Geisow & Barrett (eds.) Computing in biological science

6

Database techniques for two-dimensional electrophoretic gel analysis

PETER F. LEMKIN and LEWIS E. LIPKIN

1. INTRODUCTION

This article discusses GELLAB [1–6], a system for computer aided analyses of two-dimensional electrophoretic patterns. The 2D polyacrylamide gel electrophoresis (PAGE) technique [7] has been a rapidly developing biochemical tool, applicable to a wide variety of problems in molecular biology, basic biochemistry, genetics and clinical research [8]. The 2D PAGE technique can be used to separate hundreds to several thousand polypeptide components as a matrix of spots because the variables which determine electrophoretic mobility in each of the two dimensions are effectively orthogonal to each other. Isoelectric focusing over a pH gradient determines extent of movement in the first dimension. In the second dimension the sodium dodecyl sulfate (SDS) interaction with polypeptides results in a mobility which is a function of molecular weight.

The greatly-increased number of 'spots' detectable in 2D PAGE [7] is a major reason for efforts at computerized analysis [4,8–12]. As a result of these automation attempts, there has been an increase in complexity of intermediate analysis results that it would strain the limits of unaided human analytical ability. The need for analytic assistance is further increased by the added complication of non-linear spatial warping of corresponding moieties in comparable gels.

A 2D PAGE electrophoretic gel is a complex of distinct polypeptides, each one of which is characterized by position relative to other polypeptides ('spots') and density. However, unlike a geographic map, proximity of polypeptides on a gel is no particular indication of related genesis or biological function. Nonetheless the large number of discrete spots in a gel and the similarity that is preserved among gels from

a similar source allows one to track many proteins through the effects of experimental variables. Hence, comparing biological specimens by comparing their corresponding gels for quantitative or qualitative differences has become an important means of determining protein manifested metabolic differences.

We are progressively more concerned with the generation of data structures, strategies and tactics for their employment in the analyses of *sets* of gels. Such comparisons, both qualitative and quantitative, among multiple gels might reflect, for example, successive values of a dose or time variable in an experiment or the clinical course of a patient.

In dealing with this material, human factor considerations place a practical limit on the number of spots for which manual density information can be obtained. Manual techniques such as optical flicker or dual-color comparison between two local regions on separate gels are useful for local alignment, especially in cases with obvious spot differences [9]. However, this method forces the user to deal with the gels sequentially in that only pairwise gel comparison can be made. This makes the process time consuming and difficult for the observer to visualize a pattern directly over a set of gels. Such manual comparison methods are probably capable of supporting a complete search for all major polypeptide differences, but the bookkeeping needed to identify the same spot in several gels makes computer aid attractive. Beyond some relatively small number of spots, some computer aid in matching, 'remembering' and retrieving images of preserved spot correspondences is seen as indispensable in thoroughly analogous sets of gels. An added benefit of this is that after the spots have been isolated, located and tagged, the machine can use this information to produce a variety of representations. These include pictorial, diagrammatic, numerical, etc., that aid the user to see patterns difficult to grasp when attention is focused on small regions. Final output of GELLAB includes labeled gel image maps and mosaics where statistically interesting spots have been marked as well as numeric spot data lists to support these findings.

Figure 1 illustrates the major steps performed in the GELLAB 2D gel analysis procedure. Gels are first accessioned (assigned a unique number) and related experiment information is entered at this time into the system. The gel images are then acquired (digitized and stored). Then spots are segmented in each gel followed by spots being paired between each gel and a standard gel using a small set of manually defined landmarks. Finally a multiple spot data base (DB) is constructed and analyzed. Final output of such a system takes several forms. This includes labeled gel image maps (superimposed on the original gel images) where statistically interesting spots have been marked as well as numeric spot data to support these findings. Various 1-, 2-, and 3D functions of spot data base features may be plotted both on interactive displays and later, on paper.

1.1. Gel characteristics

For the most part, a spot's resultant geometric position in a gel bears no relation to function or the origin of the protein it represents. Closely related polypeptides may be separated by considerable distances while functionally unrelated materials could

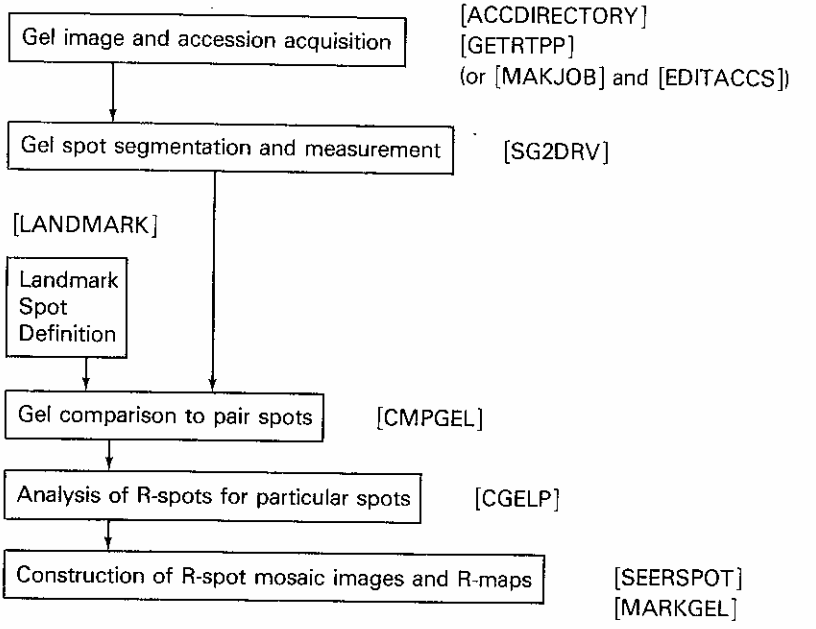


Fig. 1. Block diagram of the 2D-gel analysis GELLAB system. Programs associated with major steps of GELLAB are indicated in '[...]'. Gel images are acquired by scanning with a vidicon TV camera interfaced to a picture memory and saved on computer disk files. Accession information about the experiments which produced the gels is also used to update an accession file. The gel images are then segmented and measurements made of the detected spots. Landmark spots common to the *R*-gel and all other gels are then interactively selected. The landmark spots are aligned for all of the gels with a representative gel (*R*-gel) using gel image flicker alignment. This landmark spot information and the spot segmentation data is then used to pair congener spots in the remaining gels with the *R*-gel. The set of gel pairings with the same *R*-gel may be merged together to form a list of sets of equivalent *R*-spots called the congener gel data base (CGL) and subjected to further analysis.

be distributed in close proximity. In contrast to more conventional images such as microscopic fields or X-rays, the image 'structure' (i.e., local adjacencies, inclusions, etc.) in gels provides little information to facilitate the analysis. Individual spots unless contaminated by artifacts or overlapped by other spots are much simpler images than, say, the image of a cell in a blood smear. Thus once a spot has been isolated, its analysis and characterization as an *individual entity* at least in a single gel, is relatively simple.

Non-congruence is a more serious difficulty (i.e., the lack of point to point reproducibility of gels). It occurs in gels derived from the same sample and from a single run on the same apparatus. This is due to a large number of preparative factors including local temperature variations, local heterogeneities in polyacrylamide texture and/or local concentration, heterogeneities of ampholine concentration, etc. All of these variables and perhaps others less understood, combine to reduce the reproducibility of mobility of polypeptide fragments in one or another dimension. This net result is a set of gels which are *not congruent* but which are related by an affine transformation. In other words, comparable spots within a set of gels have

corresponding neighbors but are not necessarily located at exactly the same distances from these neighbors in any specific instance. The set of gels show a local superimposeability, which is maintained for surrounds of varying extent. Thus it is this absence of simple direct correspondence coupled with the large numbers of spots in a set of gels that makes some automated assistance a necessity.

1.2. Classes of problems

It is necessary when viewing to carefully consider the kinds of questions which biological problems pose which will determine the nature, depth and range of the analyses to be performed.

(1) Is only one or a very few different spots present in one gel and not in its experimental pair? The paradigmatic biologic systems which pose such questions are those in areas such as bacterial genetics, where both the homogeneity of and the specificity of the product generating cell line are very high. Here the gels are used as detectors and serve simply to confirm or deny the existence of a polypeptide. Simple flicker analysis may be all that is required under favorable conditions while densitometery for this situation is a secondary consideration if one at all. A case in point is a single gene difference in an *E. Coli* mutant [9].

(2) Are there changes in any of several spots in a cell line as a function of time? These variations are often in polypeptide quantity so that densitometry is required. Moreover, the complexity of the analysis is increased so that a number of gels rather than a pair of gels must often be compared. The answer to such questions as this requires spot data structures and data base management software that are both significantly larger and significantly more complex than those required for the answer to the first question.

(3) Are there changes in several spots resulting from an applied stimulus? The less known about the outcome, i.e., the more exploratory the search, the more extensive and complex must be the gel analysis. When the cell line is only apparently homogeneous (as was the case for PHA stimulated lymphocytes [4]) and where the effect of the stimulus is both complex and a function of time, the gel analyses become correspondingly more extensive and complex. In such situations, where many new spots may result, there seems no alternative to automatic spot pairing using a computer.

(4) Is there a 'finger print' of morphologically homogenous but biologically and functionally different cell groups as for example differences among various lymphoblastic tumors? Here, especially if stimuli are required to elicit differences, the number of gels grows to an m (number of classes) times n (number of temporal samples) times p (number of levels of stimuli) number of comparisons. This assumes minimum problems in the reproducibility of the gels whereas often multiple gels of the same sample are run. Particular interest must be focused not only on differences but on subgroup similarities as might seem to be indicated by the clustering of density ratios in the histogram tables (cf., Table 10).

(5) Are known polypeptides present in normal or abnormal quantities in a body fluid? Here are essential problems of clinical chemistry, but multiplied by the large

number of spots present in the gel. At first, the answer to questions of this type might seem simpler than those above. The comparison of a single gel's contents to some internal or external standard cetainly involves future developments in the area which has been called by Anderson 'molecular anatomy' [8]. Because of the need for extensive bookkeeping in multiple gel analysis, it is likely that some of the data structures we present here will be an aid in this development.

Gels may be thought of as complex objects similar to a geographic map with individual polypeptides appearing in distinct morphologic regions. They are not however in any way certain reference points where unlike the geographic map, adjacency of morphologic polypeptides in the gel is no particular indication of related genesis or biological function. However, characteristic patterns are obtained with carbamylation and other biochemical treatments. Comparing biological specimens by comparing their corresponding gel maps is one means of determining major protein differences although this is tedious when done manually. Given a number of gels, polypeptide concentration values may be modeled as a density distribution for each set of corresponding spots and the analysis performed on the distributions.

1.3. Hardware and software implementation

This has been described in [9,13–17]. Briefly the system consists of a DECSYSTEM-2020 computer controlling a one-of-a-kind Real Time Picture Processor (RTPP) constructed in our laboratory [14–17]. The RTPP uses a PDP8e as a display processor controller for the sixteen 8-bit gray values (256 gray values 0 = white, 255 = black) 256 × 256 picture element (pixel) frame buffer image memories. A Quantimet 720 image analyzer is used for video TV camera input and TV display output. These memories may be configured as four 512 × 512 pixel images for use with the GELLAB programs. The RTPP has a TV camera and image (A/D) acquisition hardware which enables a TV frame to be acquired in 1/10 second into the frame buffer memories. The time shared 2020 (under TOPS10 monitor system) processor has 512 K words of 36-bit memory and three 160 Mbyte disk pack drives as well as two magtape drives. The RTPP's picture memory, as well as the PDP8e control teletype, has been interfaced to the 2020 via a UNIBUS – thus implementing a distributed picture processor. The GELLAB system is one of several image processing projects using the above hardware.

In addition, gel images scanned elsewhere on an Optronics scanner have been used on our system with a RT-11 magtape intermediary. The procedure is to scan the gel onto a RT-11 PDP11 disk file and then later to transfer one or more of these files to magtape. A GELLAB SAIL program called MT-11 is able to read images from these magtapes onto the DECSYSTEM-10 (or -20) file system.

The set of GELLAB programs SG2DRV, CMPGEL, CGELP, MARKGEL, SEERSPOT, DWRMAP, LMSEDIT, etc., are written in the SAIL programming language [18]. This language is currently implemented *only* on DECSYSTEM-10 and DECSYSTEM-20 computers. It has distinct advantages in its ease of algorithm expression, macro expansion, string, list, set and associative processing and record

structure operations. Since SAIL strongly encourages structured programming it is an ideal environment in which to implement a set of complex interacting algorithms.

A subset of the GELLAB system has been constructed to run on any DECSYSTEM-10)or -20) *independent* of our special purpose hardware. Of course, interface software must be written to read and write image files to the local image display or plotting hardware systems. This export version of GELLAB will be discussed in more detail in the Appendix.

1.4. Image acquisition

Data acquisition is accomplished by scanning backlighted gels or gel autoradiographs with the Vidicon camera [9]. The Vidicon camera has a Nikon-N auto 1 : 2 28 mm objective lens routinely set to *f*8 with the autoradiograph film mounted 69 to 42 cm away and backlighted on an Aristo type T-12 uniformly illuminated light box (Port Washington, NY). The effective resolution of the image ranges from 250 to 170 microns/pixel (picture element) although other lenses can and have been used. A type 1009 NBS Neutral Density (ND) wedge is mounted at the bottom of the illuminated area. For use with 120 mm film size negatives (for use with the silver stain), a 55 mm micro-Nikkor lens is fixed at *f*8 and set for a distance of 55.5 cm.

The scanned images are acquired using the GETRTPP program of GELLAB and saved directly on the DECSYSTEM-2020's disk. At this step of acquisition, gels are assigned accession numbers and experiment information about the gels is entered. This information is used in tracking the gels throughout the analysis. An image consists of an 512 × 512 pixel array. The upper left hand corner of the image is defined to be (0,0) while (511,511) represents the lower right hand corner. The actual dynamic range of the gray scale data is slightly greater than 7-bits, however of this probably only 6-bits of gray scale resolution is actually valid. The maximum dynamic range of a vidicon is about 0 to 2.0 OD. Gels can be checked against the ND wedge using analog video detector circuits of the TV system to determine whether any spots are greater than the TV's dynamic range. Since most spot information is less than 2.0 OD there is generally no problem of gray scale distortion for photographically non-saturating spots. Fifty images are easily scanned in about one hour or less.

Although the transfer function of a Vidicon TV camera is non-linear, it may still be used to perform densitometry under some conditions. These include: (1) the majority of the material to be measured is below the saturating end of the camera's dynamic range; (2) the non-linear gray scale to OD transfer function be well behaved over the range of data to be measured; and (3) all calculations involving integrated OD/spot are done in the density domain. Even then, appreciable errors on very dark spots can occur due to saturation, noise and to optical problems (e.g., glare and under-representation of dark high OD pixels).

1.5. Calibration

Using the GETRTPP or EDITACC programs on the RTPP (see Fig. 1) the user

interactively defines the computing window (the region in the gel image where the spots are located) taking care to omit the ND wedge and gross artifacts in the gel. Using these programs, the ND wedge (in the gel image) is calibrated by computing a gray scale histogram of a 25 pixel wide window positioned by the user across the wedge. Peaks in the smoothed histogram are then matched automatically with the actual OD values of the wedge. The wedge gray value peak information and a gel computing window position (interactively defined) are automatically updated into the gel accession file along with the gel experiment information. A piecewise linear function can be generated for the ND wedge as a function of gray value. This permits image gray values to be mapped to density (OD). Table 1 shows part of a typical accession file. Figure 2 illustrates a gel with the ND wedge 2a, its manually defined computing window 2b and ND wedge sample window 2c, and the resultant ND wedge histogram calibration curve 2d. The piecewise linear OD calibration function is drawn (black) over the histogram.

1.6. Gel segmentation: nature of the image

Gel image accessioning and acquisition is only the first stage in making spot information available for automated processing. Provision must be made for separating out 'pictorial information of interest' (in this case the spots) from 'noise' and 'background' before spot positions and spot properties can be compared. Consequently, a spot extraction algorithm must be capable, under a wide variety of actual gel image conditions, of (1) detecting, (2) defining the extent of, and (3) measuring the density of a spot.

The segmentation problem is one of the more important and ubiquitous, in the general field of image processing. Almost all real images resist simple gray scale thresholding as a solution to pictorial partitioning or segmentation and despite the simplicity of spot morphology, 2D gels are no exception. The thresholding operation applied to an image retains all values higher than a certain gray value while all others are set to white.

1.7. Spot morphology

The vagueness of spot morphology and inhomogeneity of gel background complicate these images. Spots often touch each other resulting in overlapping spots. The 'tails' of spots may extend for a considerable distance into an overlapping region. Spots have no distinct boundary, but occur most often as an effectively continuous Gaussian-like distribution which Lutin asserts that this distribution tends to be symmetric with respect to isoelectric point (pI) but is skewed in molecular weight (MW) [10]. In practice, this holds only for ideal non-conglomerate spots. In addition, spots may appear round, oblong, or take on various continuous shapes particularly when there is excessive loading of material in the gel. In all cases, except in the case of extreme overloading, however, the center of a spot is its darkest part. Spots may sometimes be obscured in certain regions of the gel which are susceptible to streaking in both MW and isoelectric axes. Spots can occur within these streaks which are of interest.

