CLIN. CHEM. 28/4, 861-866 (1982)

Design and Implementation of a Prototype Human Protein Index

J. Taylor, N. L. Anderson, A. E. Scandora, Jr., K. E. Willard, and N. G. Anderson

This paper describes information-handling aspects of the TYCHO I analysis system (*Clin. Chem.* **27:** 1807–1820, 1981), which analyzes two-dimensional electrophoresis gels, matches the individual protein spots with those in a reference pattern, and stores various information—including spot measurements, identifications, treatment profiles, set memberships, and comments—in a computerized database. This and additional information such as amino acid composition and cellular localization is then accessible from an interactive program that includes a pictorial user interface and presents much of the data in graphical form. Use of the TYCHO I system is illustrated by examples drawn from analyses of gel patterns from human leukocytes.

Additional Keyphrases: *TYCHO system · two-dimensional electrophoresis · computerized data acquisition and handling · proteins in leukocytes*

The collection and organization of information from two-dimensional (2-D) electrophoresis gels into an accessible and useful form is a large and difficult problem. Such patterns typically exhibit hundreds or even thousands of separate protein spots. It is the goal of the TYCHO I analysis system to measure as many spots as feasible, to identify each spot by matching the pattern with other patterns in the experiment, and to make the measurements and identifications easily accessible to the researcher. Intercomparison of multiple gel patterns within an experiment and comparison of results between or among different experiments present a substantial problem in record keeping and in gleaning interesting and important information that may be present in the data.

Basic aspects of the TYCHO I system have been described in a previous paper (*1*). Here we describe in greater detail the organization of the experimental data and the programs that maintain the database of information comprising the prototype Human Protein Index (*2*). Particular attention is given to those parts of the system that allow the user to update and interrogate the database. Examples in this paper are drawn from a prototype protein index, constructed by using gel patterns from human leukocytes.

Implementation and Examples

The Computer System

The TYCHO I analysis system is implemented with a PDP-11/60 minicomputer (Digital Equipment Corp., Marlboro, MA 01752). Figure 1 shows a block diagram of the major hardware parts of the computer system, detailed elsewhere (*1*). A Floating Point Systems AP-120B array processor is attached to the bus of the minicomputer to handle the heavy computational loads of the analysis process. An Optronics P-1000 rotating drum densitometer is used to scan the autoradiographic films. The digitized images are temporarily stored on 300 Mb disk storage systems until they have been processed. A Grinnell GMR-27 color raster scan device is used to display the images. It contains nine planes of 512 × 512 bit refresh memory and is equipped with a trackball-controlled cursor for interactive work. The system is equipped with a Tektronix 4025 graphics terminal and a Tektronix 4663 flatbed plotter for additional graphics capability.

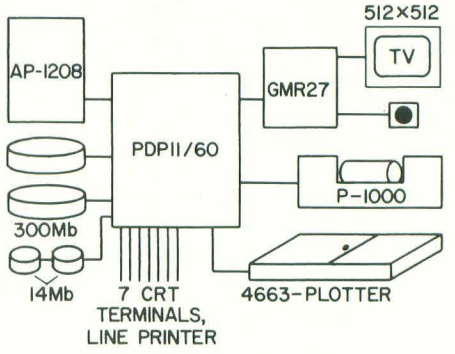

Fig. 1. Block diagram of the major hardware components of the TYCHO I analysis system

Design of the Prototype Protein Index Database

The data of the prototype human protein index are organized in a set of hierarchically arranged files as shown in Figure 2. Each solid rectangular box in the schematic diagram represents a different computer file, and the lines connecting the boxes represent the primary data paths. The boxes enclosed by dashed lines at the top level of the figure represent files relating to a single experiment. The dashed ovals enclose the files for tissue-specific and other protein-index databases. The translation table, represented by the circle in the middle of the figure, is used to establish correspondences among spots from different tissues or experiments. The flow of information generally proceeds from top to bottom, with data being culled at each level to eliminate extraneous information. The structure shown in Figure 2 is not rigorously imposed on the researcher; he may invent other data organizations for use with this same system if he so desires.

The database file types. As now implemented, the database contains five different file types:

1) Image files. Image files contain 2D gel images in digitized form.[1] Both scanned and processed images are stored in these files. Because image files are very large as compared with the file types described below, they are considered expendable and are deleted when the researcher is satisfied that the models to be described below adequately represent the patterns being analyzed.

