Quantitative Analysis and Pattern Recognition of Two-Dimensional Electrophoresis Gels

Gregg Ridder, Ed VonBargen, David Burgard, Harvey Pickrum, and Emma Williams

We describe a system, both hardware and software, that provides quantitative analysis and data reduction of two-dimensional electrophoresis gels. Image-analysis techniques are used to determine spot intensities and to match spot patterns among many gels. A pattern-recognition program is used to extract the useful information contained in the spot lists. The application of this technology to a study of supernates from bacterial cultures is described.

Additional Keyphrases: image-analysis techniques • computerized data handling • characterizing bacteria from information on culture supernates • strain differences

We have used two-dimensional electrophoresis (2D) to characterize strains of Staphylococcus aureus by profiling the proteins in their culture supernates. In evaluating the 2D silver-stained gels, it quickly became apparent that visual inspection of the gels was totally inadequate. Visual inspection is only semiquantitative, making subtle quantitative changes impossible to detect by eye. Also, "information overload" is quickly experienced when gels containing more than 1000 spots are to be analyzed and more than two gels must be compared. Because there are so many variations in protein patterns among the strains, samples from many strains must be analyzed if one is to determine which proteins are unique to those strains capable of causing a specific disease.

Our objective was to program a moderately priced ($100,000) commercially available image-processing unit to semi-automatically transform the protein patterns from stained gels into a quantitative list of matched spots. The image-processing system would have much greater capability than a simple scanning device, but not the expense of a large system with the many personnel required to program and maintain it. We also wanted to implement a pattern-recognition program to extract only the useful information from all the spot lists. The entire system we implemented allows us to get useful results from a relatively small number of gels (<200 per year) at a minimum cost and effort. The characteristics of the hardware and software are described and the application of the entire system to a toxicology study is discussed.

Materials and Methods

Culture Supernates

Supernatant fluids from cultures were prepared by incubating the various strains of Staphylococcus aureus overnight in a dialysate from heart-infusion broth, with shaking, in air. By dialyzing the broth with use of a 3.5-kDa cutoff membrane, no proteins originally present in the medium would be detected on the 2D gels. After incubation, the bacteria were removed by centrifugation and the culture-supernate proteins were removed by precipitation with ammonium sulfate, redisolved, and then dialyzed against polyethylene glycol/water (20/80 by vol) overnight. The resulting protein concentrates were passed through a G-25 Sephadex column, and lyophilized.

Electrophoresis

The proteins in the culture supernate were reconstituted in distilled water and mixed with a solubilization buffer containing, per liter, 20 g of sodium dodecyl sulfate (SDS), 10 g of dithiothreitol, and 100 g of glycerol, and buffered at pH 9.5 with 50 mmol of 2-(N-cyclohexylamino)ethane sulfonic acid buffer per liter. The supernates were then subjected to the isodalt two-dimensional electrophoresis technique (1). A wide-range (pH 3.5 to 9.0) mix of LKB ampholytes was used for the isoelectric focusing and a 100 to 200 g/L linear polyacrylamide gradient was used for the SDS electrophoresis. The gels were silver stained by the procedure of Guevara et al. (2). Figure 1 shows an example of the resulting gels. Automated analysis of these gels must accurately identify and measure each spot on the gel and make statistical comparisons of all the proteins contained on many of these gels.

Image Analysis

The stained gels were analyzed using a vicom (San Jose, CA 95134) image-processing system for which we had written programs to develop matched spot lists. The hardware has been described elsewhere (3), and our configuration is shown in Figure 2. All of the programs to analyze the gels were written in pascal and are described separately below. The programs are menu driven and allow for operator interaction through the keyboard for program control.
and through the data tablet for graphic information input. The entire process consists of eight steps which are described by the following programs:

**Data acquisition:** This program digitizes the image of the stained gels into a $512 \times 512$ (16 bits deep) picture element (pixel) matrix in computer memory. A Model 7120 vidicon television camera (COHU, San Diego, CA 92122) is used. The system contains enough memory to hold four gel images simultaneously. The operator can load a reference image from the disk into one memory plane and flicker between this reference image and the live television image of the gel to be digitized in order to position the gel so that it is in close registration with the reference.

Sixteen scans of the gel are averaged to increase the signal/noise ratio of the image. This transmission image of the gel is converted into an absorbance image by subtracting the logarithmic transformation of the gel image from a logarithmic transformation of an image of the light box alone. This procedure is used to compensate for the uneven illumination of the light box and the irregular response across the face of the television camera, and to make the absorbance value of each pixel proportional to the concentration of the protein in the gel at that position. Once the absorbance image has been determined, the image is smoothed by convolution with a $7 \times 7$ spatial filter with a gaussian-shaped bandpass curve \((4)\). The final image is stored on disk for immediate access by the programs and on magnetic tape for archival purposes.

