MLC*). Members of some prominent protein sets are also indicated, with their numbers in the set: surface proteins (*Surfac*), mitochondrial proteins (*Mitcon*, labeled M), heat shock protein (*ShockH*), calcium-regulated proteins (*Calgon*), and variable proteins (*Varble;* control unknown). The scale running horizontally at the level of actin shows the positions of creatine kinase charge standards (ref. *12*)

sured by the simple *r*-factor defined above, and visualized in log/log plots of spot integrated densities (volumes).

The variation in results arising from errors in different phases of the analytical system is shown in Figure 7 and Table 1. If the same autoradiogram is scanned twice on the densitometer and these two images are fitted starting with the same prototype pattern, the resulting spotlists differ by an *r*-factor of approximately 1–3% (Figure 7*A*). If two films, each exposed to the same gel for five days, are compared, the *r*-factor obtained is 3% (Figure 7*B*). Comparison of exposures of the same gel differing in duration by a factor of 4.4 (nine and 22 days) yields an *r* of 4% (Figure 7*C*) after scaling. If two different gels of the same specimen (and from the same batch of gels) are compared, the difference is in the range 8–10% (Figure 7*D*). Thus the systematic error ascribable to film variation and densitometric noise seems to be about $r = 5\%$; if gel-to-gel (within-batch) variation is added to this, the total systematic error is about $r = 8$–10%.

When we compared patterns of conventionally prepared lymphocytes from four unrelated individuals (same batch of gels), the overall *r*-factor, including interindividual differences, was about 16–18% (Table 1). In contrast, when we compared fibroblast cell lines obtained from different individuals but maintained for many months in tissue culture before testing, the *r*-factor was only 11–15% (Figure 7*E*). In each of these comparisons the differences exceeded the systematic error, with the lymphocytes exhibiting a greater degree of interindividual variation than the established cell lines.