TABLE 1

Example of part of a gel accession descriptor file

```
ACCESS. #/PATIENT/BIRTHDATE/RACE&SEX/EXP DATE/EXP #/CULTURE REAG/AMPH,GEL/
INTRVL BEFR LBLNG/LBLNG ISOTOPE/DURTN LABEL/DURTN OF EXPSR/STUDY/
FILE #/TAPE #/OPT. BACKUP TAPE #/ CAMERA,LENS,DISTANCE/EXPRMNTR*
ND:.05,.20,.35,.50,.66,.80,.95,1.10,1.25,1.41,1.56,1.72,1.87,2.02,2.17
ASBSPT.DA = ASBESTOS E-SPOT FILE
        .
        .
        .
0250.1/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 DAY/ALUMINUM,T0,CONTROL,BOTTLE#1/
L00153/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 41 67 89 110 127 144 159 171 182 189 195 205 0 0 0 0 53 425 78 425
0250.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T0,CONTROL,BOTTLE#1/
L00157/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 65 88 109 126 144 159 170 181 190 196 208 213 0 0 0 31 460 51 400
0251.1/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 DAY/ALUMINUM,T0,CONTROL,BOTTLE#2/
L00161/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 29 56 67 79 101 119 137 153 166 177 186 193 200 208 0 0 2 478 49 410
0251.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T0,CONTROL,BOTTLE#2/
L00165/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 67 88 110 126 144 159 171 182 191 197 208 0 0 0 0 2 446 45 390
        .
        .
        .
0258.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T24,CONTROL,BOTTLE#9/
L00221/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 65 88 110 127 144 159 171 182 191 197 208 213 0 0 0 2 491 53 405
        .
        .
        .
0264.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T24,AMOSITE,TOXIC,PHAGOCYTIC,BOTTLE#15/
L00269/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 41 67 89 110 127 144 159 171 182 190 197 208 213 0 0 0 2 459 72 405
0265.1/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 DAY/ALUMINUM,T24,AMOSITE,TOXIC,PHAGOCYTIC,BOTTLE#16/
L00273/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 62 84 105 123 140 156 168 178 187 193 208 0 0 0 0 0 2 446 40 440
0265.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T24,AMOSITE,TOXIC,PHAGOCYTIC,BOTTLE#16/
L00277/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 37 63 86 108 125 143 158 169 180 189 194 201 210 0 0 0 6 454 23 360
        .
        .
        .
```

Example of part of a typical gel accession descriptor file for the P388D1 data base. Each data record contains four lines. The first four lines of the file define the record field descriptors and ND wedge values. The descriptors are separated by '/' and terminated with a '*'. The fourth line of each record is the set of gray value peaks corresponding to the ND wedge calibration. The last four value of that line are the computing window for that gel [$x1$:$x2$,$y1$:$y2$].

Polypeptides in the gel are not visible by themselves and must therefore be visualized in order to perform an analysis. At least four methods are currently used which include: Coomassie blue staining, autoradiography (on radioactively labeled proteins produced by growing the tissue culture in radiolabeled amino acids), silver staining [19–20], and fluorescent dyes. These spot detection methods have

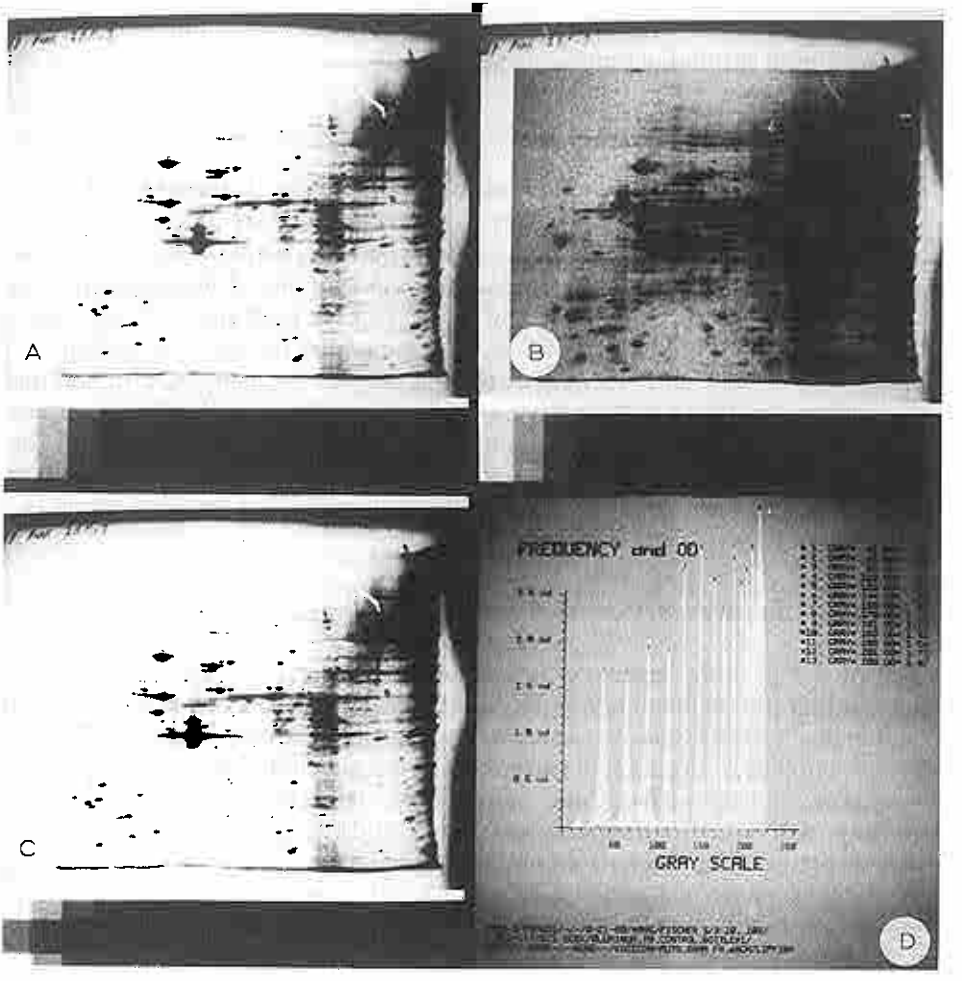


Fig. 2. Typical 2D gel with ND wedge for P388D1 macrophage like cells. (a) Original T0 gel image scanned at 250 microns/pixel; (b) image with computing window noted to define segmentation region; (c) ND wedge sample window defining wedge calibration region, and (d) smoothed ND wedge sample gray value histogram with computed piecewise linear calibration function superimposed. The ND value and gray value frequency are on the ordinate and gray value on the abscissa (0 to 255).

widely different dynamic ranges and stochiometry, as well as application for different types of biological material.

Care must be taken to ensure that autoradiographic film is used in the linear portion of the density versus log (exposure) curve otherwise saturation of some spots will occur. The dynamic range of spot detection may be covered using a series of increasingly long autographic exposures of the same gel. The Vidicon or other imaging detector is subject to similar saturation problems.

Because some spots will be recorded as saturated, it is useful to know which ones and furthermore to be able to track these spots throughout the entire analysis process. This is done in GELLAB. Spots saturating in one gel might not do so in

another so that alternative measurements could be made as for example in the case of multiply exposed autoradiographs.

1.8. A segmentation model

In any locally determined (e.g., non-thresholding) feature extraction process, some explicit or implicit model of the pictorial objects is necessary. The algorithm as presented here embodies some of the ideas of the underlying spot model. The spot extraction methods previously reported use various spot models to aid the process [8,10–12]. A first order model is the triple (x,y,d) consisting of the spot's centroid (x,y) in cartesian space and its total integrated density d (a measure of polypeptide concentration). This triple appears adequate for many types of multiple analyses where the object of the analyses is to measure the amounts of polypeptides present. Segmentation is a method of spot extraction which results in obtaining this triple as well as other features. Our spot segmentation algorithm is based on a shape and density independent model and takes into account the realities of touching and overlapping spots.

1.9. Role of the segmenter in overall gel analysis

As shown in Fig. 1, the segmenter is applied following gel image acquisition. It is important that the segmentation procedure be made as automatic as possible with minimum manual intervention because of the large number of spots on a gel.

We present here a specific spot extractor which is able to handle a wide variety of spot shapes, density and cluster morphology. However, any spot extractor generating an ordered list of spot triples (x,y,d) could be used in the first stage of the GELLAB analysis. Parameterization of the segmentation algorithm permits a wide variety of gel stains to be handled, and produces different types of output which can be put to varied uses.

2. SEGMENTATION

2.1. The segmentation algorithm

The segmentation algorithm is a sequence of procedures applied to a locally averaged image. The first of these is the digital analog of the spatial second derivative; it is used to construct an image called the central core image consisting of the centers of spots. Second derivative (cf., Fig. 3) information delimits the extent of outward propagation resulting in an algorithmic limit on individual spot extent. Initial spot candidate generation is parameter independent. The decision function which later separates noise from valid spots is adjusted by user defined parameters. Auxillary information required by the segmenter, such as picture file name, and ND wedge calibration and computing window is obtained from the accession file. The segmentation algorithm, SG2DRV, is shown in flow chart form in Fig. 4.

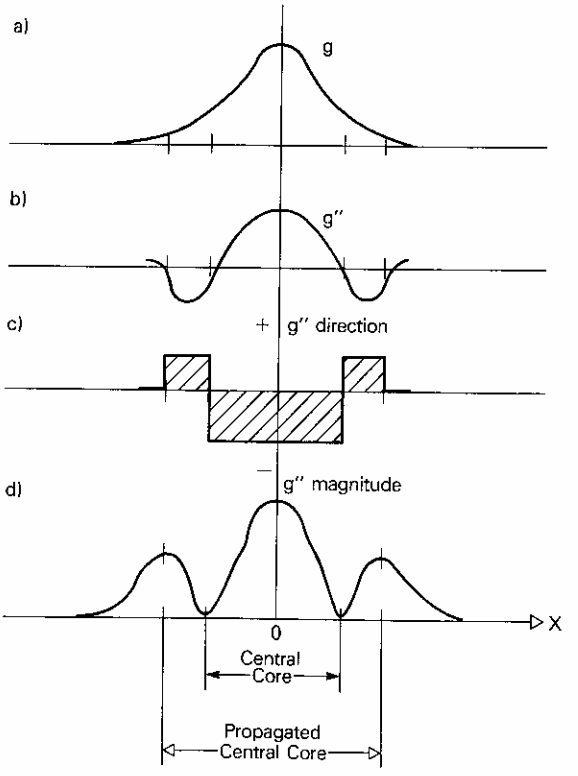


Fig. 3. 1D representation of a Gaussian like function g (a), its 2nd derivative g'' (b), its direction SIGN (g'') (c), and absolute magnitude (d) functions. In the central core region, the direction of g'' is less than 0 and changes sign in the propagated central core region. The outer extent of the propagated central core region is indicated by a second maxima in the g'' magnitude function.

2.1.1. Principle

Let g be a image gray scale point function whose mode, median and mean are all more or less central with respect to the extrema and let its second derivative be g''. The central region of a spot has a negative g'' direction and a g'' magnitude maximum. Beyond the mid-region where the direction of g'' changes sign, there is a second smaller peak in the magnitude of g''. Our segmentation procedure is based on finding these two maxima in 2-dimensions. The approximation to the boundary is operationally defined by the second maxima in the g'' magnitude function.

2.1.2. Smoothing

The original image is first smoothed to remove some of the high spatial frequency noise. This is illustrated using a 3 × 3 convolution filter [10]. Let matrix M_{ij} be defined as:

$$M_{ij} = \begin{matrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1. \end{matrix}$$

Then, for a central pixel (x,y), each smoothed pixel $f(x,y)$ is defined as:

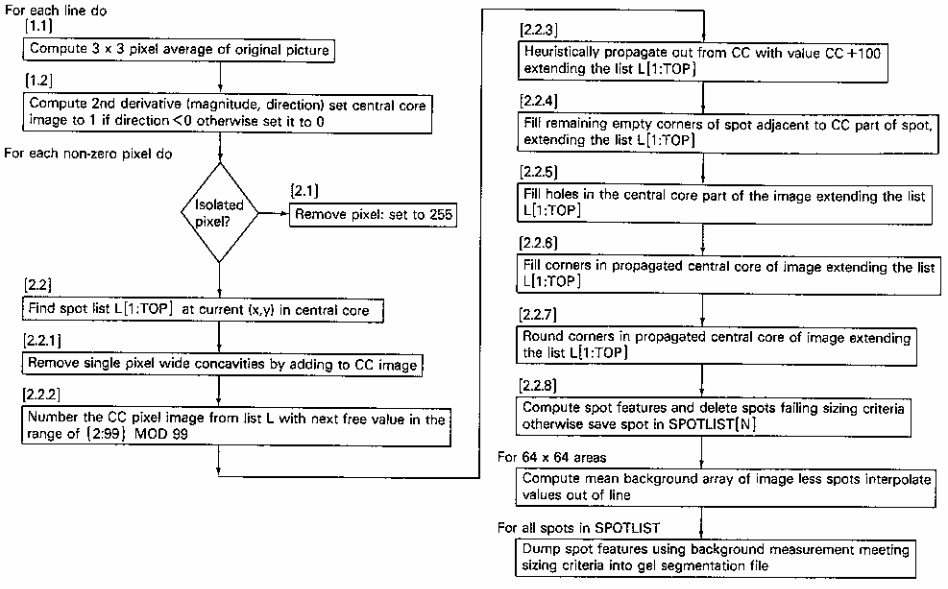


Fig. 4. Block diagram of the 2D-gel spot segmentation procedure performed during 2 passes through the image. Substeps k performed during passes 1 and 2 (through the image) are denoted by 1.k and 2.k respectively. The averaged and central core images are computed during pass 1 while pass 2 processes each spot to completion.

$$f(x,y) = (1/16) \sum_{i=1}^{3} \sum_{j=1}^{3} M_{ij}{}^{*} g(x + i - 2, y + j - 2).$$

It is applied over the entire picture, pixel by pixel in a top down left to right fashion. Each pixel in the 3 × 3 pixel neighborhood (defined by the center pixel) is multiplied by the corresponding 3 × 3 filter (pixel for pixel) and the total divided by 16. The result is saved in an *averaged* image. This filter removes enough of the high spatial frequency noise so the 2nd derivative analysis algorithm may be more successfully applied. Spot shapes are not distorted to any noticeable degree. The actual spot density measurements are made on the original image data. Larger filters, a 5 × 5, and a 7 × 7 [21], are also used.

$$\begin{matrix} 1 & 1 & 2 & 1 & 1 \\ 1 & 2 & 4 & 2 & 1 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \end{matrix} \quad \text{divided by } 52$$

$$\begin{matrix} 4 & -6 & -12 & -14 & -12 & -6 & 4 \\ -6 & 9 & 18 & 21 & 18 & 9 & -6 \\ -12 & 18 & 36 & 42 & 36 & 18 & -12 \\ -14 & 21 & 42 & 49 & 42 & 21 & -14 \\ -12 & 18 & 36 & 42 & 36 & 18 & -12 \\ -6 & 9 & 18 & 21 & 18 & 9 & -6 \\ 4 & -6 & -12 & -14 & -12 & -6 & 4 \end{matrix} \quad \text{divided by } 441$$

2.1.3. Central core and magnitude 2nd derivative images

The 2nd derivative is computed as the vector (dx^2, dy^2) using the following difference formulae [22] (in a similar manner to the convolution filter of 2.1.2). These filters are applied to the averaged image just being computed.

$$dx^2 = \begin{matrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{matrix} \qquad dy^2 = \begin{matrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{matrix}$$

The *magnitude* image of g'' is approximated by the city block distance – the sum of the absolute values of dx^2 and dy^2. The direction image is not actually computed. Instead a *central core* image pixel is defined as having a '1' where $(dx^2 < 0)$ and $(dy^2 < 0)$, and being '0' everywhere else. Both the average and central core images are computed during the first raster scan through the image.

2.1.4. Extracting a spot

In a second raster pass through the image only those pixels coded as '1' are processed. Isolated pixels, defined as being 4-neighbor unconnected, are marked for deletion by setting them to a 255 code. Otherwise, each time a '1' code is encountered, a spot pixel list (SPL) is computed as in step [2.2]. A push down stack is used to keep track of all unexpanded pixels. An unexpanded pixel is defined as one found to be a neighbor of an expanded pixel but which has not been checked (i.e., is not on the SPL or push down stack). Each unexpanded pixel is expanded and checked to determine whether any of its 4-neighbor pixels have a '1' code and are not already in the SPL. Unexpanded pixels so identified are put into the push down stack while the pixel being investigated is saved in the SPL. The algorithm keeps processing the push down stack until it is empty. The spot will then be processed to completion using the SPL which will grow as the spot is propagated to the region approximated by the 2nd derivative magnitude function's second local maximum.

2.1.5. Removing concavities

Single pixel wide artifactual concavities which occasionally occur and are removed by checking each SPL pixel C for the following 4 neighborhood conditions. (By neighborhood we mean here the central pixel in question and its 8 adjacent neighboring pixels, see [23].) If a condition is found to be true, the '0' valued pixel in the central core image is changed to a '1'. The SPL is also updated. In each of the following cases, a '0' and '1' *must* occur and a '–' means 'don't care'.

$$\begin{matrix} 1 & 0 & 1 \\ 1 & C & 1 \\ - & - & - \end{matrix} \quad \text{or} \quad \begin{matrix} 1 & 1 & - \\ 0 & C & - \\ 1 & 1 & - \end{matrix} \quad \text{or} \quad \begin{matrix} - & - & - \\ 1 & C & 1 \\ 1 & 0 & 1 \end{matrix} \quad \text{or} \quad \begin{matrix} - & 1 & 1 \\ - & C & 0 \\ - & 1 & 1 \end{matrix}$$

2.1.6. Numbering the central core image

In the central core image, the spot is then assigned the next sequential number in the range of [2 : 99] modulo 100. All SPL pixels in the central core image for that spot get that number. It is very unlikely but in an extremely densely populated spot image, it is possible for two adjacent spots to have the same value for successive lines. This notation problem is easily solved by alternative coding schemes using larger numbers.