**Background subtraction:** The streaks and uneven background associated with the electrophoresis and staining processes are removed by a program similar to the approach described by Anderson et al. \((5)\). In this routine the minimum element in a $51 \times 1$ vertical or horizontal line segment surrounding each pixel is assigned to the corresponding background image pixel. These images are then expanded by assigning the maximum element in the respective $51 \times 1$ line segments to the expanded background image. A third background image is created by assigning to each pixel the maximum value found in the corresponding pixels of the expanded vertical or horizontal line segment images. This third image is then smoothed by a $3 \times 3$ low-pass filter and subtracted from the gel image.

**Boundary determination:** Spots are differentiated from background areas or residual streaks, and overlapping spots are separated from each other by convolution of the gel image with $3 \times 3$ vertical and horizontal line filters (Figure 3). These line filters produce directional second-derivative images similar to the process used by Lemkin and Lipkin \((6)\) to produce the "central cores." The complete process is to \((a)\) perform two successive logarithmic transforms on the gel image, to accentuate the gradients of the spot edges; \((b)\) calculate the vertical line filter image and scale it to the full intensity range; \((c)\) create the binary image of the line filter image by selecting a threshold value such that single pixels attributable to noise begin to show; \((d)\) repeat the first two steps, to create the binary image of the horizontal line filter image; \((e)\) perform a boolean "and" operation with the two binary line filter images to create a binary mask; and \((f)\) add to the mask a binary image of the original gel image by using a very high threshold to detect the flat tops of those peaks whose intensities are beyond the dynamic range of the camera. The resulting mask undergoes one dilation to expand the boundaries slightly and is used to define boundaries of all the spots on the gel.

**Image editing:** This program allows the operator to inspect and, if necessary, alter the binary mask produced by the boundary program. The operator may zoom and scroll the image to get a closer look at any part of the gel, and either erase or draw mask pixels where they are required.

**Spot integration:** Once the boundaries have been determined this program will integrate the densities of all the pixels within a boundary, calculate the \(x,y\) position of the centroid for each spot on the gel, and create a complete spot list containing this information on the disk.

**Spot matching:** A reference image and spot list is created by copying the image and the associated spot list of a gel judged to be the most representative of the group of gels to be matched. All of the gels will be matched to this reference, and spots not initially present on the reference will be added manually. Thus the reference will in the end contain all of the spots on all of the gels. Ultimately, new matched spot lists will be created from the old spot lists and will contain only assigned spot numbers and spot intensities.

The algorithm for determining matches is similar to that of Garrels \((7)\). A neighborhood size and acceptable error limits of matching are defined by the operator such that the computer makes matches over the largest possible areas without making immediately obvious mistakes. The process is started by the operator establishing a series of landmark matches, using the data-tablet input and allowing the computer to find as many matches as possible. All of the matches are displayed by tie-lines in the graphics overlays of the reference and test images. When the computer can find no more matches, the process is restarted with new landmark matches made by the operator. This iterative approach is continued until all possible matches are made. The software allows the operator to zoom in on any part of the reference and test images, adjust the registration of the two images, and flicker between them in order to inspect or edit any matches made. The amount of operator effort is dependent on the similarity of the protein content of the samples and the reproducibility from gel to gel.

**Spot comparison:** When all the spot lists are matched, this program allows the operator to compare the intensities of the matched spots of any two gels and to see which spots are still unmatched. The program also performs a statistical evaluation of the slope, intercept, correlation coefficient, standard deviation of the data about the regression line, and
average mean difference (5) as an estimate of overall similarity of the two gels. Any data point on the plot can be identified on the original gels by use of the cursor, resulting in the display of the spot list data and graphic location of the spot on the gel images.

Pattern recognition: Determining correlations of the data in all the spot lists with any known data on biological activity is done by using a software package, ARTHUR 81 (Infometrics, Seattle, WA 98125). This package consists of many FORTRAN routines and is run on the VAX 11/730 computer. It includes programs of the principal-component analysis and cluster analysis type to help determine any inherent structure within the spot lists (5).

We transmit our matched spot list data over telephone lines to the VAX 11/730 computer. The format of the data is in conformity with the program-input routines and they then are run through a series of analyses selected by us, to extract the relevant information. For example, if the samples can be assigned to particular categories (e.g., culture supernates from control or disease-associated strains of bacteria), a typical series of program steps consists of (a) auto-scaling the data, so that spots with the larger intensities are not weighted more in the analyses; (b) determining which spots are best markers for the sample categories on the basis of their calculated Fisher weights; (c) selecting the fewest spot numbers necessary to discriminate accurately among all sample categories; and (d) running a k-nearest neighbor (KNN) analysis or principal component with the selected spot numbers, to determine the accuracy of the markers in discriminating sample categories.