Exposure of human peripheral lymphocytes to phorbol

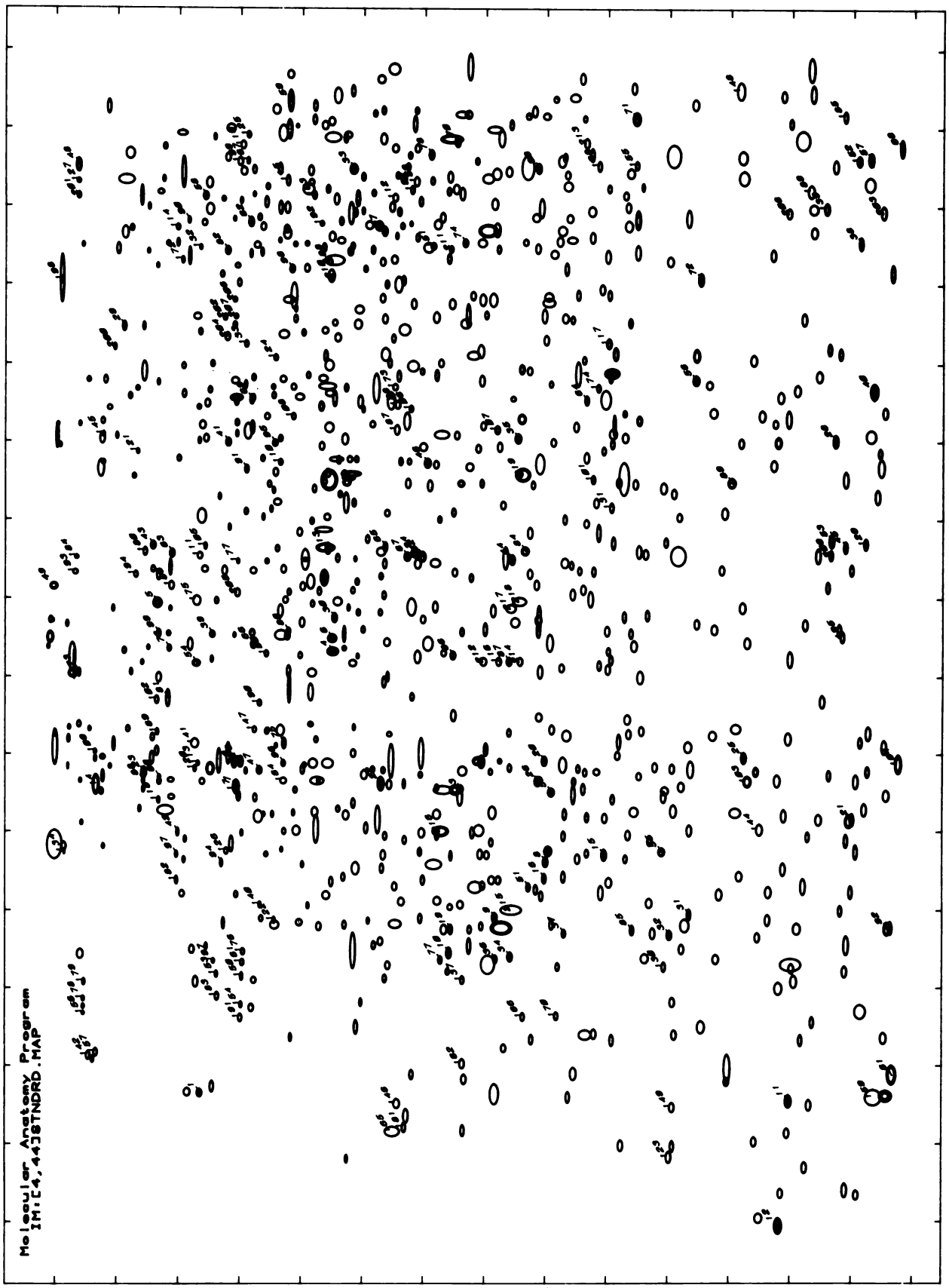

Fig. 9. Diagram of a spot list obtained from the image of Figure 8
About 1000 spots in this pattern have been designated with lymphocyte master spot numbers (only the first 220 numbers are shown here, for clarity.) These numbers are the keys to information about their respective spots in the lymphocyte protein index data base

Fig. 10. Color-television display, showing a synthetic image made from a spot list representing the pattern of Figure 8
The spot list is the same as that shown in Figure 9. Identified by use of a color overlay, the spots known to be mitochondrial proteins have been highlighted in red. Numbers adjacent to red spots are numbers in the *Mitcon* set

myristate acetate, a potent tumor promoter, produces numerous changes; such cells yield an *r*-factor of 37%. A somewhat simpler pattern of change, induced in the lymphoblastoid cell line GM607 by a 10-min heat shock at 45 °C (Figure 7*F*), gives an *r* of 22%. Even viewed as an average over all the spots, these treatments produce changes much larger than the differences observed between cells from different individuals.

Very large differences are observed when different cell types are compared; peripheral lymphocytes differ from a lymphoblastoid cell line (GM607) by 37%, and from a fibroblast line (1494) by 41%. Lymphocytes treated with phorbol myristate acetate differ from GM607 by an *r* of only 27%, indicating that the pattern of gene expression has been generally altered to resemble more closely the lymphoblastoid cell than the untreated lymphocyte (*r* of 37%, above). If, in addition to the spots measurable in both gels, one includes in the difference measure those spots detected in only one of the gels (i.e., spots unique to one of the samples), then the difference values are substantially greater.

Discussion

The results presented here indicate that the image-processing, spot-modeling, and pattern-matching procedures described are useful for the routine analysis of complex two-dimensional electrophoretic data. An average error of r = 8–10% in the measurement of 500–1000 individual proteins [including errors attributable to gel (of the same batch), film,