2.1.7. Propagating the central core

The numbered spot is then propagated with the value ($C + 100$) from the central core (C) value of the spot to the *propagated central core* region. This propagation from a central core edge point is performed in each of the 4-neighbor directions until it is terminated based on various constraints. (Whereas the 8-neighbor definition of a neighborhood included the corner pixels, the 4-neighbor definition does not [23].) The SPL is updated with the new pixels. The heuristic propagation termination conditions are:

(1) The 2nd derivative magnitude is increasing (starting 1 pixel out from the central core), outward from the central core indicating a second local maxima.
(2) The 2nd derivative magnitude outward from the central core has the same value twice in a row indicating a noisy edge.
(3) The propagation would impinge on another central core pixel.
(4) The propagation would extend beyond the computing window.
(5) The propagation would impinge on an isolated pixel.
(6) The gray value outward from the central core is increasing instead of decreasing indicating that the spot is overlapping a much larger spot.

2.1.8. Corner filling

This type of heuristic propagation sometimes forms small rectangular empty corner regions in the four corners of the spot. Such corners can be filled with propagated central core values. Both 0 and 255 (isolated pixel) corner values are candidates for filling (by changing to N, i.e., $C + 100$) if the central pixel is its central core. The four corner cases are expressed as neighborhood conditions as follows. In the corresponding positions of the neighborhood surrounding a pixel in question, C is the central core value, N is the propagated central core value, E is either 0 or 255, and '–' meaning 'don't care'.

$$\begin{matrix} E & N & - \\ N & C & - \\ - & - & - \end{matrix} \quad \text{or} \quad \begin{matrix} - & N & E \\ - & C & N \\ - & - & - \end{matrix} \quad \text{or} \quad \begin{matrix} - & - & - \\ - & C & N \\ - & N & E \end{matrix} \quad \text{or} \quad \begin{matrix} - & - & - \\ N & C & - \\ E & N & - \end{matrix}$$

2.1.9. Hole filling the central core

In very large saturating spots, the center of the spot may not be detected as such and thus not segmented. The spot will have a doughnut topology. This problem is repaired by filling any artifactual holes in the central core region. The leftmost and rightmost horizontal coordinates for each line of the central core are found and saved as run length codes [23].Then any 0's in the central core image between these points are changed to central core values.

2.1.10. Concavity filling of propagated central core

Occasionally, concavities may appear in the propagated central core image. These are filled by applying the same hole filling algorithm as in step [2.1.5] but for *all* spot pixels.

2.1.11. Round corners

The propagation algorithms applied above tend to leave the corners rather sharp. These are rounded out by applying the following neighborhood conditions (as in step [2.1.8]) to each central core pixel. If an exact match is made, then the 0 pixel on the diagonal is propagated and thus rounds out the corner (value N being $C + 100$).

0	N	–	–	N	0	–	–	–	–	–	–
N	C	–	–	C	N	N	C	–	–	C	N
–	–	–	–	–	–	0	N	–	–	N	0

2.1.12. Spot features and initial sizing

After the final SPL is computed, it consists of the pixels in the central core and propagated central core. Several features are computed using density values mapped from the average image. A preliminary spot sizing is performed to remove most of the background noise spots in step [2.1.8] where a 254 code is also placed in each deleted spot pixel in the propagated central core image.

2.1.13. Background correction

A background density correction is performed during a third pass through the image using a zonal notch filter algorithm similar to that described in [37]. A running average of the averaged image is computed (see [38] for algorithm description) for a $n \times n$ movable averaging window masked by the *complement* of the central core image. That is of background pixels which are not isolated pixels, deleted spots, central cores or propagated central cores. The result constitutes the background image and may be saved in the 2nd derivative magnitude scratch image since it is no longer needed. The mean background for a spot is then estimated by reading the background image at the centroid of the spot.

2.1.14. Computing corrected density and secondary sizing test

The features presently used to determine acceptance of a spot include: spot area (in square pixels), total integrated corrected spot density and range of pixel OD seen in the spot. The last is useful for eliminating small noise spots from the image. The corrected spot density D' is computed from D taking the background estimate into account. Those spots which meet the criteria are saved in the Gel Segmentation File (GSF). Spots failing the final density sizing criteria are deleted as before by having 254 placed in the central core image of each pixel. The gray value numeric data for the final set of spots may be optionally saved in an output image.

It is possible for the spot feature sizing parameter limits to be made more restrictive to eliminate some of the smaller noise objects. Table 2 shows part of a gel segmentation file for a typical gel.

2.2. Example of gel segmentation

We now show some image segmenter output which illustrates the wide range of effectiveness of the algorithm. Other results of segmentation are deferred to later. The segmentation algorithm appears to be applicable to a wide range of gel

TABLE 2

Example of Gel Segmentation File

```
SG2DRV :Version March 17, 1981 - 5:12AM
Today's date is 3/27/1981, 10:00:18 AM
User:[61,1]
Gel Segmentation File is: P20250.GSF
0250.2/P388D1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,TO,CONTROL,BOTTLE#1/
L00157/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 65 88 109 126 144 159 170 181 190 196 208 213 0 0 0 31 460 51 400
Switches: /7X7LOWPASS /ALLOWTOUCHINGEDGES
Window [31:460,51:400]
Area sizing limits (13.00: 500.00)
Density sizing limits (1.00: 500.00)
Density range sizing limits (.05: 2.70)
Saving output image in [61,2]Z00157.PIX
Mean background matrix in ND (std dev)
 .13(+-.03) .12(+-.03) .12(+-.03) .12(+-.03) .13(+-.04) .16(+-.05) .20(+-.06)
 .14(+-.03) .14(+-.03) .17(+-.04) .19(+-.04) .23(+-.04) .23(+-.05) .32(+-.12)
 .15(+-.03) .16(+-.03) .23(+-.05) .25(+-.05) .29(+-.06) .34(+-.05) .30(+-.05)
 .15(+-.03) .18(+-.04) .21(+-.04) .23(+-.04) .25(+-.05) .27(+-.05) .25(+-.05)
 .15(+-.03) .16(+-.04) .16(+-.03) .18(+-.04) .21(+-.04) .21(+-.04) .21(+-.04)
 CC#  1 M.E.R[ 436:444,  55: 59] D.R.=[  .13: .42] D/A= .259 MnB= .128
  1st MOM[ 440.26, 56.35] A=   22  D= 5.70 D'= 2.88 (D'/totalD')%=  .06%
  Sx= 1.68 Sy= 1.10 Sxy=  .79     V= 5.51
 CC#  2 M.E.R[ 448:453,  63: 69] D.R.=[  .28: .68] D/A= .527 MnB= .128
  1st MOM[ 451.63, 66.02] A=   25  D= 13.18 D'= 9.98 (D'/totalD')%=  .19%
  Sx= 1.36 Sy= 1.57 Sxy=  .99     V= 10.49
 CC#  3 M.E.R[ 436:439,  63: 69] D.R.=[  .21: .41] D/A= .327 MnB= .128
  1st MOM[ 438.01, 66.47] A=   20  D= 6.55 D'= 3.99 (D'/totalD')%=  .08%
  Sx=  .94 Sy= 1.63 Sxy=  .88     V= 4.47
        .
        .
        .
 CC# 685 M.E.R[ 144:150, 397:399] D.R.=[  .16: .34] D/A= .234 MnB= .128
  1st MOM[ 147.03, 398.31] A=   19  D= 4.44 D'= 2.01 (D'/totalD')%=  .04%
  Sx= 1.88 Sy=  .83 Sxy= 1.03     V= 3.72
Total of    686 accepted D spots accumulated density= 7872.56, area= 20625
Total of    685 accepted D' spots accumulated density= 5232.56, area= 20625
Total of   7265 omitted spots accumulated density= 8518.47, area= 42756
Omitted/Accepted density =  108%
```

Part of the gel segmentation file (GSF) output of the SG2DRV program for a ^{14}C-labeled P388D1 macrophage like cells autoradiograph gel *ACC#* 250.2. Segmenter parameters and some of the spot feature list data are presented. *CC* is the spot connected component number, *MnB* is mean background density, *A* is spot area, *D* is uncorrected total spot density and corrected density *D'* is computed as $D - (A)(MnB)$. *D/A* is the mean density and (*D'*/Total*D'*)% is *D'* expressed as a percentage of total gel spot density. 1st MOM is the spot's centroid while *D.R.* is the density range of pixels seen in the spot. *Sx*, *Sy* and *Sxy* are the standard deviation and covariance of spot size with *V* being the Gaussian volume estimate of density for this region. Density values are in OD calibrated in terms of the associated ND wedge in the image.

magnifications and densities, i.e., autoradiographs of varying exposures and spot detection modalities, etc. It is also capable of resolving touching spots and other image complications over a wide range of conditions.

The segmenter has been applied to various types of gels (both autoradiographs and silver stain) of different types of material scanned at different magnifications with satisfactory results. At least 500 gels have been segmented using this program. Figure 5 is a composite photograph showing two P388D1 gel images before and after segmentation. As can be seen in Fig. 5b and 5d, the segmentations of spots were successful in a vast majority of cases. The number of spots segmented in these two gels were 547, and 672 respectively.

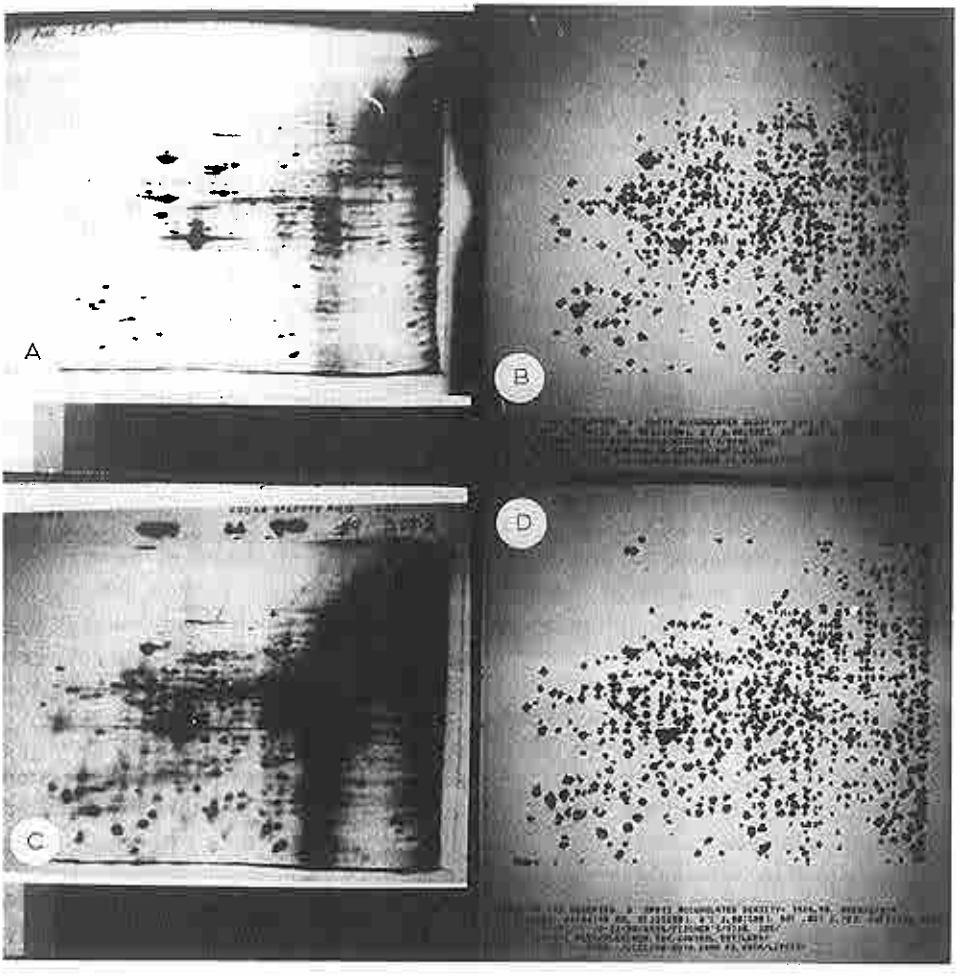


Fig. 5. Composite photograph showing two different P388D1 mouse macrophage gels before and after segmentation has been performed. (a) T0 gel 250.2; (b) segmentation of (a); (c) T24 control gel 258.2; and (d) segmentation of (c).

A set of sizing parameters for various classes of gels have been empirically determined. The parameters currently used for the 250 microns/pixel autoradiographs are: total spot area range [13–500] square pixels, total integrated spot density range [0.1–500] OD, spot pixel density range [0.03–2.7] OD. For the 170 microns/pixel autoradiographs, the total integrated density minimum limit is increased to at least 1.0 OD. For silver stained gels at 250 microns/pixel, the total integrated density minimum is 1.0 OD and the minimum spot pixel density range up to 0.15 OD.

Typical times for segmenting 250 microns/pixel resolution gels on the DECSYSTEM-2020 (KS-10 CPU) are on the order of 8 to 18 min/gel for gels on

the order of 1000 spots depending on the segmentation options (such as low pass filter size, etc.) as well as number of spots. Running times of these programs on a KL-10 CPU TOPS-20 system seem to be of the order of 10 times faster than for the KS-10 processor. The 250 microns/pixel gel images have the computing window set to about 2/3 to 1/2 the area of the gel because of the need to include the ND wedge in the image. These computing times increase slightly when performed over the full 512 × 512 pixel image for the higher resolution gel images. The running times also increase somewhat with the number of spots being segmented.

2.3. *Some other spot extraction algorithms*

We presented the requirements for a gel image segmenter in the context of this particular problem domain. The segmenter we have adopted for analyzing 2D gels is only one of many applicable to this problem. We will show reasons for selection of our approach after consideration of some other segmenters and spot extractors which have been applied to gel images.

A large number of image processing techniques have been brought to bear on the detection and extraction of objects from images [22–23]. Some of these, using global techniques such as thresholding, fail when applied to gels because of inhomogenity of the image. Others using local algorithms which are globally applied have been much more successful.

A threshold segmentation technique [12,21] analyses the gel image density histogram to find the mean background value and then thresholds the image at 0.05 to 0.10 OD above this value to detect faint spots. Darker clusters of spots are segmented together using this technique into a single 'spot'. Each 'spot' is then tested to find maxima and minima regions within it to determine whether it should be iteratively split into subspots. Spots found in this way are then expanded to the region defined by a second derivative of the gray scale data.

Another segmentation technique [8,24] detects spots by first scanning in a raster direction and then orthogonally to find spot maxima. Ellipses are fitted over the detected spots to approximate their boundaries and density information measured. This group has also developed algorithms for spot extraction using convolution techniques [25].

Another method of spot detection is based on finding spots in the image starting at the darkest pixel in the image and then removing fitted spots at each step of the search [10]. The algorithm assumes that the spots are approximately Gaussian in x and skewed Gaussian in y. It fits parametric curves to the spots and then estimates the volume density from the parameters.

Spots may be detected by assembling detected line segments of spots from successive lines using a procedure called 'chain assembly' [11]. The edges of adjacent chains are smoothed and Gaussian curves fit to estimate the spot.

2.4. *A distribution-free density independent spot segmenter*

The algorithm we have presented uses the central core model of a spot. The central

core was defined as the region of negative slope of the second derivative function of the image. Such a region occurs within the first minimum of g'' surrounding the spot's center. This model seems to work robustly on real gel data. It has failed only on those few spots of such a huge extent, saturated, or so noisy that no simple model exists for the spot. In addition, it is independent of assumptions as to exact spatial distributions of density as well as of orientation.

Because the shapes of gel spots corresponds to the physical diffusion process used in their generation, no explicit boundaries are present. Manual measurements made at what observers subjectively define as the boundary have resulted in up to 50% error in total integrated spot density in the case of fuzzy spots. We have chosen to algorithmically define a spot as its propagated central core which is found to be reproducible.

A common problem of many gels is artifactual streaking. One may attempt to remove streaks before processing or alternatively, the spots may be segmented in the context of the streaks. The central core algorithm finds spots regardless of whether the streak is present or not. Therefore streak removal by pre-processing is not necessary.

The central core algorithm finds a spot's peak if it is present and if there is sufficient resolution, in both the spatial and density domains of the image, will resolve overlapping peaks. The cases where it fails can be understood if one looks at the difference formula for g''. This discrete approximation to g'' can not resolve spatial position differences less than about 5 pixels. Thus gels scanned at higher resolution will show fewer unresolved touching spots after segmentation than gels scanned at lower resolution.

Two overlapping saturating spots will also sometimes be unresolved. Because the plateau effect occasioned by saturation obscures the second peak in g'' which is necessary for spot separation. If one wishes to keep track of and modify the segmentation of saturated spots it is necessary to employ a different morphologic analysis. In any case, saturated spots are tracked throughout the entire analysis process. Spots saturating in one gel might not do so in another so that substitute measurements could be made in the case of multiply exposed autoradiographs.