2) Spot-list files. Each spot-list (or spot-parameter) file contains the model parameter values describing a single pattern. Each model consists of a set of 2D gaussian distributions, with each distribution corresponding to a protein or protein fragment (*3*, *4*). The individual distributions are characterized by five parameters: an amplitude, x and y positions, and x and y widths. The distributions are restricted to be aligned along the x and y axes. A spot-list file exists for each gel, for each experiment (called an "experiment spot-list file"), and for each tissue type being studied (called a "master spot-list file"). Except for the experiment and master spot-list files, these files

Molecular Anatomy Program, Argonne National Laboratory, Argonne, IL 60439.

[1] Nonstandard abbreviations: 2D, two-dimensional; PIF, protein-index file; ESN, experiment spot number; MSN, master spot number; and SPF, MSFE, PLRESP, MERLST, names of computer programs.

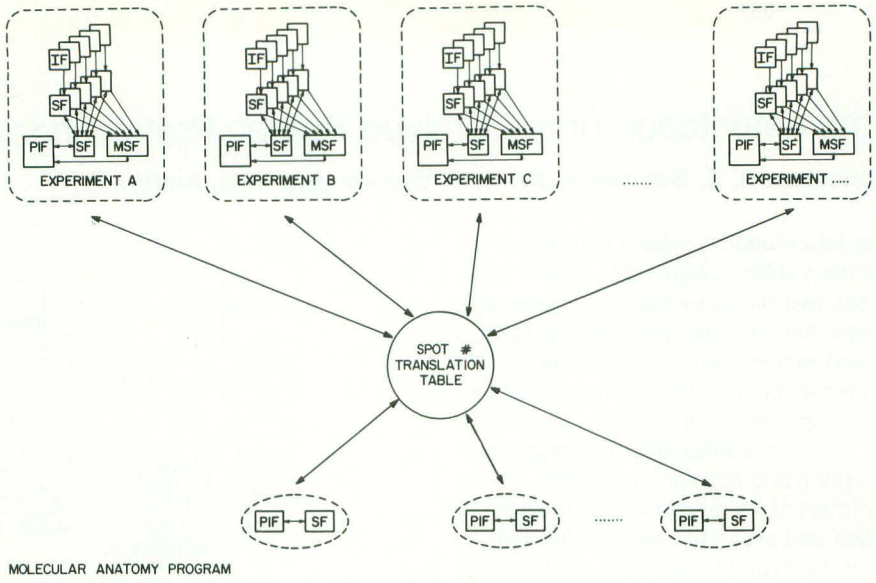

Fig. 2. Major file components used by the TYCHO I analysis system

Blocks marked *IF* are image files, those marked *SF* are spot-parameter files, those marked *MSF* are merged-spot files, those marked *PIF* are protein-index files

are originated by the spot-detection part of the image-processing system and are updated by the optimization system. They are read and written by using direct calls to the operating system executive. Programs exist both to list these files and to display their synthesized images on the display monitor. They may also be converted to image form for other purposes.

3) Merged parameter files. Files of this type contain the spot-parameter values for a whole experiment, organized so that all information for a single spot is accessible by reading a single record. These are secondary files in the sense that they are built from the spot-list files and contain no additional information other than the proper correspondence among spots on different gels. They are generally kept for the duration of the experiment, with selected records transferred to a protein-index file, where they are kept permanently and known as "treatment-profile entries."

4) Protein-index files. The protein-index file (PIF) contains spot identifications, researcher observations, spot population lists, and other types of information. This file has been implemented by using a keyed access file structure. For reasons of data integrity, the only access to this file is restricted to a specialized database management program. All other programs desiring to retrieve or enter information must request this service from the database management program.

5) The spot-number translation-table file. This file, represented by the circle in Figure 2, contains the information necessary to convert the spot numbers from one master or experiment spot file to another. The assignment of master and experiment spot numbers will be described in greater detail below.

Processing of Data

Typically, we process two-dimensional electrophoresis data in several steps. First, the autoradiograph is scanned and its image processed; then a model is developed and optimized. These models are then compared with others of the same experiment. Results of the experiment are then examined by use of interactive programs, and selected information is added to the PIF for the particular tissue type. Briefly, these steps are as follows.