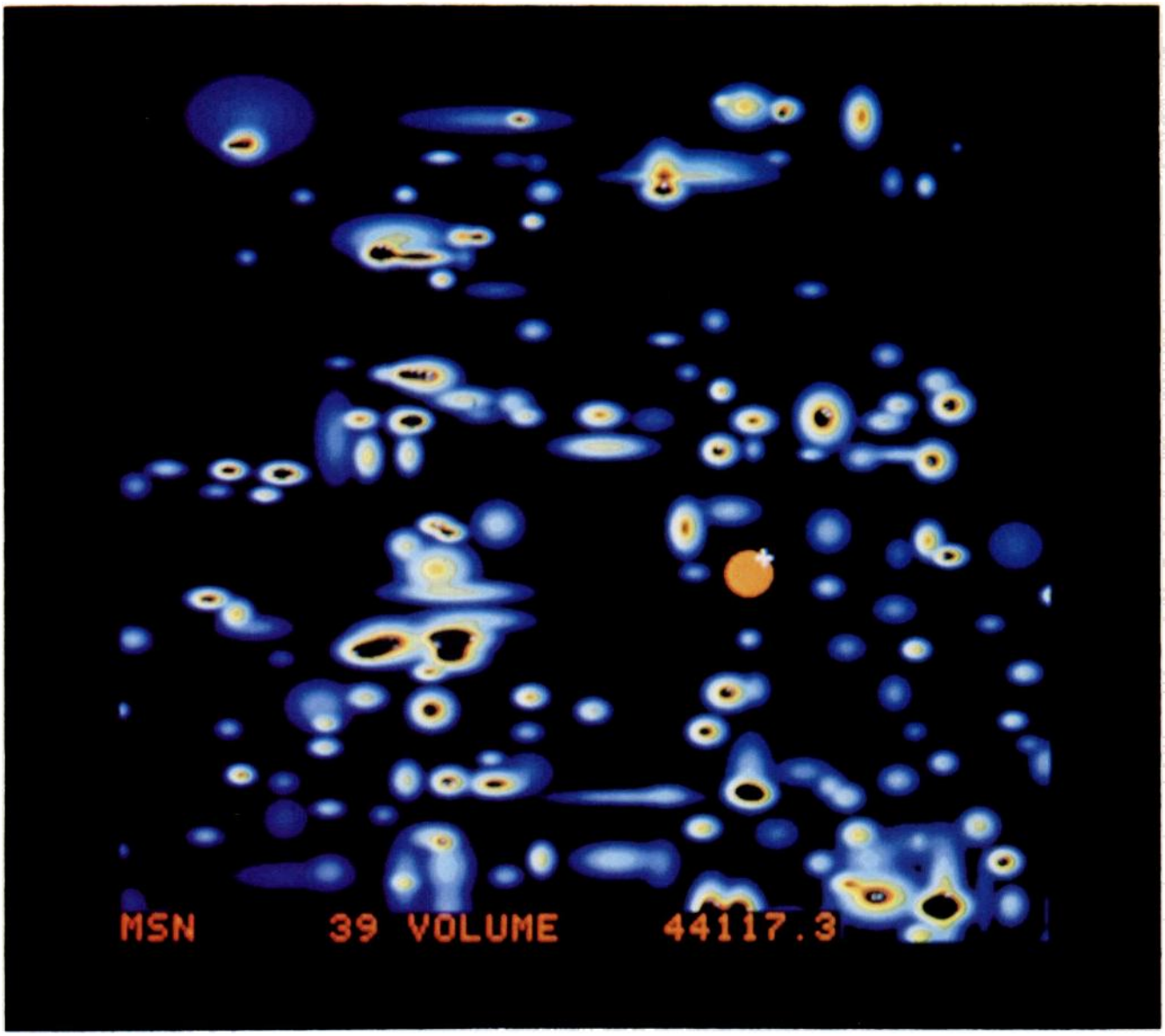

Fig. 11. Threefold enlargement of a section of the synthetic image in Figure 10 (using a different pseudo-color scheme), showing method of access to the data base

A cursor (white "+") driven by a trackball has been used to indicate a single spot (*highlighted in red*); the spot's master number (*MSN*) and abundance (*volume*) are shown automatically at the bottom of the screen

densitometer, and computer processing] seems acceptable at this stage, though it can probably be improved. Indeed, such an error compares favorably with that currently achieved in some clinical tests in which only a single enzyme is measured.

Our choice of methods for analyzing data from two-dimensional gels has been determined more by specific features of this type of data than by general image-processing considerations. Because of the physical chemistry of the gel system used, it is to be expected that "good" spots have a gaussian distribution of density in both dimensions. This feature, demonstrated by Garrels (22) in the data from fluorographed gels, has been confirmed by us for autoradiographs of gels produced by use of the ISO DALT system. We therefore decided to analyze gels as collections of such gaussian forms, using a least-squares optimization technique to bring the model into close agreement with the observed gel pattern. From the outset it was clear that this approach involved much numerical computation. In the TYCHO system an array processor is used to perform most of this computation, at a speed comparable to that of a mainframe computer, but at much lower cost. Nevertheless, the arithmetic load presented by these procedures would argue against their ultimate widespread usefulness were it not for two important factors: (*a*) the cost of computation is the one cost in technology that routinely decreases by large factors with time, and (*b*) the optimization method used can be specifically tailored to routine analysis of a particular type of sample with great savings in processing time. By taking a master pattern of spots to be quantitated and "impressing" this upon the image (as