Results

Data Acquisition

The accuracy with which density measurement of dark spots on a light background can be made with use of this system can be shown by measuring an image of a calibrated photographic-density-step tablet. In Figure 4 the measured density values are plotted in arbitrary integrated-density units against the true values and show a nonlinear response, which is useful only to about 2.0 A. The nonlinearity is undoubtedly due to the problems of unavoidable stray light and "blooming" associated with the use of a vidicon camera.

The enhancement of the signal/noise ratio due to averaging and the use of the gaussian spatial filter is shown in Figure 5. Without this enhancement the boundary determination routines do not perform well, because they are based on detecting small differences in the intensities of adjacent pixels.

Background Subtraction

Our silver-stained gels typically have large, uneven background densities and streaks resulting from various electrophoretic anomalies. The results of the background subtraction process as used on a representative gel are shown in Figure 6. By visual inspection the process appears to work well in most cases. Occasionally problems do occur where over-ranged spots are associated with long, heavy streaks, resulting in density being removed from the spot.

Boundary Determination

To test the accuracy and reproducibility of the boundary detection process, we created a synthetic image to include a series of gaussian shaped spots representing variations in protein concentration, a series representing variations in spot width, and a series representing variations in the amount of overlap of two adjacent spots. The largest spot size consisted of approximately 50 pixels, or roughly the size of a spot 2 mm in diameter when the entire area of a gel 17 cm square is being analyzed. Random noise was introduced onto the image to create a more realistic situation. Figure 7 shows a plot of the integrated densities of the spots representing variations in protein content as measured by the image-analysis system vs the corresponding true values. A linear response and a slight negative intercept results. This error in accuracy contains a proportional component of about 9% and a small constant component equivalent to 3% of the 2 mm spot.

The width of the spot appears to have a significant effect on the boundary determination, in that the widest spot is underestimated by as much as 20% more than the narrowest spot.

The boundaries of overlapped spots were efficiently split in every case where two peaks were discernible.

We estimated a 1% reproducibility in the analysis of all the spots on the synthetic image by analyzing three images with different sets of random noise introduced.

Comparisons of Two Gels

Because we are interested primarily in relative rather than absolute differences among gels, the most useful test of the imaging system is to compare the reproducibility with which the same gel can be analyzed twice. To determine the performance of this system, we have calculated the average mean difference, described by Anderson et al. (5) as a measure of similarity between the results of two analyses. Approximately 8 µg of bacterial culture-supernate proteins were separated by 2D and silver stained. This gel was analyzed twice and the intensities of the matched spots were plotted against each other (Figure 8a). An average mean difference of 8.9 was calculated, which represents the limit in reproducibility of the measurements of these silver-stained gels owing to variations in scanning, background subtraction, and boundary determinations.

To determine if the reproducibility of the analysis would allow us to measure the variations in producing culture supernates, we analyzed two samples from duplicate culture supernates, using the same strain of Staphylococcus aureus. The plot (Figure 8b) shows greater variation and a larger average mean difference, 16.9, in the replicate cultures than was seen in the replicate analyses of the one gel. A compari-
son of samples of culture supernate of two different strains of *Staphylococcus aureus* (Figure 8c) results in an even larger average mean difference, 33.3. The analyses of 16 different strains of *Staphylococcus aureus* showed degrees of variation among the strains similar to those illustrated in Figure 8c. These results show that the reproducibility of the image analysis is sufficient that it can be used to determine significant differences in replicate cultures. Also, the reproducibility of the culturing and sample-preparation processes is sufficient to determine significant differences among the various strains of bacteria.

Comparison of Many Gels

One of the goals of the project was to identify all of the proteins in the culture supernates that qualitatively or quantitatively discriminate strains associated with a disease from strains isolated from healthy individuals. To identify these proteins, we assigned the gels to either control (no. 1) or disease-associated (no. 2) categories. The classified matched-spot lists were all used to create a data base for the *Arthurs81* program.