The autoradiographs are scanned with an Optronics P-1000 rotating drum densitometer. The pixel spacing is either 100 or 200 μm, and the optical density range is 0 to 2.0. The measurements are made to eight-bit accuracy. Typical scanned images contain up to four million bytes of data. The images are then analyzed with a sequence of specialized transforms and operations that are programmed into the array processor. These operations include film response correction, noise filtering, streak and background removal, and spot detection. The products of these operations are a processed image and a preliminary set of parameters for a model of the image. The parameters of the model are then optimized in a least-squares sense. This procedure has been described in a previous paper (*4*). Recently the performance of the optimization has been improved by transferring more of the calculations to the array processor, but the basic technique remains the same.

Handling of Sets of Gel Patterns

Most experiments involving the 2D electrophoresis method generate many separate patterns, and generally each of these patterns should be compared with all other members of the set. The number of pairs to be examined increases rapidly with the number of patterns, and studies with 100 or more separate gels are essentially not feasible with this method. The obvious remedy is to compare each gel pattern with a single master pattern, which must contain all of the spots being studied in the whole experiment. Because no single gel is guaranteed to contain all of the spots, the master must be a composite of many patterns. One might argue that a composite could be generated by pooling all the samples and running an extra gel, but this is not always acceptable, because proteins appearing in only a small fraction of the samples may not be detectable in the pooled image. Pooling of samples is particularly inconvenient for progressive studies in which additional samples are examined in an experiment over an extended period.

Our approach is to use only the model patterns (represented by the data in the spot-list files) for the comparison and to generate an experiment pattern (also a model) incrementally from pairwise comparison with each of the other patterns in the experiment. The first experiment pattern is made by copying one of the patterns in the study and assigning a unique experiment spot number (ESN) to each individual spot. Each pattern (model) in the study is then stretched into registration with the experiment spot file pattern (5), and homologous spots are identified and tagged with the appropriate ESNs. The experiment pattern is updated as necessary to add missing spots that are discovered in subsequent gels. Thus, at the end of the analysis the experiment pattern consists of a composite of many gel patterns. It contains every distinct

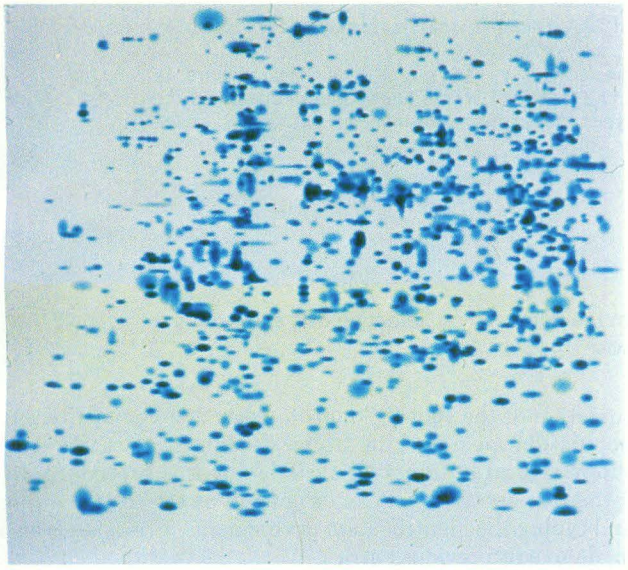

Fig. 3. Computer display of a typical human leukocyte pattern as represented by a gaussian model

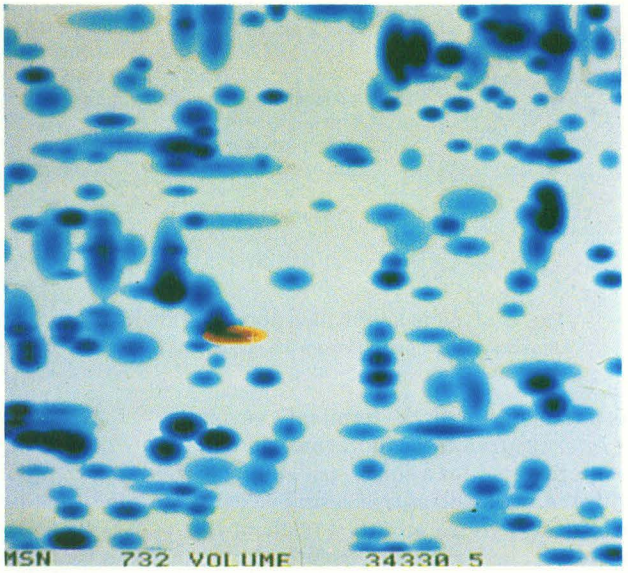

Fig. 4. Closeup view of part of the pattern in Fig. 3 as represented by a gaussian model