described in *Materials and Methods*), it is possible to measure a given set of spots quite accurately with only a few iterations of gaussian fitting. We think that such an approach may represent the fastest and most nearly accurate processing method for application in a clinical setting. To the extent that measurements of enzyme amounts can replace measurements of activity, the two-dimensional technique could with one analysis replace a large number of enzymic determinations.

Another principal feature of the analysis strategy we have chosen is the adoption of master-numbered spotlists as the standard format for results. The master numbers associated with particular spots by the matching procedure identify those spots as reproducible entities (analogous to EC numbers for enzymes), the abundance or behavior of which can be correlated with experimental variables or diagnostic parameters. Figure 8 shows a typical lymphocyte two-dimensional pattern; Figure 9 shows a map in which approximately 220 of the principal proteins present in it have been assigned standard master numbers. Although master numbers could be assigned in various ways to reflect relative abundance (*27*) or position in a particular pattern, we have elected to use a non-systematic allocation procedure, to avoid implying, within the numbering scheme, any abundance relationships among spots that would later turn out to be untrue for other cell types, sample-preparation procedures, or gel systems. The master numbers we use therefore serve no purpose other than identifying particular proteins within a permanent system of computer-compatible nomenclature.

On the other hand, the requirements of thought, discussion, and, ultimately, publication necessitate the use of an additional and parallel system of evocative spot names capable of conveying information of biological interest. Often the important information concerning a spot consists of the knowledge that it belongs to the set of proteins sharing some chemical, structural, or functional characteristic. An example is the set of proteins whose production ceases in cells treated with antimitochondrial drugs such as nonactin or valinomycin (*35*). We refer to this set by the mnemonic "Mitcon" (because the proteins it contains turn out to be the major mitochondrial proteins) and assign each of the Mitcon proteins a number in the set for easy reference (Figure 10). Mitcon:5 is thus the fifth in a series of mitochondrial proteins; its master standard number, on the other hand, is 71 (MSN:71). Mnemonics are used because of their ability to convey a fairly specific impression while still being sufficiently brief to serve as a shorthand notation. So far, about 20 protein sets have been given set names of this type (*36*).

The notion of named protein sets defined by special characteristics is useful in data analysis as well as nomenclature. As is evident from the relative abundance plots of Figure 7 and the results comparing *r*-factors, overall difference measures are of limited usefulness in experimental work. Once various important functional sets are defined, it becomes possible to ask specific questions having to do with complex regulational effects—for example, does lymphocyte transformation cause a general increase in the synthesis of mitochondrial proteins as compared with cytoskeletal molecules? Large overall *r*-factor differences such as those between lymphoctyes and lymphoblastoid cells may thus be dissected to give sets of proteins showing greater and lesser difference. If the expression of a set of proteins is altered in one situation, the relative size and polarity of change it shows in other circumstances may be revealing. A large set of lymphocyte proteins whose expression is altered by treatment with phorbol myristate acetate also differs between lymphocytes and lymphoblastoid cells, and in the same sense (N.L.A., unpublished observations). This explains why lymphocytes so treated are more similar to lymphoblastoid cells (r = 27%) than either is to control lymphocytes (r = 37% and 37%).

As a data base of interesting information concerning individual proteins accumulates, the general problem of useful human interaction with such a data base emerges. It is fortunate that two-dimensional electrophoretic separations present us at the outset with data in a form we can easily appreciate. The usefulness of star charts, terrestrial maps, and a variety of other forms of visual information is based on our having a natural facility for interpreting diagrammatic two-dimensional displays. It therefore seems expedient to retain the two-dimensional pattern as an interface between man and his proteins. What is required is an efficient means of extracting information about particular spots. Figure 11 illustrates a computer-based method for retrieval of master number and abundance (spot volume) data by means of a cursor pointing out one spot in a color-television image of part of a two-dimensional pattern. The image is synthesized from a spotlist by reconstructing the gaussian spots of a fitted gel, and thus there is no requirement for storage of image data. Ultimately, the entire data base of information about the spots will be accessible through this type of interaction; specific spots may be interrogated for all relevant information (which will appear on the adjacent computer terminal) or specific sets of proteins, defined by some characteristic, may be highlighted (as in Figure 10) and the intersections, unions, complements, etc. of various sets examined. The synthetic image displayed in color with contrasting overlays for highlighting thus appears to be an efficient and pleasant medium through which to interact with a Protein Index-type data base.