In practice, spots are frequently somewhat distorted so that an idealized Gaussian shaped spot may rarely be found. The segmenter works well for such spots because their extent is defined by the second maximum of the g'' magnitude over *all* of the spot's edge.

The sensitivity of the spot detector (e.g., stain) varies among gels so it is necessary to normalize spots in each gel in order to compare them. Normalization will be discussed later. However, the segmenter performs one type of normalization which is useful for well segmented gels. In addition to reporting each spot's total integrated density (D) and its background corrected density (D'). The segmenter also reports D' divided by the sum of D' for all spots accepted expressed as a percentage. These density features are illustrated in Table 2.

Scans of the same gel at two different resolutions provide an opportunity to investigate differences in segmenter behavior for the same set of spots. Use of PIXODT, image debugger [1], leads to the conclusion that the few instances

where spots are incorrectly merged is due to either (a) lack of spatial resolution or (b) gray scale resolution (spots were close to or at saturation) or were very noisy. Overall, the correlation was good with most spots being correctly segmented.

3. SPOT PAIRING

We have treated the problem of spot extraction within a single gel using the SG2DRV program. Now we consider the first step in locating a particular spot in a set of gels, i.e., pairwise matching of the spot in two gels. This can be done by shifting one of the gels until the spot overlaps and recording the cartesian coordinates of the spot in each gel. This spot by spot pairing is a prerequisite to detecting whether individual polypeptides change with respect to experimental conditions. Furthermore, pairing of spots within a set of gels taken two at a time is the means whereby a multiple gel data base is gradually constructed.

Referring back to Fig. 1, gels are first acquired, then spots are segmented using the SG2DRV program. This resulted in a gel segmentation file (GSF) consisting of a list of spot (x,y, density) triples (as in Table 2). Note that in the table each spot has a spot index (which can be used to refer to the spot), a (x,y) centroid and a density measurement given in several formats. An algorithm for spot pairing between two gels using a small set of landmark spots to locally align subregions is presented which uses the GSF spot list files as data. The CMPGEL program implements this algorithm producing a gel comparison file (GCF).

3.1. Partitioned search in pairing spots

A major problem complicating spot localization is the local distortions in the gel such that neighboring spots in one gel will likely be neighbors in another gel while the intervening distances between them vary to some degree. Several semiautomated methods for aligning corresponding spots in two gels are discussed.

In one, a gel is transformed locally to the distortions of a second gel [12,21,25]. Sets of three evenly spaced corresponding landmark spots are manually defined for both gels covering regions to be transformed. A linear approximation to the affine transformation of this region is performed to translate, rotate and stretch the image locally. After the transformation, spots from the two gels are mapped to the same domain. Then a least squares fitting procedure matches those pairs less than a specified distance apart in the two gels.

An interactive mode using a color display allows comparison of spots between gels while geometric correction is done locally using a linear interpolation in localized regions of a pair of gels [8,24]. The corrected images are then used to produce protein maps for the local regions, can be investigated independently [24] and serve as a basis which to discuss a subset of spots.

3.2. A landmark driven spot pairing algorithm

We present an alternative view of the primary pairing algorithm. This involves in

effect constructing a projected image composed from the two members of the gel pair. In the actual matching, computations are performed only on (x,y,d) data in a single plane, the representive gel plane. Central to the algorithm is the establishment of landmark spots that serve to 'anchor' the other spots in its vicinity. Essentially, landmark spots are manually aligned in the two gels at which point the computer automatically aligns all other spots with the corresponding spots in the other gel. The procedure is simple and is easily extended to align any number of gels.

Such partitioning by landmark region increases the efficiency of integral spot matching by providing an empirical basis for the partitioning of a gel image into tractable corresponding subregions.

A landmark spot should be selected according to particular criteria. It is a morphologically distinctive spot present in all gels such that neighboring spots and the landmark spot form a consistent morphologic structure. Moreover, this morphologic structure should be easily recognized across the set of gels. The landmark spot should not be a touching spot. The set of landmark spots are selected to fairly easily cover the regions of interest of the gel fairly. From 10 to 25 landmarks are generally selected depending on the quality of the gel with fewer required for better gels. This set of spots is called the landmark set. In practice, the operator aligns the landmark spots in the two gel images using the flicker algorithm [9], which permits one of the gels to move while keeping the other constant. Viewing time for each of the images may be independently set and varied until the user is satisfied that the two images of the same spot are locally 'superimposed'. The superimposed spot of interest is noted to the computer and the next landmark spot processed in the same fashion. The flicker procedure is described in [9] and landmark acquisition in [2].

GELLAB also offers alternative facilities for generating landmark spot data without using our special purpose RTPP interactive hardware. Program DWRMAP draws a labeled outline-plot of the segmented gel from the GSF file with the darkest spots (sorted by density) labeled in the plot with a table also given on the side of the plot. Spots are represented by an oval proportional to spot density. By manually comparing such R-map plots, corresponding lists of landmark spots can be defined using the CC#s of each gel. Program LMSEDIT then allows the manual entry of a landmark set as a list of pairs of CC#s from the two gels.

The landmark region surrounding a landmark spot is defined as a polygonal region having higher pairing certainty for spots closer to the landmark spot. The half-radius of certainty Ri for landmark i is a distance defined to be half the distance from the nearest landmark spot landmark i. Spots within the half-radius of a landmark set have a higher probability of being aligned (since the landmarks have 'perfect' integral alignment) than if the spot were outside of this radius. The landmark spots in each gel are compared with the two GSF spot lists and the best segmented spot's centroid is used rather than the coordinates manually produced. If no spot can be found for a landmark within specified error bounds (currently the dT2 distance; see below), then the manual landmark coordinates are used. The CGELP program to be discussed (cf. Section 4) has an operator VALIDLANDMARKS which computes a table and statistics of all valid landmarks for all gels in a multiple gel data base. It is

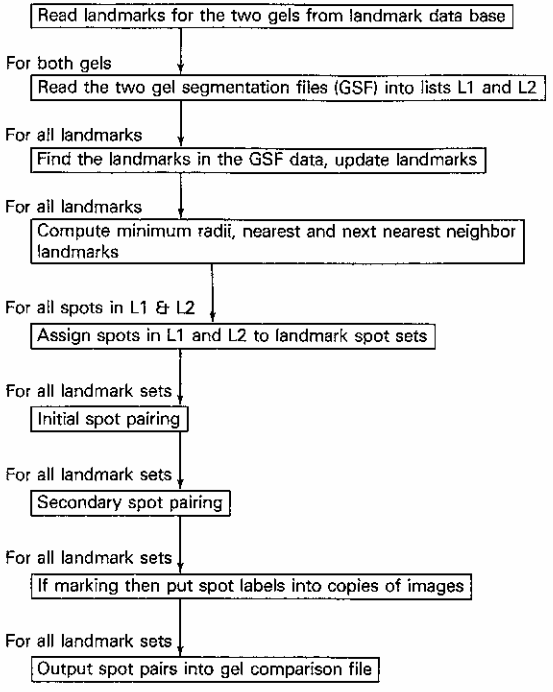


Fig. 6. Block diagram of the 2D-gel comparison procedure where spot segmentation lists are merged using a small set of landmarks to produce partitioned landmark sets of paired spots.

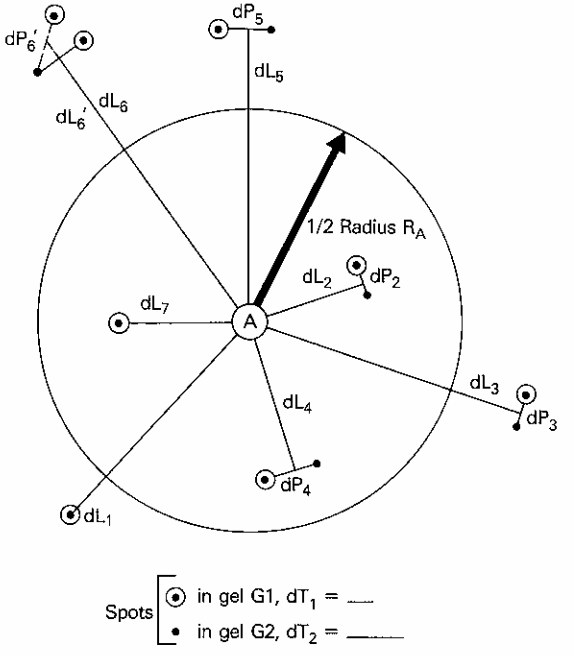
possible to ascertain how reliable the particular pairing actually was by backchecking paired spots in a set of pairing-labeled images to be discussed.

Partioned search has the added advantage that landmark regions contain an order of magnitude fewer spots than the total gel space. Therefore the combinatorics of performing the spot matching is greatly decreased as well.

3.3. *Algorithm for landmark-oriented spot pairing between two gels*

The spot pairing algorthim is illustrated in flow chart form in Fig. 6. It is implemented as the CMPGEL program. Pairing is performed in two passes through the landmark sets data using the primary and secondary pairing procedures. Each procedure operates on one landmark set at a time, in both gels.

In the primary pairing algorithm (Fig. 7), spots are first mapped to the Cartesian coordinate system defined by shifting the landmark spot to (0,0) relative to the origin in the two gels G1 and G2. Each spot in G1 is provisionally paired to the spot that is its nearest neighbor (by minimum Euclidean distance) in the projected image of G2. Because of possible asymmetry of the two landmark regions, the reverse comparison is also performed so that each spot in G2 is provisionally paired with its nearest neighbor spot in G1. This nearest neighbor distance is denoted dP (pair



Fig, 7. During primary spot pairing labeling assignment, each potential nearest neighbor spot pair in a landmark set is assigned one of four labels: SP, sure pair; PP, possible pair; AP, ambiguious pair; US, unresolved spot. The labelings are defined by the following cases: [1] US – unresolved spot (no *dP*); [2] SP – $dL2 < Ra$ and $dP2 < dT1$; [3] PP – $dL3 > Ra$ and $dP3 < dT2$; [4] PP – $dL4 < Ra$ and $dP4 > dT1$ and $dP4 < dT2$; [5] PP – $dL5 > Ra$ and $dP5 < dT1$; [6] PP – $dL6 > Ra$ and $dP6 < dT2$. For the other spot AP' – $dL6' > Ra$ and $dP6' < dT2$ and $dP6' > dP6$; [7] US – unresolved spot (no *dP*).

distance). The distance from the landmark spot to the mean locus of the two spots in the provisional pair is denoted dL. Two user specified parameter distances are empirically defined: dT1 and dT2. Spots closer than dT1 to the landmark spot are relatively well paired. Spots greater than dT2 are poorly paired and possibly should not be paired. The current values of dT1 and dT2 (5 and 10 pixels respectively) were determined empirically, by examination of the nearest neighbor values of many sets of paired gels under gel resolution range of 170 to 250 microns/pixel. Figure 7 shows various pairing which can occur. Four types of pairing labels can be defined. These are sure pair 'SP', possible pair 'PP', ambiguous pair 'AP' and unresolved spot 'US'. The primary spot pair labeling assignments are defined in Fig. 7.

The primary pairing algorithm is a simple first order model not taking some spots on the periphery of the landmark region into account. These spots may be misclassified as an AP or US whereas they would be a SP and PP classification in another adjacent landmark region. To correct these few misclassification errors, a secondary pairing algorithm is applied in order to possibly re-pair AP and US spots in the next-nearest landmark set using AP and US spots from those sets. The resultant re-paired spot pair (either a SP or PP if it meets the threshold criteria) is then placed in the landmark set with the smallest dL value.

3.4. *CMPGEL output*

Finally, after spot pairing, the program can optionally draw the labels into copies of the original images. The paired spot data, (x,y,d) sorted by landmark sets, are then output into the gel comparison file (GCF). Other information regarding the identity of the two gels and gel segmentation files as well as the manually defined landmarks is part of the permanent preface to the GCF. The estimated landmark spots from the GSF found in the GSFs are also reported as is the Euclidian distance from them to those manually defined by the user. If this distance is greater than dL from a landmark for either G1 or G2, then that landmark spot is so marked and the GSF spots are partitioned using the manually defined coordinates, landmark spot sets. At the end of the GCF is a statistics summary for both the primary and secondary pairing reguarding the number of each of the four pairing assignments.

3.5. *Example of spot pairing*

The spot pairing algorithm just described has been in use over the past year and has been applied to over 500 gels. Figure 8 shows the landmark spots for the pair of P388D1 gels with each of these spots marked with a small plus sign and the landmark name to its right. Table 3 illustrates a typical landmark set entry for the 250.2/258.2 pair of gels.

After the spot labels have been assigned, copies of the original images may be overwritten with the label names for all spots in the spot lists. SP, PP and AP labels appear in the marked image as 'S', 'P', and 'A' while US appears as a small '+' (because using the larger 'U' symbol would crowd the image where it is noisy). Figure 9 shows the labeled pairs for the P388D1 gel pairs where landmark spots in these images have a box around them. Table 4 shows part of the output of a typical GCF for gels 250.2 and 258.2. The first part of the table illustrates the landmark registration and parameters while the rest shows some representative paired spots as well as pairing statistics.

Secondary pairing for gels 250.2/258.2 increased the (SP+PP)/total spots percentage from 62.7% to 64.0%. In general we observe a 1.5% to 3% increase in (SP+PP) labeling which seems to indicate that most pairing is performed (as expected) during primary pairing. When two very different gels (with widely different number of total spots extracted) are compared as is expected a fairly high number of US and AP spot labels result. This does not mean that the spots that are SP or PP paired are not paired well. For similar gels, the (SP+PP) ratios are in the range of 65% to 85% depending on the artifactual noise of the gels.

3.6. *Global gel matching*

In general, the global matching of two gels would involve a D'arcy Thompson type transformation [26] of one of the gels to bring it into congruence with the other. The set of points on the grid are displaced by a continuously differentiable 2D distortion function in this transformation. One approach to gel analysis is to remove

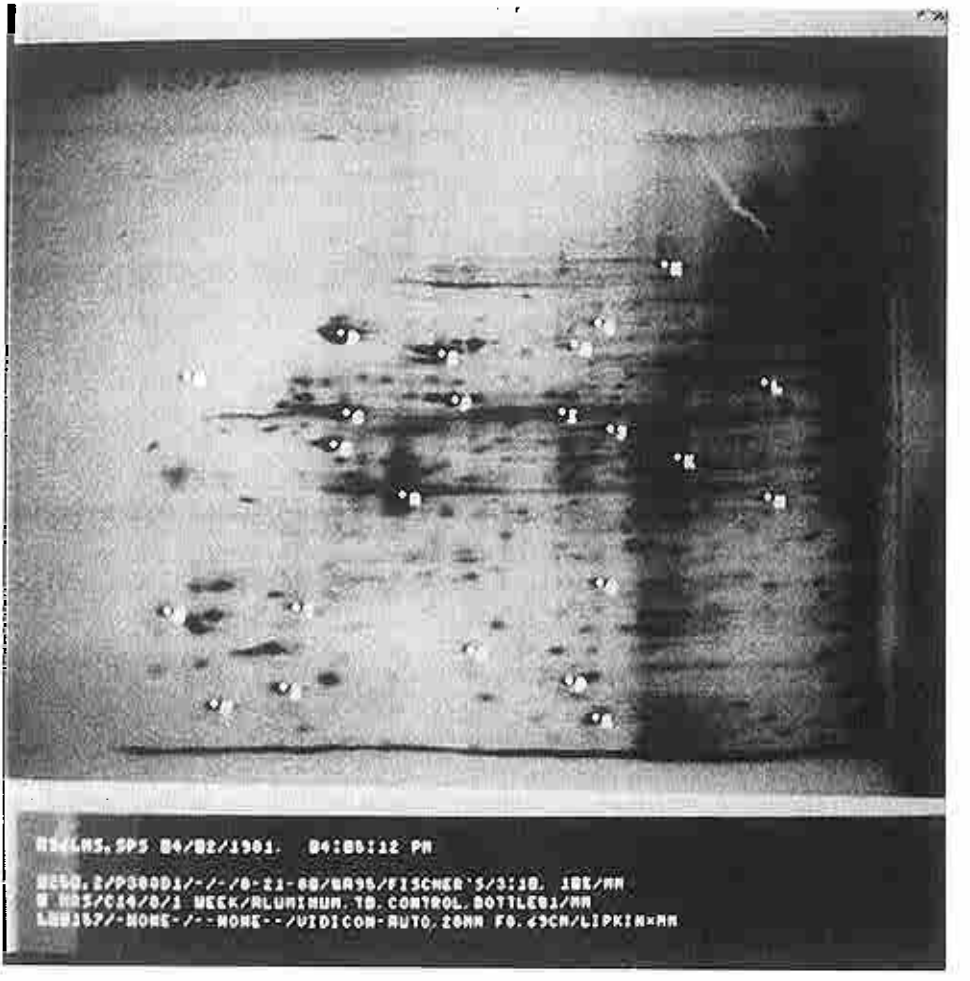


Fig. 8. Landmarks used in pairing the two P388D1 gels superimposed on R-gel 250.2. The landmark is defined at the small '+' sign with its landmark set name its right.

the distortion (as is currently done with satellite images) and then perform a point by point comparison [22]. This correction implies some knowledge of the complete inverse distortion function that is by and large lacking for 2D gels although estimates are computed using landmark triples [12,21,25].