The spot highlighted in red is one that the user has selected by the trackball for some special operation

spot detected in the entire set of gels. Its form is simply a list of five parameters per spot, which can easily be converted to an image for display purposes. The assignment of ESNs to each protein spot for all gel patterns allows measurements to be added to the database, tagged by spot identification. Comparisons among several gels are then relatively easy.

A protein-index file and a master spot-list file are maintained for each tissue or body fluid under investigation. The master spot-list files are constructed from the experiment spot files in the same way that the experiment spot files are constructed from the individual pattern spot-list files, i.e., by accretion. The spot numbers for the spots in the master spot-list file are called "master spot numbers" (MSNs). We store spot reference data in the PIF, using ESNs for experiment PIFs and MSNs for the tissue-specific PIFs. The spot-numbering system for the ESNs and the MSNs are connected by entries in the spot-numbering translation file.

Comparison of data from several patterns usually requires that the spot-intensity data be scaled or normalized for each gel pattern. The method used here is to scale the volumes and adjust the amplitudes accordingly. The scaling constants are usually not known *a priori*, because variation in the production of gel patterns is still too great to provide numbers for absolute normalization. The scaling must be undertaken with great care in order not to mask out the very effects that the researcher is studying. Various scaling and normalization options are available in this system, and the user must choose the most appropriate. The simplest method is to require that the sum of the volumes of all the matched spots for each of the patterns be the same as the corresponding sum of its experiment spot-file pattern. This sum may also be based on a subset of matched spots (see *Spot populations*, below). Alternatively, the researcher may elect to fit a linear, quadratic, or cubic function to the volumes of the spots in the experiment spot file as a function of the volumes of the spots in the pattern being normalized. The spot volumes are then replaced with their "scaled" values. Outliers may be excluded if desired. The normalization function is plotted on the graphics terminal for inspection, and the user may accept or reject that normalization.

The User Interface to the Database

The exact form or implementation of the database itself is not particularly crucial. What is more important is the user interface to the information stored in the spot-list, merged-parameter, and protein-index files. The primary user interface of the TYCHO I system is designed to be highly interactive, easy to use, and powerful. It consists of one controlling program (called SPD) and six slave tasks. Data from a spot-list file, the merged parameter file, and a protein index file are accessible. An image of the pattern under study, synthesized from the gaussian distribution functions, is displayed on the image display monitor. A full 512 × 512 pixel image can generally be calculated and displayed in less than 10 s, and subsequent display operations require considerably less time. The user may display either the entire pattern (Figure 3) or some closeup portion of the pattern (Figure 4). In addition, he may elect to display the actual scanned image from which the model was derived. A toggle switch on the cursor-control box of the display-unit controls which image is being displayed. All images are displayed in pseudo-color (i.e., artifactual color). Several major functions and concepts are built into the SPD program. These are described in the following paragraphs. In the description below, the acronym MSN is used generically to mean either an ESN or an MSN, depending on the context.

Choosing a spot. Many operations require the designation of a particular spot. Because it is unreasonable to require the user to memorize the MSNs, the program allows him to choose a protein spot by positioning the trackball-controlled cursor. So that there can be no doubt about which spot is being selected (as might be the case with overlapping spots), the chosen spot is highlighted with a contrasting color. Simultaneously its MSN and integrated volume appear on the screen. (The red spot in Figure 4 illustrates this operation.) The user then pushes a button next to the trackball to signify that this is indeed the spot he wishes to designate.

Editing the spot file. The SPD program allows the user to edit the spot file to remove artifacts and correct any inaccuracies from the automatic spot detection, and to interrogate or add to the protein index file. Spots may be deleted singly, by using the trackball interaction described above, or in groups, by using the flagged sets (which will be described later). It is sometimes necessary to add a spot that was missed by the automatic detection system. This is done by positioning the trackball, creating a spot in the image, and then adjusting

```
Spot Number                    36
This spot is the one in which CSG and NLA found a charge variant in fibroblast
line 1493.  There is a JBC paper on the variant.
     In a GM607 heat shock experiment (E96), this form of the nonmuscle
tropomyosin showed much reduced synthesis following heat shock compared to
the adjacent spot (MSN = 9).  The effect is transient and approximately equal
rates of synthesis were resumed after 24 hr.


Membership of                  36
   CYTOSK
   MAMINV
   MANCON
```

Fig. 5. Typical information about a spot, made available from the database on selecting that spot with the trackball
The acronyms at the bottom are the populations of which this spot is a member

its parameters with the terminal's keypad.