For purposes of reference only, a combination of numbered diagrams (Figure 9), cross indices (master number vs position, spot characteristic vs master number), and color maps highlighted with protein sets (Figure 10) can make the same data more widely accessible in published form. The results might appear something like an equal blend of an atlas of the world and the Oxford English Dictionary.

Using the tools described here as part of the TYCHO system and experimental approaches presented elsewhere (*20*, *36*), we now have the means to attack directly the study of gene expression (and pathologial variations in it) against the background of the real complexity of the cell.

We wish to thank particularly our colleagues in the Molecular Anatomy Program for valuable input during the course of TYCHO's development; J. W. Tippie and the Minicomputer Systems Group, Applied Mathematics Division, for high-quality advice and assistance from the outset; Anne Gemmell for excellent technical help; the American Cancer Society for its interest and encouragement; and the Department of Energy for providing us continually with state-of-the-art equipment. This work was supported by the U.S. Department of Energy under contract No. W-31-109-ENG-38 and by the American Cancer Society under grant No. AD43.

References

1. Raymond, S., and Weintraub, L., Acrylamide gel as a supporting medium for zone electrophoresis. *Science* **130**, 711 (1959).

2. Ornstein, L., Disc electrophoresis. 1. Background and theory. *Ann. N.Y. Acad. Sci.* **21**, 321–349 (1964).

3. O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021 (1975).

4. Klose, J., Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissue. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243 (1975).

5. Scheele, G. A., Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *J. Biol. Chem.* **250**, 5375–5385 (1975).

6. Iborra, G., and Buhler, J.-M., Protein subunit mapping. A sensitive high resolution method. *Anal. Biochem.* **74**, 503–511 (1976).

7. Vesterberg, O., and Svensson, H., Isoelectric fractionation, analysis, and characterization of ampholytes in natural pH gradients. IV. Further studies on the resolving power in connection with separation of myoglobins. *Acta Chem. Scand.* **20,** 820–834 (1966).

8. Weber, K., and Osborn, M., The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis. *J. Biol. Chem.* **244,** 4406–4412 (1969).

9. Laemmli, U. K., Cleavage of the structural proteins during the assembly of the head of bacteriophage T_4. *Nature* **227,** 680–685 (1970).

10. Anderson, N. G., and Anderson, N. L., Analytical techniques for cell fractions. XXI. Two-dimensional analysis of serum and tissue proteins: Multiple isoelectric focusing. *Anal. Biochem.* **85,** 331–340 (1978).

11. Anderson, N. L., and Anderson, N. G., Analytical techniques for cell fractions. XXII. Two-dimensional analysis of serum and tissue proteins: Multiple gradient-slab electrophoresis. *Anal. Biochem.* **85,** 341–354 (1978).

12. Anderson, N. L., and Hickman, B. J., Analytical techniques for fractions. XXIV. Isoelectric point standards for two-dimensional electrophoresis. *Anal. Biochem.* **93,** 312–320 (1979).

13. Anderson, N. G., and Anderson, N. L., Molecular anatomy. *Behring Inst. Mitt.* **63,** 169–210 (1979).

14. Anderson, L., and Anderson, N. G., High resolution two-dimensional electrophoresis of human plasma proteins. *Proc. Natl. Acad. Sci. USA* **74,** 5421–5425 (1977).

15. Anderson, N. G., Anderson, N. L., Tollaksen, S. L., et al., Analytical techniques for cell fractions. XV. Concentration and two-dimensional electrophoresis analysis of human urinary proteins. *Anal. Biochem.* **95,** 48–61 (1979).

16. Edwards, J. J., Anderson, N. G., Nance, S. L., and Anderson, N. L., Red cell proteins. I. Two-dimensional mapping of human erythrocyte lysate proteins. *Blood* **53,** 1121–1132 (1979).