The density values at each spot number were auto-scaled to give a mean value of zero and a standard deviation of 1.0 across all 16 gels. The program then calculated the Fisher weights for each protein as an indicator of their ability to discriminate the two categories of strains. Five proteins out of 720 evaluated were identified with Fisher weights above 1.5. We have chosen a Fisher weight of 1.5 as the lowest acceptable level of discrimination. The ability to discriminate the two categories of strains can be seen by plotting the
intensity values of the two most-correlated proteins, spot numbers 316 and 320, against each other for each gel (Figure 9). It can also be seen that these two proteins are strongly correlated with each other, because the amount of 320 increases with the amount of 316. On the gel these two proteins do appear to be related by virtue of their closeness in pl and molecular mass. Other proteins of the five identified by the program are also highly correlated with 316. Two of the five others have negative correlation coefficients, with no apparent relationship by virtue of position on the gel.

To find any subsets of the disease-associated strains based on any inherent differences of the protein patterns alone, the program selects the protein most highly correlated with the two categories, decorrelates the other proteins from this chosen protein, reweights the data, and then selects the protein whose new Fisher weight is highest. The result of this analysis identified no new proteins with Fisher weights >1.5. With use of only the data from the two most highly correlated spots, both KNN cluster analysis and principal-component analysis were able to classify all of the gels correctly.

Because two of the control strains were able to produce small amounts of protein no. 316 we reclassified the strains as non-producers of no. 316 (category no. 3) or producers of no. 316 (category no. 4) based on the 2D gel analyses. Of the 10 proteins with Fisher weights >1.5 for these categories, spot number 316 was identified, as expected, as the most highly correlated protein, but a new subclass of 316 producers was now identified by a second protein. As shown in Figure 10, by using the combination of 316 and the second protein (spot number 157), we identified three categories of samples that could not be clearly discriminated by the analysis of any one protein alone. The information contained in spots number 316 and 157 was sufficient for correct classification of all the gels by KNN and principal-component analyses. The biological significance of this second protein and those highly correlated with it is being determined in animal experiments.

Discussion

We have developed and applied a computer system for the semi-automated analysis of the large amount of information in complex 2D electrophoresis gels. This system allows 2D electrophoresis to be effectively used as a comprehensive screening tool in biological investigations. The use of pattern-recognition techniques was required to reduce the information in the matched-spot lists.

The advantages and disadvantages of several features of our computer system need to be considered. The first feature is the use of a television camera for data acquisition. The advantages are that wet gels can be analyzed easily, the camera is relatively inexpensive, and scanning is very fast. The disadvantages are that the sensitivity of the response is usable to only 2.0 A, owing to stray light and blooming, and the resolution of 512 × 512 pixels is just adequate to scan an entire gel. The vicom system has the benefits of being a small, stand-alone system that is very fast at the convolution of spatial filters through parallel processing. Also, using 16-bit deep pixels allows all mathematical procedures to work directly in the image memory, making the algorithms simple and fast. Unfortunately, the image-analysis software available for it is minimal and had to be written by us. The pattern-recognition software, ARTHUR81, is comparatively inexpensive and contains a very wide range of statistical options; however, it must be run on a separate computer.

Fig. 8. Plot of the logarithms of the matched spot intensities from (a) duplicate analyses of the same 2D gel of bacterial culture supemates, (b) duplicate cultures of the same strain of bacteria, and (c) culture supemates from two different strains of bacteria

Fig. 9. The separation of control (1) and disease-associated (2) strains of bacteria determined by the amounts of spots no. 316 and no. 320 contained in their culture supemates

Only two control strains (labeled *) contain detectable amounts of spot no. 316

Fig. 10. Identification of subpopulations of strains that produce spot no. 316 (labeled 4) based on the amount of spot no. 157 present in the culture supemates

CLINICAL CHEMISTRY, Vol. 30, No. 12, 1984 1923
We have found the system to be very useful for analyzing as many as 40 gels at a time. This is a reasonable number, since batch-to-batch variations in gels currently would probably be the ultimate limiting factor. The speed with which we can analyze this number of gels depends on how similar the samples are in protein content and how reproducible the spot patterns are. Our experience with the various bacterial culture supernates, whose protein contents are very different in most cases, is that a week is required to completely analyze 16 gels because of the operator intervention required to add new spots to the reference spot list and check the matches for accuracy. About three days is still required for analysis of 20 more similar gels, simply because of the number of steps required. Further improvements in the software may shorten this time significantly, but the greatest need for improvement is to increase the gel-to-gel reproducibility, especially in the isoelectric focusing dimension.

The system has allowed us to extract relevant information from very complex spot patterns in a reasonable amount of time. The information we obtained on the culture-supernate gels was impossible to obtain by visual inspection. Similar results have been obtained in other experiments in our laboratory to determine changes in selected tissues of animals treated with known toxins or carcinogens. For effective use of 2D electrophoresis to study biological processes, image analysis and pattern recognition are essential tools.

References