Interrogating and adding information to the PIF. The PIF is interrogated for a single spot simply by selecting the spot with the trackball, as was described earlier. All information previously entered into the PIF about that spot is then displayed at the terminal. This information includes the identification of the spot (if known), researcher comments, and population memberships (to be described later). Figure 5 shows typical information available by this procedure. Information about a spot can be entered by selecting the spot and typing the identification or comments. Provision is also made to edit previous entries.

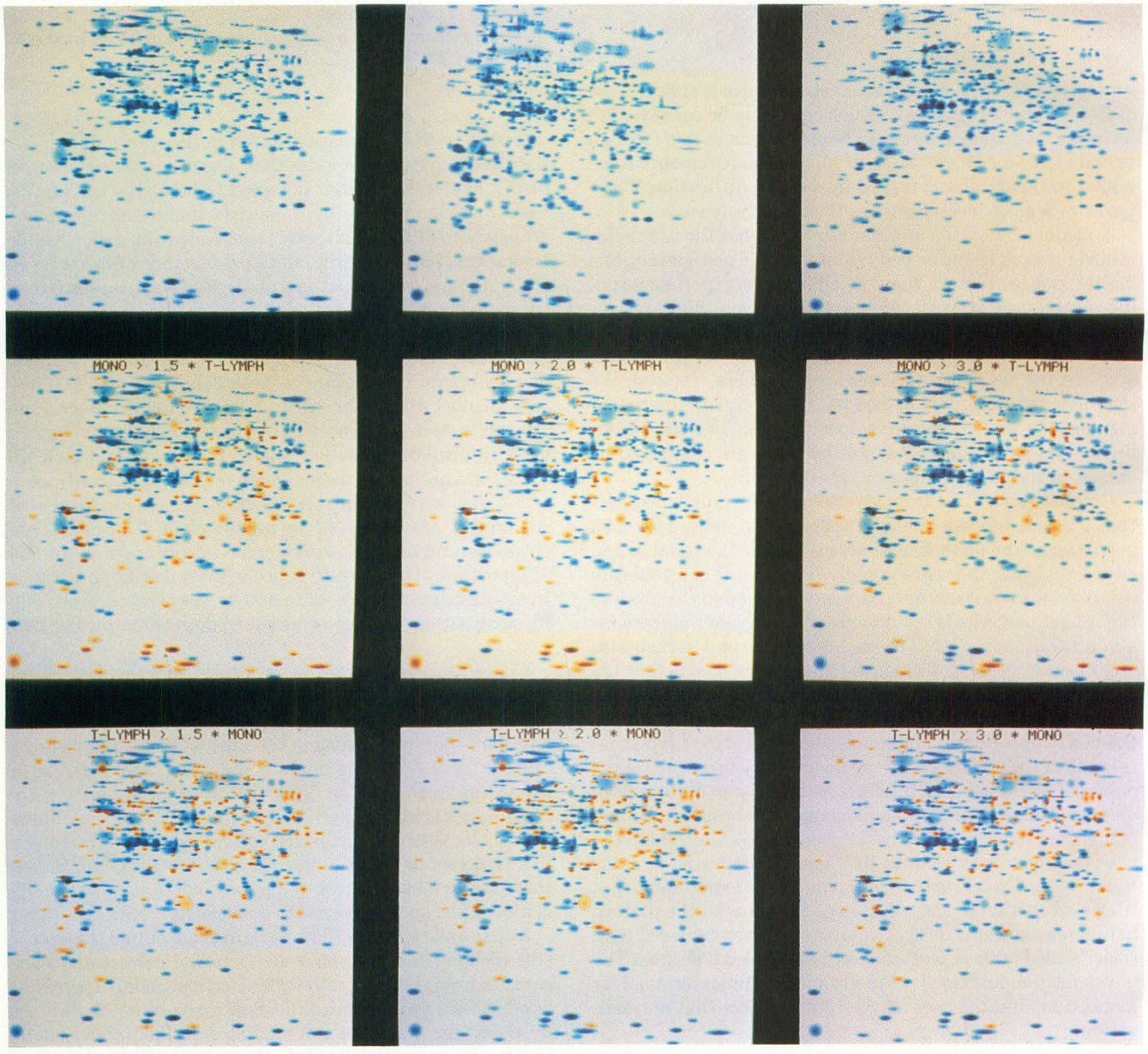

Fig. 6. A panel showing information available from computer programs SPD and MSFE
All frames are synthetic images from gaussian models. The data are from the cell-sorting experiment described in the text. The top row, left to right, shows the computer display of the unsorted pattern, the monocyte pattern, and the pattern from the T-cell groups. The second row shows sets of proteins that are more abundant in the monocytes than in the T-lymphocytes by the factors shown in each frame. The bottom row shows similar data for the T-lymphocytes