The amount of computation required to perform the D'arcy Thompson image transformation for every pixel in the image is considerably greater than for simply pairing spots which may be thought of as sublandmark registration. Since the actual gel analysis requirement is to pair spots for comparison, it is more efficient to simply pair locally corresponding spots rather than to transform the gel images themselves and then compare all of the spots.

The landmark driven pairing algorithm has an additional advantage, i.e., it is also

TABLE 3

Example of a landmark set from the landmark data base file

```
/ CMPGEL: VER# 9/23/80 - 9:09AM
/ INTO SYS@:JUNK@@.DA FROM GSF FILES: P20250.GS AND P20258.GS
/ SURE!PAIR THRESHOLD= 5, POSSIBLE!PAIR THRESHOLD= 10
01/11/1981,  11:14:04 AM
 LANDMARK #A  G1[211, 262], G2[199, 265]
 LANDMARK #B  G1[173, 235], G2[160, 236]
 LANDMARK #C  G1[180, 219], G2[166, 220]
 LANDMARK #D  G1[177, 176], G2[162, 177]
 LANDMARK #E  G1[231, 185], G2[221, 186]
 LANDMARK #F  G1[239, 211], G2[229, 211]
 LANDMARK #G  G1[318, 168], G2[312, 168]
 LANDMARK #H  G1[305, 182], G2[297, 183]
 LANDMARK #I  G1[292, 218], G2[285, 219]
 LANDMARK #J  G1[325, 226], G2[317, 227]
 LANDMARK #K  G1[363, 241], G2[354, 244]
 LANDMARK #L  G1[410, 200], G2[400, 201]
 LANDMARK #M  G1[355, 138], G2[350, 137]
 LANDMARK #N  G1[413, 261], G2[399, 268]
 LANDMARK #O  G1[322, 308], G2[311, 311]
 LANDMARK #P  G1[307, 362], G2[293, 368]
 LANDMARK #Q  G1[321, 382], G2[308, 388]
 LANDMARK #R  G1[248, 346], G2[235, 349]
 LANDMARK #S  G1[149, 367], G2[134, 371]
 LANDMARK #T  G1[154, 321], G2[133, 324]
 LANDMARK #U  G1[ 89, 324], G2[ 65, 324]
 LANDMARK #V  G1[115, 375], G2[ 94, 381]
 LANDMARK #W  G1[ 99, 198], G2[ 71, 196]
```

An example of a typical landmark set entry in the landmark data base file. Entries are accessed by gel name pairs, (e.g., by the gel name pairs (250.2,258.2) or (258.2,250.2)). G1 (G2) corresponds to gel 250.2 (258.2) The G1 $[x,y]$ 2-tuples are the positions of the spots in gel i for the specified landmark and similarly for G2.

applicable to automatically defined landmarks, should they become available. One interesting technique that has been used to circumvent the gel distortion problem is double labelling [27]. Spot pairing is greatly simplified if two gels are congruent. One gel sample is labeled with ^{14}C and the other with ^{3}H-labeled amino acids when the samples are grown in deficient media. The two samples are then added together just prior to running the gel. Two-step autoradiography is performed for the ^{14}C- and for ^{3}H-labeled gel. Unfortunately, this technique can only be used with material which can be radioactively double labeled. An analogous role could be played by fluorescent labeling of known spots for stained gel images.

3.7. Biases due to landmark selection

The partition of the plane into variable sized polygonal landmark regions is based essentially on the local spread of landmarks. A priori one would think that pairing would be more likely to be correct in regions of high landmark concentration. However, a small radius of confidence may have undesirable effects on pairing, i.e., spots that would otherwise be matched as sure pairs might be entered into the probable pair category. It is possible for a spot to be paired to be found in the *next* to

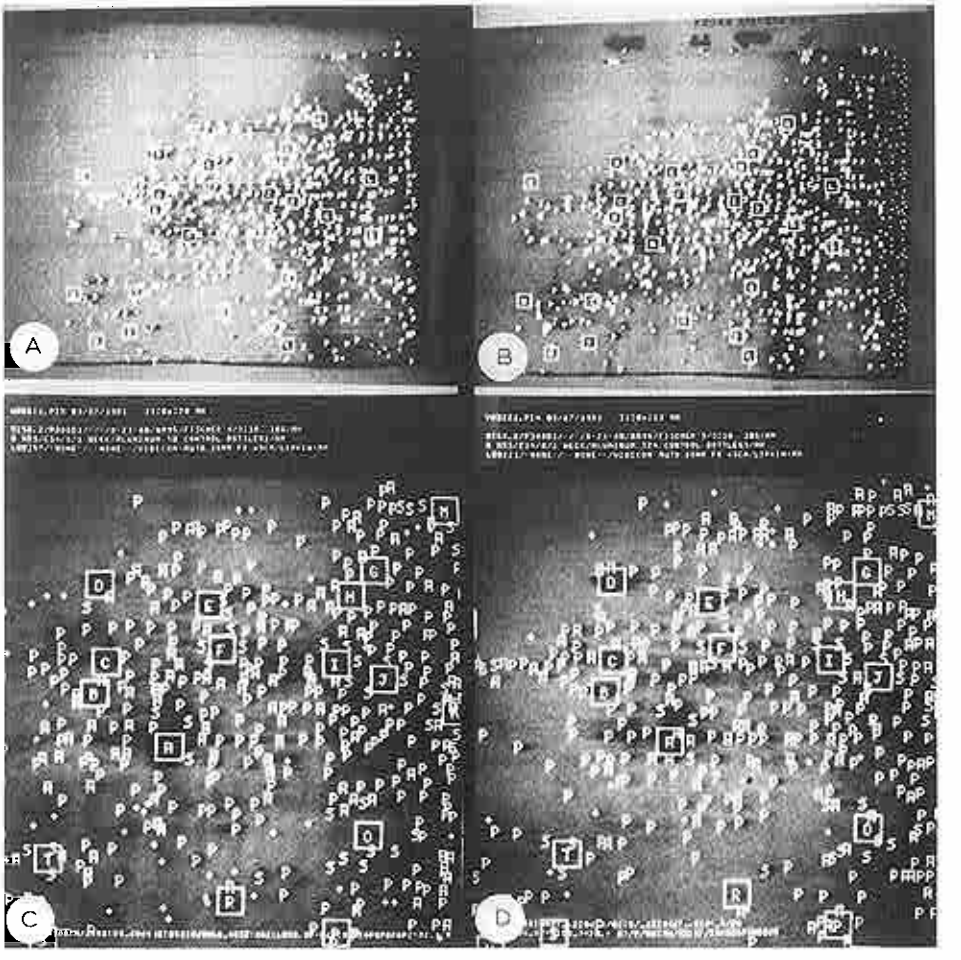


Fig. 9. CMPGEL labeled marked images with pair labels: S, sure pair; P, possible pair; A, ambiguous pair; '+', unresolved pair. Gels 250.2 and 258.2 are labeled in (a,b), a 2× magnified central region of these images is in (c,d).

next nearest neighbor landmark set rather than the landmark or next nearest landmark sets. In digital space, the problem of a possible shift of a spot from one landmark region to another, such that pairing would be affected, as a result of increasing the concentration of landmarks is obscure and does not seem easily treated.

The consideration of correctness and completeness of the primary pairing algorithm is not simple although the algorithm in itself is quite straightforward. Performance should not be gauged exclusively on the results when gels of widely different spot numbers are compared. On the other hand, comparisons of closely similar gels should yield good results.

TABLE 4

Example of Gel Comparison File

```
CMPGEL[50,32]: Version Jan 26, 1981 - 1:50PM
Today's date is 03/27/1981,  11:01:56 AM
User: [61,1]
Gel Comparison File is: C20258.GCF from P20250.GSF and P20258.GSF
0250.2/P38bD1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,TO,CONTROL,BOTTLE#1/
L00157/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 65 88 109 126 144 159 170 181 190 196 208 213 0 0 0 31 460 51 400
0258.2/P38BD1/-/-/8-21-80/#A95/FISCHER'S/3:10, 10%/
0 HRS/C14/8/1 WEEK/ALUMINUM,T24,CONTROL,BOTTLE#9/
L00221/-NONE-/--NONE--/VIDICON-AUTO,28MM F8,69CM/LIPKIN*
 40 65 88 110 127 144 159 171 182 191 197 208 213 0 0 0 2 491 53 405
Distance sizing limits (dP1 =  5.00, dP2 =  10.00):
Switches:  /MARK
LMSLL.LM from gel ACC#'s 0250.2 and 0258.2
The (Representitive)R-gel is: 0250.2
 LANDMARK #A  G1[211, 262], G2[199, 265]
 LANDMARK #B  G1[173, 235], G2[160, 236]
 LANDMARK #C  G1[180, 219], G2[166, 220]
        .
        .
        .
 LANDMARK #W  G1[ 99, 198], G2[ 71, 196]
G1[A, 437][ 208, 263],E.Diff= 3.2, G2[A, 502][ 198, 266],E.Diff= 1.4-OK
G1[B, 358][ 173, 236],E.Diff= 1.0, G2[B, 413][ 158, 237],E.Diff= 2.2-OK
G1[C, 287][ 177, 219],E.Diff= 3.0, G2[C, 326][ 164, 220],E.Diff= 2.0-OK
        .
        .
        .
G1[W, 228][  96, 199],E.Diff= 3.2, G2[W, 260][  68, 198],E.Diff= 3.0-OK
R[A]=  22 to nearest LMs[B,B], next nearest LMs[C,C]
R[B]=   9 to nearest LMs[C,C], next nearest LMs[A,A]
R[C]=   9 to nearest LMs[B,B], next nearest LMs[A,A]
        .
        .
        .
R[W]=  42 to nearest LMs[D,D], next nearest LMs[C,B]
Marked gel comparison files are: U00221.PIX and V00221.PIX on [61,2]
G1 HAS  685, G2 HAS  825 SPOTS
TOTAL DENSITY G1= 5232.56, G2= 9928.50
OMITTED TOTAL DENSITY G1= 8518.47, G2= 11116.49

LM[A] G1 HAS   46, G2 HAS   45 SPOTS
#A G1: 549[-22, 43]&G2: 619[-25, 34] PP,DP=9.5,DL=48,D1=1.4,D2=1.6od
         .25Maxd1 .29Maxd2  15A1  13A2 .19MinD1 .23MinD2
         1.40sX1 1.18sX2 .81sY1 1.05sY2
#A G1: 352[ 1,-29]&G2: 403[ -2,-32] PP,DP=4.2,DL=32,D1=5.1,D2=11.0od
         .44Maxd1 .56Maxd2  20A1  34A2 .34MinD1 .41MinD2
         1.60sX1 1.64sX2 1.06sY1 1.64sY2
#A G1: 369[ 0,-23]&G2: 429[ 6,-23] AP,DP=6.0,DL=24,D1=10.5,D2=17.5od
         .54Maxd1 .74Maxd2  36A1  43A2 .30MinD1 .36MinD2
         2.11sX1 1.53sX2 1.34sY1 2.25sY2
#A G1: 376[ 5,-20]&G2: 429[ 6,-23] PP,DP=3.2,DL=24,D1=9.0,D2=17.5od
         .57Maxd1 .74Maxd2  26A1  43A2 .32MinD1 .36MinD2
         1.06sX1 1.53sX2 2.01sY1 2.25sY2
#A G1: 394[-5,-17]&G2: 450[ -7,-19] SP,DP=2.8,DL=20,D1=23.7,D2=29.2od
         .73Maxd1 .97Maxd2  75A1  56A2 .25MinD1 .35MinD2
         2.96sX1 2.02sX2 1.87sY1 2.06sY2
#A G1: 514[-30, 24]&G2:   0[  0,  0] US,DL=39.0,DL=39,D1=1.4,D2= .0od
         .25Maxd1 .31Maxd2  16A1  19A2 .18MinD1 .24MinD2
         1.08sX1 1.14sX2 1.35sY1 1.38sY2
#A G1:   0[  0,  0]&G2: 616[-20, 34] US,DL=33.0,DL=33,D1= .0,D2=1.6od
         .23Maxd1 .28Maxd2  15A1  14A2 .18MinD1 .23MinD2
         1.15sX1 1.27sX2 1.31sY1 .92sY2
                .
                .
                .
LM[B] G1 HAS  19, G2 HAS  30 SPOTS
#B G1: 306[-23,-15]&G2: 355[-26,-14] PP,DP=3.2,DL=30,D1=5.4,D2=1.7od
         .41Maxd1 .30Maxd2  29A1  13A2 .20MinD1 .25MinD2
         1.63sX1 1.06sX2 1.52sY1 .91sY2
#B G1: 401[  4, 37]&G2: 549[  1, 38] PP,DP=3.2,DL=38,D1=1.7,D2=2.0od
         .28Maxd1 .30Maxd2  14A1  15A2 .23MinD1 .25MinD2
         1.16sX1 1.05sX2 1.15sY1 1.14sY2
                .
                .
                .
PAIRING STATISTICS
------------------
        After Initial pairing:
US 194
SP 178
PP 732
AP 348
(SP+PP)/(US+AP+SP+PP)= 62.7%
```

The current pairing algorithm defines a sure pair (SP) as being within the landmark radius R/i for a given landmark i. We have found that most of the possible pairs (PP) are actually paired and should be pooled with the sure pairs as well-matched spots. Large numbers of ambiguous pairs and unresolved spots result when comparing two widely different or noisy gels. Currently, nothing is done with these (AP and US) spots. Although they are tracked through the data base. A possible extension to GELLAB processing would be to incorporate additional procedures to further process the AP and US spots such as merging AP fragments with the spots they belong with. Conglomerates of spots sometimes appear as single spots and other times as several spots, e.g., actin complex, so that merging spots is an attractive idea under the right conditions.

We have found that highly populated spot regions should have somewhat more landmarks, however landmarks should not be 'on top of' each other. Other criteria in landmark selection include using fewer landmarks if the regions have little distortion and line up fairly well. A landmark spot should be well defined morphologically and non-touching being part of a locally consistent pattern in all of the gels to be compared.

Manually landmarking a pair of gels takes from 3 to 30 min depending on the comparability of the gels with an average time being about 4 to 7 min. These times depend on gel quality (the single most important factor) and the set of landmarks selected (which must be in all gels to be compared). CMPGEL processing times are on the order of 2 min.

The ability to pair most of the spots in a set of gels enables examination of larger gel data bases where subtle shifts and correlations in the spot data can be more easily detected. Figure 10 shows a log density – log density scatter plot of two normalized paired (SP and PP labels only) P388D1 gels. Most of the spot pairs are close to the 45 degree line. Some of the outliers are real and some are due to noise in the entire gel-image processing system. We will now consider techniques for further resolving noisy data using multiple gels and means for facilitating the checking of outliers.

```
        After secondary pairing:
US 186
SP 178
PP 751
AP 337
(SP+PP)/(US+AP+SP+PP)= 64.0%
```

This is an example of the first part of the gel comparison file (GCF) output of the CMPGEL program applied to gels 250.2 (T0) and 258.2 (T24) in the effect of time experiment. The manually selected landmark spots are listed followed by the number of spots in each gel. The best spot estimates of the landmark spots from the GSF data are then given. The Euclidian distance betwen the segmented and manually defined landmark spots indicates how well these spots fit the landmark estimation. Half-radii and next-nearest neighbor landmark spots are then listed followed by total spot densities listed for each gel for both included and noise (omitted) spots. The second part of the GCF contains labeled landmark sets GCF with pairing statistics for gels 250.2 (G1 is the R-gel) and 258.2 (G2). In each paired spot, the bracketed 2-tuple is the relative cartesian distance from the spot to the landmark. The segmentation spot index preceeds the '[' for both G1 and G2. The spot reliability labels are (SP,PP,AP,US). Distance DP is defined as that between spots in a pair and DL is the distance of a pair from the landmark. D1 (D2) is the D' (background corrected) density measurement of the spot in G1 (G2). Od is the maximum OD value seen for either spot in the pair. MaxD1 (MaxD2) and MinD1 (MinD2) are the maximum and minimum density values seen for any pixel in the spot.

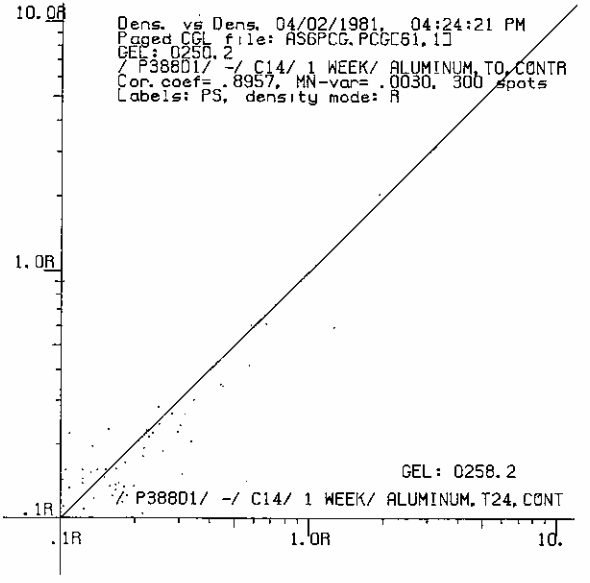


Fig. 10. Density vs. Density scatter plot (on a log–log scale) of P388D1 gels 250.2 (T0) and 258.2 (T24). The density values were first normalized using least squares (cf., Section 4) to gel 250.2.