17. Anderson, N. G., Anderson, N. L., and Tollaksen, S. L., Proteins in human urine. I. Concentration and analysis by two-dimensional electrophoresis. *Clin. Chem.* **25,** 1199–1210 (1979).

18. Giometti, C. S., Anderson, N. G., and Anderson, N. L., Muscle protein analysis. I. Development of high resolution two-dimensional electrophoresis of skeletal muscle proteins for analysis of microbiopsy samples. *Clin. Chem.* **25,** 1877–1884 (1979).

19. Anderson, N. L., and Anderson, N. G., The potential of high resolution protein mapping as a method of monitoring the human immune system. In *Biological Relevance of Immune Suppression*, J. H. Dean and M. Padarathsingh, Eds., Van Nostrand–Reinhold Co., New York, NY, 1981, pp 136–147.

20. Anderson, N. L., Edwards, J. J., Giometti, C. S., et al., High-resolution two-dimensional electrophoretic mapping of human proteins. In *Electrophoresis '79*, B. Radola, Ed., W. de Gruyter, New York–Berlin, 1980, pp 313–318.

21. Anderson, N. G., and Anderson, L., Automatic chemistry and the Human Protein Index. *J. Autom. Chem.* **2,** 177–178 (1980).

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

23. Lutin, W. A., Kyle, C. F., and Freeman, J. A., Quantitation of brain proteins by computer analyzed two-dimensional electrophoresis. In *Electrophoresis '78*, N. Catsimpoolas, Ed., Elsevier/North Holland, Amsterdam–New York, 1979, pp 93–106.

24. Bossinger, J., Miller, M. J., Vo, K.-P., et al., Quantitative analysis of two-dimensional electrophoretograms. *J. Biol. Chem.* **254,** 7986–7998 (1979).

25. Lipkin, L. E., and Lemkin, P. F., Data-base techniques for multiple two-dimensional polyacrylamide gel electrophoresis analysis. *Clin. Chem.* **26,** 1403–1412 (1980).

26. Vo, K.-P., Miller, M. J., Geiduschek, E. P., et al., Computer analysis of two-dimensional gels. *Anal. Biochem.* **112,** 258–271 (1981).

27. Lester, E. P., Lenkin, P., Lipkin, L., and Cooper, H. L., Computer-assisted analysis of two-dimensional electrophoresis of human lymphoid cells. *Clin. Chem.* **26,** 1392–1402 (1980).

28. Zimmer, H.-G., Kronberg, H., and Neuhoff, V., Quantitative evaluation of chromatograms. *Proc. Fourth Int. Joint Conf. on Pattern Recognition*, IEEE Computer Society, 1979, pp 834–836.

29. Alexander, A., Cullen, B., Emigholz, K., et al., A computer program for displaying two-dimensional gel electrophoresis data. *Anal. Biochem.* **103,** 176–183 (1980).

30. Capel, M., Redman, B., and Bourque, D. P., Quantitative comparative analysis of complex two-dimensional electrophorograms. *Anal. Biochem.* **97,** 210–228 (1979).

31. Butler, J. W., Haasl, J. R., Hodges, D., et al., *ALICE Reference Manual*, TM-231, Applied Mathematics Division, Argonne National Laboratory, Argonne, IL 60439.

32. Anderson, N. G., Anderson, N. L., and Tollaksen, S. L., Operation of the ISO-DALT System. *Report ANL-BIM-79-3*, 1979, Argonne National Laboratory, Argonne, IL.

33. Taylor, J., Anderson, N. L., Coulter, B. P., et al., Estimation of two-dimensional electrophoretic spot intensities and positions by modeling. In *Electrophoresis '79*, B. Radola, Ed., W. deGruyter, 1980, pp 329–339.

34. Taylor, J., Anderson, N. L., and Anderson, N. G., A computerized system for matching and stretching two-dimensional gel patterns represented by parameter lists. In *Electrophoresis '81*, to be published.

35. Anderson, L., Identification of mitochondrial proteins and some of their precursors in two-dimensional electrophoretic maps of human cells. *Proc. Natl. Acad. Sci. USA* **78,** 2407–2411 (1981).

36. Anderson, N. L., Studies of gene expression in human lymphocytes using high-resolution two-dimensional electrophoresis. In *Electrophoresis '81*, to be published.