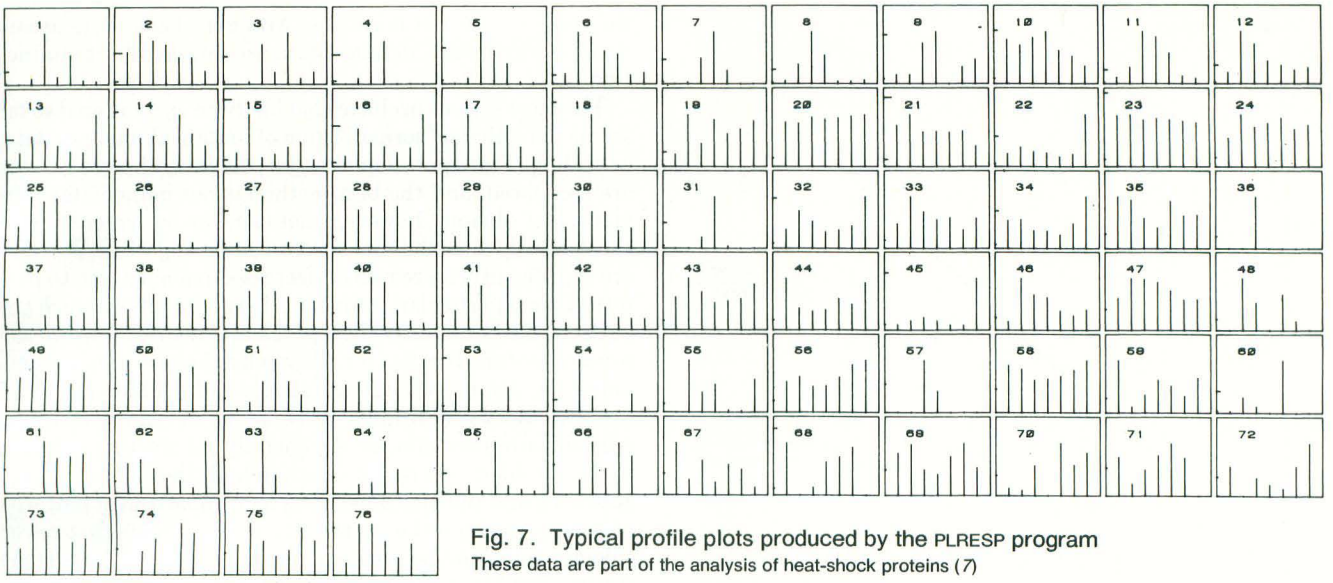

Fig. 7. Typical profile plots produced by the PLRESP program
These data are part of the analysis of heat-shock proteins (*7*)

Flagging sets of spots. Sets of spots may be flagged (*a*) by parameter value (amplitude, volume, etc.), (*b*) according to whether or not they were matched to spots in the master pattern, (*c*) according to their convergence status from the parameter optimization programs, (*d*) by membership in populations recorded in the PIF part of the database, and (*e*) by comparison with data from other gels. Flagged spots are distinguished from other spots on the color monitor by displaying them in a contrasting color. Figure 6 shows several examples of flagged spots. Flags may also be set or reset individually under manual control. Once flagged, the MSNs of the flagged spots may be stored for future reference as a spot population in the PIF file, the list of spots may be stored temporarily in a set record (internal to the SPD system of programs), or the spots may even be deleted from the spot file.