3.8. Role of pairing 2 gels in multiple gel data base analysis

When comparing corresponding spots among a number of gels, the pairing performed by this algorithm is a prerequisite for the analysis of multiple gels where one treats the values of particular spots within a set of gels. Figure 11 shows the steps in the data reduction performed in the multiple gel analysis in GELLAB. Gel segmentation files (GSF) consisting of spot lists are merged using the gel pairing algorithm into gel comparison files (GCF). These in turn are merged into a gel data base (PCG). We next discuss this last procedure and its ramifications for 2D gel analysis.

4. MULTIPLE 2D GEL ANALYSIS

Earlier we have discussed the need for computer support of 2D gel electrophoresis analysis. Such support along largely data structural lines has been shown to be essential. We have treated the problems of spot extraction and pairwise spot comparison and in the process have indicated that experiments involving time or dose variables require comparisons of spots from multiple gels. We now deal with multiple gel comparisons, the most powerful and demanding mode of application of 2D electrophoresis to biological and clinical investigation and describe a computer

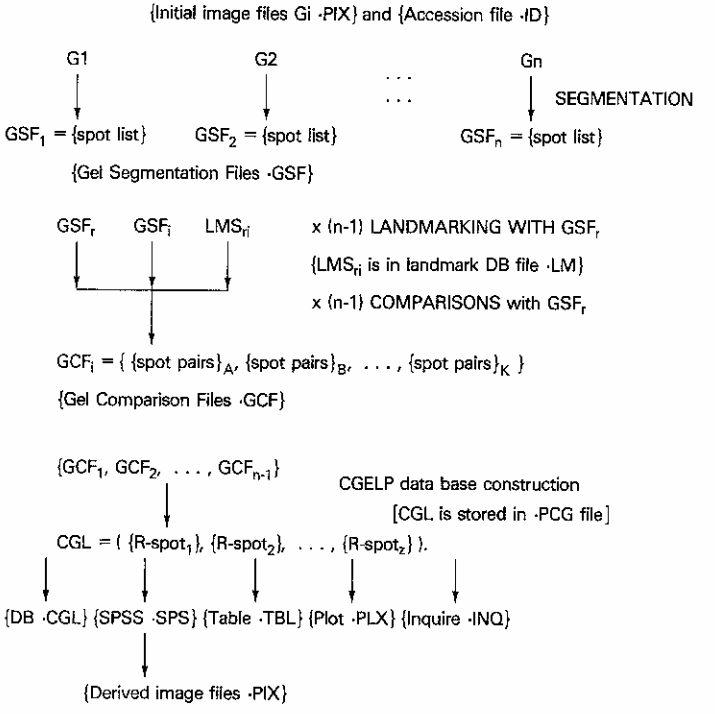


Fig. 11. Data file structures and corresponding file extensions used in the sequential steps of gel analysis. The GSFs (gel segmentation files) are produced by the spot segmentation of the gel images. The GCFs (gel comparison files) are result in comparing the GSFs using a set of landmark spots to pair spots between gels. The paged CGL data base is constructed by merging the set of GCFs. Note that the arrows indicate direction of data reduction.

program, CGELP for multiple gel analysis. An earlier version of CGELP called CGEL is discussed in [3].

Associated spots and their characteristics can be partitioned by one criterion and then repartitioned as one attempts to 'see around' the data from several perspectives. From our early efforts at gel analysis with the FLICKER system, it became evident that what was required was a system which could automatically find and measure all (or most) spots in a gel. Spots from two or more gels should be comparable which implies that the program needs to be able to partition and to cocatenate lists acquired at different times and from different gels. Without checking *all* or at least most of the spots in the set of two or more gels, no complete statement of the types of spot differences can otherwise be made. These constraints imply both a gel pairing program and a spot data management system.

In a given gel, the majority (if not all) spots, once isolated, can be characterized by (to the first approximation) a triple, comprising x and y position (centroid) and an adjusted integrated density value proportional to polypeptide concentration. Among gels, the idiosyncratic variations of these triples due to variation in gel and sample preparation, detection, etc., confound what are the 'real' variations

produced in the biological/clinical system by time, dose, clinical state, etc. We propose the concept of a *canonical gel or C-gel*, which is valid for the domain of a given experiment or a defined clinical situation. Such a *C*-gel provides information characterizing position and density distributions for all spots over all gels in the set. Further, it excludes the data idiosyncratic to detection and preparative conditions unrelated to the biologic issue. A necessary but not sufficient condition for construction of a *C*-gel is the pair wise comparison of each gel with every other gel in the set, with the condition that comparison be commutative. In other words, if there are n-gels in the set, to construct a canonical gel requires $(n - 1)$ factorial comparisons times the number of spots. Since each element of the C-gel is a function expressing the variation of the spot descriptor triple as a function of the biomedical variable, it is not easily constructed. Though not easily realized in practice, the *C*-gel provides a model object against which we may weigh a pragmatic substitute, the *representative or R-gel.*

The *R*-gel, in contrast to the *C*-gel, is derived from a single pictorial object. It is a real gel chosen from a set of gels representing a given experiment. *R*-gel selection is detailed below, but it may be considered to be what it is named, a representative (by experimenter criteria) gel which is believed to contain most if not all spots encountered in any of the members of the set. It is not necessarily an experimental control gel, but its selection by the biologist certainly reflects his knowledge of the experiment and of the resulting individual gels that constitute the set.

The *R*-gel is used as the basis against which other gels in the set are compared. Each spot in the *R*-gel is the index to a *R*-spot set. An *R-spot set* is that set of spots, with at most one from each gel in the set of gels, which corresponds to a given spot in the *R*-gel. The *list* of *R*-spots under ideal conditions includes all spots in all gels. Until biochemistry can provide essentially noise free gels and extensions to the analysis including methods for handling missing or very noisy spots in the *R*-gel, such a complete and ideal accounting is simply not attainable.

4.1. The general system of analysis

The design philosophy underlying the part of the GELLAB system that deals with multiple gels is the interactive and flexible manipulation of spot data organized by congeneric association. Paired spots and their locations and densities are recorded in a congener-oriented database denoted the 'CGL', which can be searched in a variety of ways. Various representations, numeric, diagrammatic, pictorial, textual or tabular, of this data base or of its derivatives can be rapidly displayed in order that the researcher may quickly grasp patterns and implications. Hypothesis verification is performed by interactive partitioning and testing new representations of segments of the data.

Fundamental to our system of analysis of multiple gels is the concept of congeneric polypeptides which give rise to sets of corresponding spots across gels. A congeneric set of polypeptides is one in which each member arises from a common group of biologic processes. The quantitative expression of such production may be muted or exaggerated under varying experimental conditions. But in each gel where it is

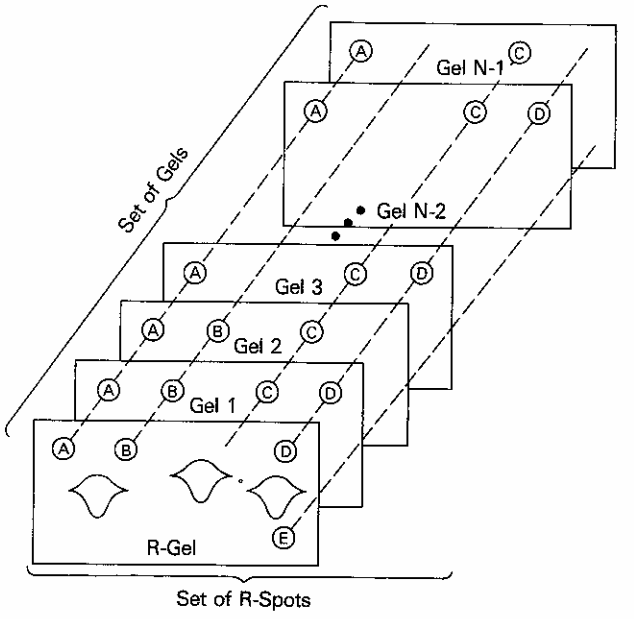


Fig. 12. Spots *A*, *B*, *D* and *E* in the *R*-gel are *R*-spots. *R*-spot set *A* consists of {*Ar*,*A*1,*A*2,...,*An* − 1} which has one member from each gel. Other *R*-spot sets such as *B* and *D* are not represented in all gels. The assumption is that spots which are members of an *R*-spot set are congeneric, i.e., the formation of the polypeptides which they denote represents the action of a chain of biologic processes common to all of them. One possible form of systematic error could occur: Suppose that in gel 3 the spot perceived as B was in fact displaced by severe local preparative distortion, so that its expected position was occupied instead by another polypeptide. This kind of false positive event is largely dependent on the pairing algorithm. The set *C*, is not an *R*-spot set even though represented in all other gels since there is no member in the *R*-gel. The presumption is that the members of set *C* are also congeneric, but whether by chance or biologically determined, the absence of a congener in the *C* position of the *R*-gel keeps the entire set outside of the data base. This problem can be resolved two ways: (a) by peforming experiments involving more than one *R*-gel (constructing other CGL data bases); or (b) extrapolating a spot *C* into the *R*-gel. This latter approach is being investigated by our group.

detected, *the spot denoting the congeneric polypeptide occupies the same relative position in the local gel morphology.*

4.1.1. The R-spot set and the list of R-spots

Formally, we represent these two constructs, the list of R-spots and the R-spot set, as follows:

(1) *A R-spot set* – This is a set of congeneric spots of a particular polypeptide, having at most one member from each gel (but definitely including a particular member of the list of *R*-spots), corresponding to a given spot in the *R*-gel. (Cf. Fig. 12 The B series of spots). The *R*-spot set may be regarded as a vector, each element of which is taken from a single plane of the three-dimensional (3D) stack of gels.

(2) *The list of R-spots sets* – All the distinguishable R-spot sets in the *R*-gel, taken

together, constitute a list of *R*-spots; i.e., all the members of the list of *R*-spots are to be found in the *R*-gel and all the spots visible in the *R*-gel are, at least potentially, members of this list of so called *R*-spots. (See Fig. 12, A,B,C, etc., in Gel *R*.) Spots that compose the list of *R*-spots are to be distinguished from the elements of a particular *R*-spot set.

The linkage and reciprocal dependency between the list of *R*-spots and a *R*-spot set is this: (1) A *R*-spot set member (i.e., a spot in the *R*-gel) must correspond to at least one other spot in the remaining $(n - 1)$ gels for it to be recorded by the CGELP program as a spot pair; and (2) a set of congener spots will not be recorded as a R-spot set if it does not have a representative in the *R*-gel (cf., Fig. 12, series C). If in (1) the spot only exists in the *R*-gel it will be recorded by CGELP. On the other hand, if it exists in a non *R*-gel, then it is not currently recorded. One extension to GELLAB, currently underway, is the use of extrapolated *R*-spots or *eR*-spots which can be inferred from the local morphology and can be used to handle this problem.

A *R*-spot set represents a presumptive empirically derived congeneric set of polypeptides. Figure 12 shows an example of a congeneric set which is not a *R*-spot set (spot set *C*). Since *R*-gels are real objects which are usually incompletely representative of the totality of protein production, it is likely that some congeneric sets will not have representation in the gel chosen to be the *R*-gel. By generating and testing several data bases built on complementary *R*-gels, this problem can be handled at present (until the *eR*-spot GELLAB extension is completed).

4.1.2. Local morphology

We have found that for gel analysis a most effective strategy is to concentrate on sets of local morphologies (both within and across gels) rather than to treating one object at a time. Even if the task is defined as detection of the presence or absence of a single spot, some consideration of local morphology is necessary for any decision to be made by machine or requiring human confirmation.

Recognition and identification (as opposed to detection) is quite difficult because of the absence of fixed shape and size of the individual spot. In dealing with spot identification, we are actually concerned with problems of local morphology, in which we are aided by the machine to (1) establish the proper region of regard, (2) maintain a local coordinate system, and (3) perform pictoral and numeric comparisons.

4.1.3. The congener spot data base

We have discussed procedures that have been preparatory in that they deal with operations on individual spots or spot pairs. After constructing the set of *R*-spot sets, we are now in a position to use these data so as to construct a data base which can be ordered as a function of biological, clinical, experimental or temporal variables. The richness of the data base does not limit us to any one of these as the facilities which we now describe allow a multiplicity of orderings. A variety of representations may be chosen which may best be determined by the nature of the experiment. The biology demands that the analytic process be limited in its 'attention' to the set of congeneric

spots, one from each gel, a process that transcends the constraints of the individual gel. The CGELP data management system permits this type of analysis to be applied successively to the majority of such *R*-spot sets.

The types of operations performed consists of many computational or representational operations on the list of *R*-spot sets or sublists of *R*-spot subsets. The latter subsetting may be automatically accomplished based on an experiment dependent characteristic of a gel (from the accession file), on a statistical property of spot or *R*-spot set features, etc. Alternatively, the user may construct at will a working set of gels taken from the entire set of gels. A wide variety of representations of the data, both image and numeric, is available with many modes of display including superimposition on the original image. Important data structures include:

(1) The set of working gels used to restrict the CGELP operations to a subset of the gels in the data base. Only gels in the working set are used in the computations.
(2) The gel subsets structure which is used to manipulate gel subsets in easily redefining working set or gel classes.
(3) The classification sets which contain the names of the gels in each of up to 9 classes. Thus, the user can, depending on the problem he is dealing with, classify gels by temperature, by disease, by metabolic condition, etc.
(4) A 'search results list' of *R*-spots set number names which were found by one or more of the various available search options (or explicitly defined) is available to many of the CGELP operators.

In dealing with real data, it is frequently necessary to create a working subset of gels taken from the original data base in response to different questions. The same data may be used to analyze different aspects of the same experiment by being partitioned in various ways. A related requirement is the facility to declare classes of gels and to create further subsets based on class membership. As an aid to manipulating subsets of gels, gels may be put into named subsets and treated as an entity.

4.1.4. Solution strategies

The properties that characterize spots, the principle of local morphology and the different objectives of different users require of an analysis system which is capable of varied analysis. Such a system offers the capability of designing a solution strategy or set of strategies rather than a direct and single solution. Among the important tools available for such strategies is the experimenter directed creation of multiple representations of the same data. Many of the system procedures are essentially procedures of presentation which allow the user to alternate between say numeric position or density data and synthetic images.

These tools include *R*-maps and mosaic images which facilitate the backchecking of any *R*-spot set in both a global (the *R*-map) and a local but multiple gel (mosaic) context. A mosaic is an image constructed by concatenating, in a 4×4 array, corresponding subregions from each gel, ordered in the array by spot density, surrounding the spot of interest. The mosaic provides a powerful tool whereby the

user may be assured, on the basis of visual evidence, that a spot belongs to a given *R*-spot set. The *R*-map image provides the link between the global location of an individual spot as seen numeric *R*-spot set data or local mosaic image. The *R*-map is invaluable for rapid evaluation of the validity of spots found to be of interest by GELLAB statistical searches or manual examination of the data base. Mosaics are insufficient for establishing a spot's context because of their locality and thus the *R*-map fills this void. Numeric data, particularly functions of density presented in rank ordered tabular form is useful for evaluating magnitude differences between spots in an *R*-spot set. The gray scale numeric representation of each pixel comprising a spot in a small window of the image is occasionally useful in determining whether a spot is actually one or two or whether a spot was fragmented by the segmenter. The accession file information is always available for use with a data set or its derivatives. Any portion of it may be used as the associative key with which to regroup gels within the gel data base.

Tools such as the foregoing are invoked as needed at user discretion to establish and/or confirm membership in a biologically significant congeneric vector, i.e., a *R*-spot set. Moreover they can be used to quantitate substantive changes as a function of the biologic variable at issue.

In sum, the GELLAB set of programs represents a general method to organize and selectively compress the data of 2D gels so that the user may more efficiently perceive patterns out of the welter of individual spots. Once *R*-spot sets of interest have been found, it is a direct process to quantitate their individual components by merely printing their *R*-spot sets.

4.1.5. Analyzing multiple gels as a continuum

Each congener polypeptide visualized as a spot may be thought of as having a distribution of spot densities when sampled in a set of gels. It is expected that this distribution will cluster multimodally in the case of significant spot density differences according to the biological state of the sample. Therefore, it is important that biologically non-significant variances be controlled and minimized (false positives). Adequate numbers of gel samples must be obtained for the data base to aid in detecting these multimodal distributions.

We must assume that not all spots will be accounted for (false negatives). No automatic procedure can account for the almost infinite variety of image noise found in these gels. The semiautomation of the gel analysis may be sufficient to find spots for those biological problems where the changes are above the noise level and resolvable by the system.

4.2. CGELP spot data base analysis system

The role of the spot data base may be seen in the overview of the entire gel analysis procedure (cf., Fig. 1). The hardware environment mentioned in Section 1.3 is of some interest in understanding the processing and data structure manipulations. The GELLAB system is currently implemented using two hardware systems: the Image Processing Unit's Real Time Picture Processor (RTPP) and a Digital Equipment

Corporation DEC-2020. Figure 11 illustrates the data structures required and generated at the different stages of processing. Image acquisition and landmarking are currently performed using the interactive RTPP system with distributed processor software (2020/RTPP) – GETRTPP and LANDMARK. However, images have been acquired on an Optronics scanner and entered into GELLAB via magtape. Using the primary gel image files, accession file and landmark spot sets file, the secondary spot segmentation, gel spot pairing and CGL data base construction and analysis are performed on the DEC-2020. A cost accounting estimate of the various steps in the complete analysis of an average set of 20 gels found typical DEC-2020 times were about 25 min of CPU time/gel with average 150 K-word memory program sizes and less than 15 min of RTPP real time/gel. Using the manual landmarking approach with programs DWRMAP and LMSEDIT (cf. Section 3.2), much of the RTPP interactive time would be eliminated in cost of several more minutes of 2020 CPU time.

Such an environment imposes some practical limits on the capacity of GELLAB. Gel analysis, as illustrated in Fig. 11, is primarily a series of data reduction steps mapping gel image information into a set (of over 3000) *R*-spot density distributions for up to 128 gels. Images are reduced to spot lists, then spot lists are reduced to spot pair lists and finally, spot pair lists are reduced to a list of *R*-spot sets. Clearly, comparing hundreds to thousands of spots in over 100 gels would be a monumental task if done manually. Most of the computation is involved in the initial image data reduction phase whereas the amount of additional computation is dependent on the type of questions to be asked about a particular set of gels. For example, 60 searches on a 15 gel data base of over 600 *R*-spots with the 15 gels being repartitioned as 2 gel working sets took less than 1.5 h of CPU time.

Procedures used in this later phases of analysis are based on the analytic principles discussed above and carried out by the experimenter using an interactive program called CGELP. CGELP is an interpreter program and is used to construct a representative spot data base and then to analyze all or part of this data base in various ways.

4.2.1. Generation of the R-spot set data base

The first step in the construction of the data base is the generation of the list of *R*-spot sets. The set of $n - 1$ GCFs (Gel Comparison Files, cf., Fig. 11) are read one spot pair at a time for each gel pair where one of the spots is a *R*-gel spot. Each gel pair referenced by a 'key' for the *R*-gel spot, is formed for this pair (cf., Appendix in [3]) and the data base is tested to determine whether a *R*-spot set currently exists for that *R*-gel spot. If it does not, then a new *R*-spot set is created and both spots are put into that set. If it does, then the other spot in the pair is inserted into that set. In either case, the *R*-spot set is initially rank ordered by density, darkest first. Alternate *R*-spot set orderings may then be routinely performed as part of the analysis.

Parts of the CGELP system are illustrated here by examples of operations and results obtained on one of several projects to which it has been applied. We have chosen the work on some of the P388D1 gels for the majority of our illustrations. Table 5 lists the top level CGELP commands. A subset of CGELP commands are

TABLE 5

CGELP top level commands

```
CReate - create a CGL data base from a set of CMPGEL .GCF files.
DDplot - Draw (plot .PLX) LOG %density/%density spot plots from CGL DB.
EDit - Edit spots from the CGL data base.
EXIt - Exit CGELP to the monitor to save paged CGL data base for later use.
EXTrapolate - missing spots in Rsets from mean (dx,dy)+LM position.
Gels - Gels lists the names and total densities of the current gels.
HElp - Print this message.
HIstogram - compute histograms of functions of Rspot sets (.TBL).
Inquire - Interrogate the CGL data base for particular spots.
Limits - print the current statistical limits.
PLot - feature vs. feature plot of 2 (or 3) spot features over CGL DB.
PRotect - paged CGL data base for read-only (toggle).
REOrder - Rspot sets (after c nging density mode).
RMap - plot the Rmap surround  g the Rspot using density estimates.
SAve CGL- Save the CGL data ba  in a (.CGL) file.
SET Accession file name - chan  the default GEL.ID name.
SET Classes - Define gel class  rtition.
SET DAta base file - setup the  w or old CGL data paged data base.
SET DEnsity mode - report result  in Abs (D'), Percent, Ratio or Volume units
SET Fields - Set the list of fiel  desired for gel labeling.
SET Gel subset - define a gel subset.
SET Label - Set the 'Label' code to (S, P, A, U, E) used in searching.
SET RAtio - list of Rspots for normalizing spot densities for Ratio mode.
SET RGel - Set the name of the R-gel used in searching.
SET Statistics limits - Set statistics limits for use in searching.
SET Working gels - Define working set of gels from CGL data base.
SPss - Dump an SPSS .SPS summary file of part of the CGL database.
TAbulate - Print set Rspot dens-rank order, ratio, Mn-variation (.TBL).
TImer - print the run and cpu times for commands (toggle - normally on).
Valid landmarks - list the valid landmarks for each gel in a table.
```

The top level CGELP program commands listed here are available to the user on an interactive basis. The first part of the command that is required for it to be unique is capitalized.

described in [3] while the complete detailed description will be published in the GELLAB user's manual [39].

Table 6 illustrates a simple sequence of CGELP commands which creates a CGL data base and in the process partitions the CGL data base into T0 (initial harvesting) and T24 (cells harvested 24 h later with no media change inbetween) classes. Table 7 illustrates some typical CGL R-spot set data base entries where spots are rank ordered by density. Table 8 illustrates a rank order table for some selected *R*-spots from the P388D1 data.

4.2.2. CGELP commands

The user employs the CGELP interpretive system to analyze a set of gels as determined by successive partition parameter selection. Particularly when dealing with a new data base, the user employs CGELP 'experimentally'. Procedures are invoked. Intermediate results are displayed and examined for confirmation or rejection of the tentative hypothesis, other procedures are then invoked, and so forth. Graphics displays are performed on the user's graphics terminal (if it is a Tektronics 40××-series or DEC GT40). A plotter file may be optionally generated which may be plotted later on a hardcopy plotter or redisplayed on the graphics terminal. The nature of the interaction is highly dependent on the nature of the scientific questions to be asked of the gel data base.

TABLE 6

Example of constructing a CGL data base

```
.RUN CGELP
*SET ACCESSION FILE
        *gelll.id
*SET DATA BASE FILE
        *as5pcg.pcg
*SET FIELDS - to be used in accession file information
        *2,3,10,12,13
*CREATE
        *c20251.gcf
        *c20252.gcf
        .
        .
        .
        *c20258.gcf
        .
        .
        .
        *c20265.gcf
        *
        *spau - (the expected spot labling SP+PP+AP+US)
*VALIDLANDMARKS - list valid landmark statistics
*SET CLASSES
        *auto - class naming based on accession file field information
        *yes - change class names
        *t0 - class 1
        *t24 - class 2
        *
*SET DENSITY MODE
        *ratio
*SET RATIO LIST
        *least squares
*REORDER
*SET GEL SUBSET - define T24 gels as the toxic subset
        *yes
        *toxic
        *250.2,256.2,259.2,260.2,261.2,262.2,263.2,264.2,265.2
*EXIT - save the current data base
.RUN CGELP
*SET DATA BASE - restore saved data base
        *as5pcg.pcg
*GELS - list the gels and numberspots/gel in the data base
*SAVE
        *as5all.cgl - dump all spots in the data base
*SET LABEL - restrict the data base to SP+PP
        *ps
*TABLE - compute and save the inter-gel correlation matrix
        *mn-variation
        *as5mnv.tbl
*INQUIRE - search for the landmarks in the data base
        *landmarks
*SPSS - save the list of landmark spots in an SPSS formated file.
        *as5lms.sps
        ** - '*' indicates the search results list from INQUIRE search
;        Perform a 5% confid level F-test search on classes 1,2
*SET STATISTICS
        *0,512 - relative distance from landmark
        *0,512 - DP (distance between pair of spots)
        *0,512 - DL (distance between LM spot and mean of spot pair)
        *0,500 - mean area of an R-spot set
        *0,1000 - mean density of an R-spot set
        *0,1000 - standard deviation of an R-spot set
        *0,10.0 - std/mean density
        *0,10.0 - spot texture
        *.95 - significance level
        *0  - any number of gels in the data base
*INQUIRE
        *f-test/file
        *as5f95.inq - save search results in file as well
        *1,2
*SPSS - Save search results list in an SPSS formated file
        *as5f95.sps
        ** - use search results list
*HISTOGRAM
        *Set mean Rspot set density
        *tty
        *as5mrd.tbl
*SET WORKING SET
        *define
        *toxic
*SET CLASSES - redefine classes to include toxic and controls
        *yes - set classes manually
        *yes - change class names
        *control
        *toxic
        *
        *0 - 250.2
        *1 - 258.2
        *1 - 259.2
        *2 - 260.2
        *2 - 261.2
        *2 - 262.2
```

```
        *2 - 263.2
        *1 - 264.2
        *1 - 265.2
*SAVE - save the toxic data base in print file
        *as5tox.cgl
*EXIT
```

This example illustrates a typical CGELP command sequence used to construct a normalized CGL data base file (AS6PCG.PCG). After its construction, a *F*-test search is performed to find statistically significant spots. The CGELP commands are given in capitals and the answers to the CGELP prompts are indented and in lower case. The '.' prefix indicates a TOPS-10 monitor command while the '*' indicates a CGELP command. Comments are preceeded by '–'.

TABLE 7

Examples of CGELP data base

```
Output file: AS6NEW.CGL 03/09/1981,  12:20:16 AM
Pairing labels: PSUA
Using least square normalization.
Relative distance limits are[ .00, 512.00]
DL limits are[ .00, 512.00]
DP limits are[ .00, 512.00]
MN area limits are[ .00, 500.00]
MN density limits are[ .00, 1000.00]
S.D. density limits are[ .00, 1000.00]
Coef. variation: S.D./Mean Rset density limits are[ .00, 10.00]
Spot texture limits are[ .00, 10.00]
Class difference t-Test, F-test Rank order significance limit is .90:
Check if # gels in R-spot set [0:1000]
There are 16 gels with 546 Rspot sets consisting of 8239 spots.
Spot free store has 384977 spots available.
   1179 sure pairs,
   4072 possible pairs,
   2873 ambiguous pairs,
   115 unresolved spots.
[1]Total density[0250.2]=4739, # spots=547, Mj= 1.000     bj= 0.000
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,CONTROL,BOTTLE#1
[2]Total density[0251.2]=5638, # spots=569, Mj= .9056     bj= 1.481
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,CONTROL,BOTTLE#2
[3]Total density[0252.2]=3961, # spots=464, Mj= .8651     bj= 5.562
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AL2O3-HC,BOTTLE#3
[4]Total density[0253.2]=3022, # spots=396, Mj= .8989     bj= 6.813
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AL2O3-HC,BOTTLE#4
[5]Total density[0254.2]=3816, # spots=502, Mj= 1.134     bj= 4.716
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AL2O3-18U,BOTTLE#5
[6]Total density[0255.2]=3408, # spots=504, Mj= 1.448     bj= 3.779
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AL2O3-18U,BOTTLE#6
[7]Total density[0256.2]=4687, # spots=528, Mj= 1.006     bj= 4.449
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AMOSITE,BOTTLE#7
[8]Total density[0257.2]=5336, # spots=607, Mj= .8750     bj= 5.672
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T0,AMOSITE,BOTTLE#8
[9]Total density[0258.2]=8627, # spots=672, Mj= .5724     bj= 2.926
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,CONTROL,BOTTLE#9
[10]Total density[0259.2]=8403, # spots=733, Mj= .5616     bj= 4.770
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,CONTROL,BOTTLE#10
[11]Total density[0260.2]=6756, # spots=643, Mj= .8159     bj= 2.771
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AL2O3-HC,TOXIC,PHAGOCYTIC,BOTTLE#11
[12]Total density[0261.2]=7043, # spots=629, Mj= .8454     bj= .9399
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AL2O3-HC,TOXIC,PHAGOCYTIC,BOTTLE#12
[13]Total density[0262.2]=6550, # spots=692, Mj= .8376     bj= 1.733
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AL2O3-18U,PHAGOCYTIC,BOTTLE#13
[14]Total density[0263.2]=11268, # spots=896, Mj= .5539     bj= 3.019
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AL2O3-18U,PHAGOCYTIC,BOTTLE#14
[15]Total density[0264.2]=7001, # spots=674, Mj= .7301     bj= 4.131
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AMOSITE,TOXIC,PHAGOCYTIC,BOTTLE#15
[16]Total density[0265.2]=2387, # spots=355, Mj= 1.672     bj= 5.192
   Study: /P388D1/C14/1 WEEK/ALUMINUM,T24,AMOSITE,TOXIC,PHAGOCYTIC,BOTTLE#16

R-spot[ 34] [0250.2] XYabs=(142,228) MnD= 3.29 SD= 2.06 SD/MnD= .63 #gels=13
 ACC#[Index]C  RDensL area maxOD  D'   Lbl LM  DP  DL  Dx   Dy  Xabs Yabs txtr
------------- ------- ---- -----  ---- --- -- ---- -- ---- ---- ---- ---- ----
0250.2[ 219]1   9.10R   62   .50   9.1  PP B   2.8 35 (-32, -8) (142,228)  .12
0252.2[ 215]1   5.49R   48   .36   6.2  PP B   2.8 35 (-34,-10) (125,244)  .04
0252.2[ 193]1   4.69R   18   .62   5.3  AP B   9.8 40 (-51,  1) (108,255)  .18
0263.2[ 374]2   3.88R   37   .39   7.0  PP B   4.1 33 (-31,-12) (147,193)  .04
0255.2[ 175]1   3.33R   16   .31   2.3  PP B   8.0 36 (-32,-16) (115,220)  .03
0258.2[ 273]2   2.96R   39   .34   5.2  PP B   5.1 35 (-33,-13) (126,225)  .02
0262.2[ 270]2   2.43R   28   .29   2.9  PP B   4.0 34 (-32,-12) (144,229)  .01
0256.2[ 182]1   2.11R   18   .28   2.1  PP B   7.1 34 (-31,-15) (148,241)  .01
0254.2[ 221]1   2.04R   17   .29   1.8  AP B   9.1 42 (-74, 10) ( 67,265)  .02
0254.2[ 224]1   1.82R   15   .26   1.6  PP B   6.1 39 (-38, -9) (103,246)  .01
0259.2[ 314]2   1.74R   31   .28   3.1  PP B   4.1 35 (-33,-12) ( 99,196)  .01
0264.2[ 256]2   1.68R   20   .31   2.3  PP B   4.5 36 (-34,-12) ( 96,219)  .01
0260.2[ 259]2   1.47R   19   .24   1.8  PP B   3.2 35 (-33,-11) (139,229)  .01

R-spot[ 65] [0250.2] XYabs=(177,176) MnD= 75.81 SD= 20.83 SD/MnD= .27 #gels=16
 ACC#[Index]C  RDensL area maxOD  D'   Lbl LM  DP  DL  Dx   Dy  Xabs Yabs txtr
------------- ------- ---- -----  ---- --- -- ---- -- ---- ---- ---- ---- ----
0250.2[  99]1 125.90R  262  1.39 125.9  SP D*   .0  0 (  0,  0) (177,176) 1.51
0252.2[  95]1 108.07R  274  1.30 122.1  SP D*   .0  0 (  0,  0) (163,194) 1.39
0251.2[  95]1 102.15R  266  1.34 112.8  SP D*   .0  0 (  0,  0) ( 78,177) 1.51
0254.2[ 105]1  88.71R  240  1.21  78.2  SP D*   .0  0 (  0,  0) (145,194) 1.21
```

```
0255.2[  79]1  87.34R  200  1.13  60.3  SP D*    .0  0 (  0,  0) (149,176) 1.06
0253.2[  37]1  77.94R  204  1.20  86.7  SP D*    .0  0 (  0,  0) (186,166) 1.14
0257.2[ 104]1  72.89R  199  1.30  83.3  SP D*    .0  0 (  0,  0) (195,151) 1.35
0256.2[  93]1  71.83R  203  1.16  71.4  SP D*    .0  0 (  0,  0) (182,197) 1.10
0265.2[  59]2  69.73R  146  1.10  41.7  SP D*    .0  0 (  0,  0) (151,129)  .92
0260.2[ 146]2  66.58R  205  1.11  81.6  SP D*    .0  0 (  0,  0) (171,180)  .94
0263.2[ 233]2  60.76R  232  1.25 109.7  SP D*    .0  0 (  0,  0) (175,148) 1.21
0261.2[ 103]2  60.62R  184  1.18  71.7  SP D*    .0  0 (  0,  0) (190,157) 1.04
0262.2[ 144]2  59.47R  134  1.24  71.0  SP D*    .0  0 (  0,  0) (175,184) 1.12
0258.2[ 135]2  59.07R  212  1.17 103.2  SP D*    .0  0 (  0,  0) (161,177)  .96
0264.2[ 131]2  53.23R  187  1.17  72.9  SP D*    .0  0 (  0,  0) (129,176) 1.08
0259.2[ 184]2  48.75R  210  1.18  86.8  SP D*    .0  0 (  0,  0) (132,149) 1.08

R-spot[376] [0250.2] XYabs=(307,321) MnD= 1.79 SD= .54 SD/MnD= .30 #gels=13
 ACC#[Index]C  RDensL area maxOD  D'   Lbl LM  DP  DL  Dx   Dy   Xabs Yabs txtr
------------- ------- ---- -----  ---- --- -- ---- -- ---- ---- ---- ---- ----
0258.2[ 501]2    2.98R   36   .32   5.2  SP O   1.4 17 (-12, 12) (296,325)  .01
0263.2[ 616]2    2.66R   24   .35   4.8  SP O   1.4 17 (-14, 10) (298,286)  .01
0265.2[ 296]2    2.17R   18   .23   1.3  AP O   8.1 20 ( -9, 18) (276,284)  .00
0250.2[ 394]1    2.10R   19   .28   2.1  SP O   3.0 19 (-13, 11) (307,321)  .01
0260.2[ 464]2    1.96R   16   .28   2.4  AP O  10.0 18 ( -5, 17) (322,332)  .00
0261.2[ 448]2    1.69R   13   .32   2.0  SP O   3.6 17 (-15,  8) (323,299)  .00
0263.2[ 617]2    1.61R   15   .32   2.9  AP O   3.6 17 (-18,-25) (294,251)  .00
0262.2[ 514]2    1.59R   14   .29   1.9  SP O   1.0 18 (-13, 12) (299,332)  .00
0252.2[ 356]1    1.50R   17   .27   1.7  AP O   8.5 17 ( -5, 14) (307,345)  .01
0253.2[ 252]1    1.35R   13   .25   1.5  PP O   7.1 17 ( -6, 10) (326,316)  .00
0251.2[ 391]1    1.27R   14   .24   1.4  SP O   3.0 19 (-16, 11) (216,322)  .00
0257.2[ 436]1    1.23R   16   .23   1.4  SP O   3.0 19 (-16, 11) (322,296)  .00
0256.2[ 375]1    1.21R   16   .24   1.2  SP O   1.4 17 (-14, 10) (316,345)  .01
```

A CGL data base for P388D1 macrophages, with *R*-gel 250.2, contains 546 *R*-spots. The state of CGELP at the time the CGL DB is saved (with the SAVE command) is illustrated indicating the current partitions. Three examples of *R*-spot sets from this DB are shown: [34], [65] and [376]. Correspondences to *R*-spot [34] and [376] are missing some of the gels. A spot's *RDensL* is its normalized least squares density gel. *Dx* and *Dy* are the spot's position relative to its associated landmark. Metrics (*MnD*,SD) are the mean and standard deviation of the density measurement in the *R*-spot set. Table entry 'C' is the class partition name which in this case has the class partition of 1 = T0 and 2 = T24. *D'* is the background corrected absolute density of the spot while area and maxOD metrics are also recorded. The absolute position of the spot in the gel image is (*Xabs*,*Yabs*). *Lb*1 is the pairing label SP, PP, AP, US or EP. Note that the heuristic pairing values *DP* and *DL* are similar for most spots as are the (*Dx*,*Dy*) relative distances to the landmark spot. Because of this particular consistency, any spot in a *R*-spot set with a large deviation in one of these position features may be regarded as a possible outlier and so treated. *R*-spot [65] is a landmark spot (*D*) (denoted by the * in the *LM* field) with corresponding values of *DP*, *DL*, *Dx* and *Dy* being zero by definition. *R*-spot [65] shows significant class differences in the mean density.