Spot populations. Spot populations one of the most useful types of information stored in the database. They allow the concise summary of many experiments and an easy presentation of co-regulated groups of proteins. The user is allowed (and encouraged) to enter comments into the PIF with each population. These comments are then printed on the terminal each time the population is used to flag a group of spots.

Set records. Sixteen "set records" exist, which allow the user to store temporarily sets of flagged spots and to perform special set operations. Unlike populations stored in the PIF file, the set records do not require that MSNs be assigned to the member spots. Thus a file can be examined before it is compared with the master pattern. The purpose of these records is to allow Boolean operations on sets of flagged spots, i.e., to ask questions of the form: "What spots belong to both of two populations?". New sets of flagged spots may be defined by using Boolean set operations on any two set records. The three operations supported are the union, the intersection, and one set minus the other.

Using the Merged-Parameter Files

Several programs have been developed to examine different aspects of the merged-parameter files. Three of these—MSFE, PLRESP, and MERLST—are discussed below. The programs MSFE and PLRESP both produce spot populations and are therefore accessed through the SPD program.

Comparison of spot intensities from different gels. The program MSFE takes the current flag set of SPD and produces a subset according to certain user-defined criteria. For example, the user may wish to select all spots in the current set that are 50% more abundant on gel *a* than on gel *b*. The results of this program become the new current flag set of SPD.

Figure 6 shows a panel of several spot populations produced by this program. The data were taken from autoradiographs of human leukocyte subtypes purified by fluorescence-activated cell sorting.[2] Fresh human peripheral blood leukocytes were isolated by use of Ficoll-Paque (Pharmacia Fine Chemicals, Piscataway, NJ 08854) gradient centrifugation, reacted with monoclonal antibodies (Ortho-clone; Ortho Pharmaceutical Corp., Raritan, NJ 08869) specific for the indicated subset, and then the monoclonal antibody was tagged with a fluorescent antibody. The cells were sorted on a FACS-II (Becton-Dickinson FACS Systems, Sunnyvale, CA 94086) into several groups based on the intensity of fluorescence. After cell sorting, the purified subgroups were labeled overnight with [^{35}S]methionine and 2D gel analysis was performed as previously described (*6*). The cell groups used in this example are T-lymphocytes, which stain brightly with tagged antibody OKT.3, and monocytes, which stain brightly with tagged antibody OKM.1. The top row shows the unsorted pattern, as well as the monocyte and T-cell group. All patterns were scaled by requiring that the total abundance of the matched spots be constant. The second row shows the unsorted pattern with populations of spots where the abundance in the monocyte pattern exceeded the abundance in the T-cell pattern by three different factors, as labeled in each frame. Spots that were quantitatively uncertain because of streaks or proximity to severely overexposed regions were excluded from this analysis. The bottom row shows similar data, where the abundance of the T-cell sample exceeded the abundance in the monocyte sample. It should be remarked that these data are presented only as a rather simple example of the use of the TYCHO system. Establishment of amino acid compositions of all the spots and comparison of gels, for instance, will require analysis and comparison of many more patterns.

Treatment-profile examination. The program PLRESP is designed to facilitate the search for coregulated sets of proteins. Plots are produced on the Tektronix 4025 display terminal or the Tektronix 4663 flatbed plotter of the treatment profile for each protein. Figure 7 shows an example of such plots, taken from a study of the heat-shock proteins (*7*). Each box corresponds to a different spot, and each bar in the box

[2] Willard, K. E., et al., manuscript in preparation.

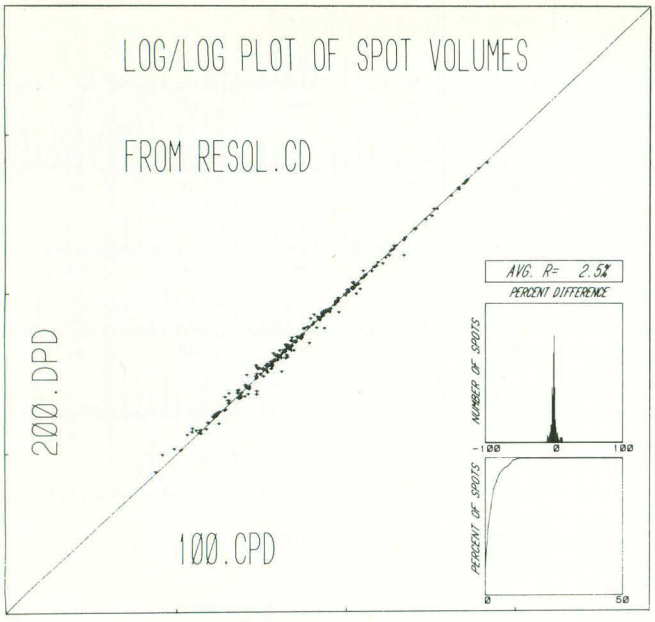

Fig. 8. A typical plot produced by the MERLST program

The volumes of spots from a 100-μm image are plotted along the *x*-axis, and the volumes from a 200-μm image of the same gel are plotted along the *y*-axis. The *insets* show the average *r*-value between the two patterns, a histogram of the percent difference, and a cumulative distribution of the absolute percent difference

is a separate measurement of the spot volume, from different gels. The user can then search for classes of similar-looking profiles. A single prototypical profile can be chosen and the analysis repeated with the correlation coefficient of each distribution with the prototype displayed. Sets of proteins with correlation coefficients above a given threshold then replace the current flag set of the SPD program.

Pairwise parameter comparison between patterns. It is often important to have an indication of the overall similarity of two patterns. The MERLST program was developed to produce scatter plots from any two patterns from the merged-list file. In addition it calculates the average *r* (correlation coefficient) value (*1*), indicating the degree of scatter. Figure 8 shows a typical plot from this program comparing the data of the same gel pattern scanned and analyzed at 100- and 200-μm resolution.

Discussion

The analysis methods used by TYCHO I have been demonstrated to be a feasible approach to the processing and comparison of 2D electrophoresis images. A hierarchical organization of the information, as shown in Figure 2, appears far more workable than simply including all measurements in a single massive file. The degree of inter-run variability in patterns still makes it far easier to compare patterns within an experiment than among experiments. The hierarchical organization minimizes the number of matching steps, both within and among experiments, and allows the researcher to cull his data and carry forward only specifically relevant information. The database is thus kept to a manageable size, and the researcher is not inundated with extraneous data at any level. Nevertheless, detailed data are retained for examination.

The most serious problem that has been encountered so far seems to be that of normalization of protein abundance data, or scaling. This operation must be performed before gel sets are compared, and the best method is not immediately obvious. It *is* obvious, however, that using an incorrect method can seriously alter the conclusions concerning which proteins are significantly increased or decreased from pattern to pattern. This is particularly true for gel experiments in which the patterns differ radically. This problem is not restricted to the present system; it exists for all systems that are used to compare gel patterns both quantitatively and qualitatively. The problem needs additional research to develop consistent normalization techniques. A generally satisfactory solution to this problem will require work in both the computational approach and the biological basis for comparison of patterns of gene expression (i.e., is there a set of proteins that never changes and could be used as a reference base?). The current system provides tools by which different techniques may be compared.

The TYCHO I system as it exists today should not be considered complete. It is under continuing development, with additional options being added as needs arise. It has already been used to analyze a number of experiments, and pertinent results have been added to the database. The prototype database (the Leukocyte Protein Index) now plays an increasingly important role in the ongoing work of the leukocyte mapping project.

This work is supported by the U.S. Dept. of Energy under contract no. W-31-109-ENG-38.

References

1. Anderson, N. L., Taylor, J., Scandora, A. E., et al., The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.* **27,** 1807–1820 (1981).

2. Anderson, N. G., The Human Protein Index. *Clin. Chem.* **28,** 739–748 (1982).

3. 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, New York, NY, 1979, pp 93–106.

4. 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, New York, NY, 1980, pp 329–339.

5. 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*, R. Allen and P. Arnaud, Eds., W. DeGruyter, New York–Berlin, 1981, pp 383–400.

6. Willard, K. E., and Anderson, N. G., Two-dimensional analysis of human lymphocyte proteins: I. An assay for lymphocyte effectors. *Clin. Chem.* **27,** 1327–1334 (1981).

7. Anderson, N. L., Giometti, C. S., Gemmell, M. A., et al., A two-dimensional electrophoretic analysis of the heat-shock-induced proteins of human cells. *Clin. Chem.* **28,** 1084–1092 (1982).