TABLE 8

Rank order density table for selected spots

```
        File: RNKTB1.TBL 02/26/1981,  10:17:23 AM
        RANK-ORDER table: <ACC#>&<LMset>&<Class #>
        Paged CGL data base file: AS6PCG.PCG[61,1]
        Using least square normalization.
        User defined spot list

Density
 125.9 |            0250.2D1
 122.8 |
 119.7 |
 116.6 |
 113.5 |
 110.4 |
 107.2 |            0252.2D1
 104.1 |
 101.0 |            0251.2D1
  97.9 |
  94.8 |
  91.7 |
  88.6 |            0254.2D1
  85.5 |            0255.2D1
  82.4 |
  79.3 |
  76.2 |            0253.2D1
  73.0 |
  69.9 |            0257.2D1
       |            0256.2D1
  66.8 |            0265.2D2
  63.7 |            0260.2D2
  60.6 |            0263.2D2
       |            0261.2D2
```

```
 57.5 |            0262.2D2          0250.2F1
      |            0258.2D2
 54.4 |                               0251.2F1
 51.3 |            0264.2D2
 48.2 |            0259.2D2          0253.2F1
      |                              0254.2F1
      |                              0264.2F2
      |                              0262.2F2
      |                              0256.2F1
 45.1 |                              0255.2F1
 41.9 |                              0252.2F1
      |                              0265.2F2
      |                              0257.2F1
      |                              0263.2F2
 38.8 |                              0258.2F2
      |                              0261.2F2
 35.7 |                              0260.2F2
 32.6 |                              0259.2F2
 29.5 |
 26.4 |
 23.3 |                                          0263.2G2
      |                                          0261.2G2
 20.2 |                    0251.2F1              0260.2G2
      |                                          0262.2G2
 17.1 |  0261.2B2          0250.2F1
      |  0260.2B2
      |  0263.2B2
 14.0 |  0262.2B2          0252.2F1              0264.2G2
      |                                          0259.2G2
      |                                          0265.2G2
      |                                          0255.2G1
      |                                          0250.2G1
      |                                          0254.2G1
      |                                          0258.2G2
      |                                          0252.2G1
 10.9 |  0264.2B2          0256.2F1              0256.2G1
      |  0265.2B2          0257.2F1              0251.2G1
      |                    0262.2F2              0257.2G1
      |                    0261.2F2              0253.2G1
      |                    0255.2F1
  7.7 |  0259.2B2          0263.2F2
      |  0258.2B2          0258.2F2
      |  0254.2B1          0253.2F1
      |                    0264.2F2
      |                    0254.2F1
      |                    0265.2F2
      |                    0260.2F2
      |                    0259.2F2
  4.6 |  0250.2B1
      |  0256.2B1
      |  0251.2B1
      |  0257.2B1
  1.5 |  0252.2B1
      |  0255.2B1
------------------------------------------------------
R-spot |       45        65        98        99       119
Class #  1=T0
Class #  2=T24
```

A rank order density table may be constructed for a small number of selected spots. Five spots were selected from the P388D1 gel data base. Each entry presents three kinds of information; the accession #, the landmark set associated with the spot and finally the class assigned to the spot by the SET CLASS operator.

4.2.3. Data base normalization

Integrel density variation makes some scheme for normalization necessary. The density data initially transmitted to CGELP is already normalized with respect to percent of total gel density. However, this is usually not satisfactory, hence other normalization modes are available. One may normalize the CGELP data base by a subset of well defined 'stable' spots common to all gels or selected for some particular reason. Other normalization methods are also available. Changes in the density

TABLE 9

Example of R-spot set searches

```
a. Landmark set constraint search (LIST LANDMARK subcommand)

        LM[ A ]=R-spot[10]
        LM[ B ]=R-spot[38]
        LM[ C ]=R-spot[58]
        LM[ D ]=R-spot[65]
        .
        .
        .
        LM[ W ]=R-spot[536]


b. T-test constraint search (T-TEST subcommand at .99 significance)

R-spot[ 45] [0250.2] XYabs=(130,266) MnD= 9.28 SD= 5.47 SD/MnD= .59  gels=16
 [45](m1,m2)= 4.72, 13.84, Lim1[ 2.17: 7.27], Lim2[ 10.28: 17.39], m2/m1= 2.93

R-spot[119] [0250.2] XYabs=(318,170) MnD= 16.55 SD= 3.79 SD/MnD= .23  gels=16
 [119](m1,m2)= 13.87, 19.22, Lim1[ 12.50: 15.25], Lim2[ 15.57: 22.87], m2/m1=
1.39
        .
        .
        .


c. F-test constraint search (F-TEST subcommand at .99 significance)

R-spot[ 45] [0250.2] XYabs=(130,266) MnD= 9.28 SD= 5.47 SD/MnD= .59  gels=16
 [45](m1,m2)= 4.72, 13.84, m2/m1= 2.93
 |m2-m1|= 9.12, t(1-a/2)SQRT(v1+v2)= 4.69, f=11

R-spot[ 51] [0250.2] XYabs=(170,202) MnD= 9.18 SD= 4.77 SD/MnD= .52  gels=15
 [51](m1,m2)= 6.02, 12.80, m2/m1= 2.13
 |m2-m1|= 6.79, t(1-a/2)SQRT(v1+v2)= 4.47, f=6
        .
        .
        .


d. Rank order constraint search (RANK ORDER subcommand at .99 significance)

R-spot[ 45] [0250.2] XYabs=(130,266) MnD= 9.28 SD= 5.47 SD/MnD= .59  gels=16
n1=     8 n2=     8 n=    16 R=   100 R'=    36 Ralpha=    43
R-spot[ 51] [0250.2] XYabs=(170,202) MnD= 9.18 SD= 4.77 SD/MnD= .52  gels=15
n1=     7 n2=     8 n=    15 R=    28 R'=    84 Ralpha=    34
        .
        .
        .
```

This table gives examples of search output with four different constraints used in the INQUIRE command linear search: landmark, *t*-test, *F*-test and rank-order test.

normalization method change the rank of the spots in relation to each other in a R-spot set. CGELP permits reordering the data base.

4.2.4. CGELP data base searching and investigation

One of several tests, statistical or otherwise, is performed as a governing condition during execution of a linear search through the CGL data base. The search results list (SRL) is a composite tabulation of *R*-spots selected by the current search meeting all conditions as to gel working set, statistical metrics (see below) and class statistics (see below) where applicable. Table 9 illustrates the results of several types of searches including finding the landmarks, *t*-, *F*-, Wilcoxson–Mann–Whitney rank order tests [28].

For the *t*-, *F*-, and *WMW* rank order-test searches, a histogram of the ratios of the two class mean densities ($m2/m1$) is computed at search completion of spots in the SRL. Table 10 illustrates this ratio histogram for a .90 significance level in the INQUIRE *F*-test search.

A conjunction of statistical limits of *R*-spot set metrics may be used to define the SRL. This can be useful for finding *R*-spot sets which: (1) consist of spots with high or low variance, (2) have primarily dark or primarily light spots, or (3) are complete in having all spots present, etc., *R*-spot set parameters tested include: Relative distance of spot from *R*-spot center, *DL*, *DP*, mean *R*-spot set area and density, standard deviation and coefficient of variation, spot texture and number of gels in the *R*-spot set.

4.2.5. Use of the search results list

As mentioned previously, the search results list is a sublist of *R*-spot sets selected by various CGELP commands and is available either for CGELP further processing or output as SPSS numeric files. SPSS numeric files can be read by the SPSS program [29], MLAB [30], or other statistical analysis packages. Other GELLAB programs (MARKGEL and SEERSPOT – see [3,39]) use the SPSS file as part of their input to generate *R*-maps and mosaic images respectively. Figure 13 shows a typical *R*-map image, produced by the MARKGEL program, of the *R*-gel in one set of P388D1 gels. The spots selected were the result of applying the *F*-test in the search with a .90 significance level. These spots also appear in the 2 class density ratio histogram Table 10. Figure 14a–d shows some typical mosaic images generated from the P388D1 gel data base.

4.3. Some results from a P388D1 data base

GELLAB has been applied to PHA stimulation of lymphocytes [31–32], the effect of asbestos on P388D1 macrophages [33–34], as well as other projects both inside and outside of NIH. We have presented some preliminary results of the effect of time on the P388D1 mouse macrophages in tissue culture here as an aid in understanding the types of questions which may be asked of a gel data base system.

5. DISCUSSION

The GELLAB system for multiple gel analysis has been defined to the point where we can now re-examine system tactics and system problems in the overall biological context (as discussed in Section 1.1).

5.1. System characteristics and limitations

Gel scanning, segmentation and pairing are all finite resolution digital processes and each introduce some error. The computer analysis of a continuous process (for all practical purposes in this case a continuous gel) is performed in a digital space at both finite spatial and finite density digitization.

When multiple gels of split samples are run there is additional variance beyond that due to gel scanning alone. Samples of the tissue cultures of the same material result in multiple gels with an additional source of variation. Sampling of a biological

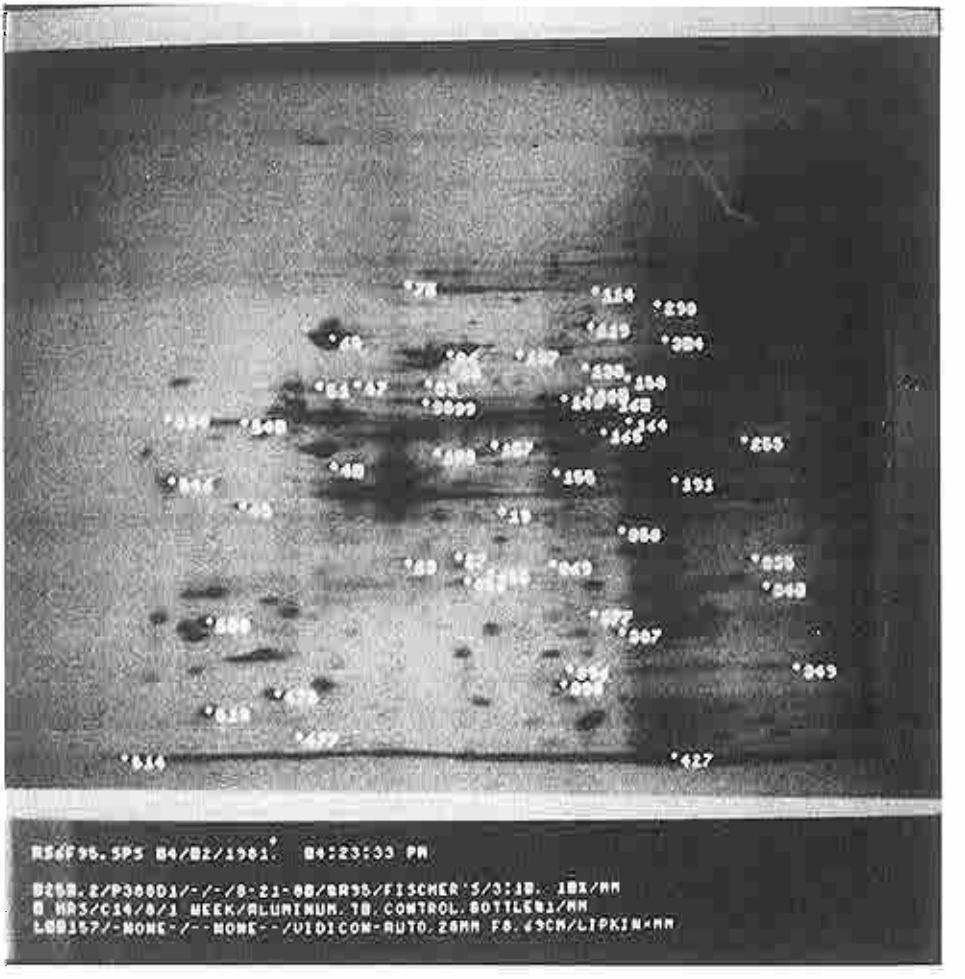


Fig. 13. An *R*-map image of a .95 significance *F*-test search on a least square normalized CGL data base for P388D1 gels was produced using the MARKGEL program. The *R*-map was performed on the *R*-gel (250.2) but could have been performed on any gel in the data base. There are three labeling options: NONUMBER = no label numbers (just the '+' at the centre of the spot); USELANDMARKS = if an *R*-spot is a landmark spot, use the landmark letter name rather than a number for it; the final and default option is to always label with a number.

process at various stages of its progression in synchronized or partially synchronized cultures is also another source of error.

Overall GELLAB system variance was explored using a set of duplicate scans of split sample gels of *E. Coli*. The reproducibility of repeated scans of the same gel at resolutions of 250 microns/pixel was the first test where correlation coefficients between gels ranged .98 to .99. In the case of duplicate gels, correlation coefficients were in the range of .97 to .99. Spots which differed markedly were checked by direct visual examination of the segmented gel image and in some cases, the central core

TABLE 10

Histogram of ratios of density means found with F-test

```
File: AS6F90.INQ 03/09/1981,  12:37:44 AM
Pairing labels: PS
Using least square normalization.
Class #  1(T0)=0250.2, 0251.2, 0252.2, 0253.2, 0254.2, 0255.2, 0256.2,
                0257.2,
Class #  2(T24)=0258.2, 0259.2, 0260.2, 0261.2, 0262.2, 0263.2, 0264.2,
                0265.2,
F-test class search at  .90 significance
Found   80 R-spots,  mean sd/mn= .46
-------------------------------------------------------------------
m2/ml   R-spot sets
  .35 104
  .40 517
  .45  22  71  87 427
  .50
  .55 415
  .60 255 335
  .65  65
  .70  98 191 349
  .75
  .80
  .85  38  99
  .90
        .
        .
        .
 1.15
 1.20 372
 1.25
 1.30  39  92
 1.35  40 107 165 315 398
 1.40  48 119 164 475
 1.45  47 163
 1.50 455 501 508
 1.55  86  88 121 143 175 358 376
 1.60 133 138 247
 1.65 120 122 229 456
 1.70
 1.75  29 166 510
 1.80 137
 1.85 387 476 499 513
 1.90 369
 1.95 158
 2.00  27
 2.05 108 116 160 538 546
 2.10  19 298 396
 2.15  51 304 540
 2.20
 2.25
 2.30 340 477
 2.35  89
 2.40 114
 2.45  41 155
 2.50 502
 2.55 140
 2.60
 2.65
 2.70 377
        .
        .
        .
 2.95  45
        .
        .
        .
 3.25 514
```

```
3.30
        .
        .
        .
4.85
4.90
4.95   70
```

A histogram of ratios of density means found during a F-test (2-class) search at .90 significance is given as an example of this histogram output. The histogram is computed after a 2-class search (for either the *t*-, *F*- or rank order-search). The ratio value is computed as ($m2/m1$) for classes 2 and 1. *R*-Spot set numbers are entered in the histogram to permit easy backchecking to the spots in the data base or *R*-map.

image was checked at the pixel level for spot definition using PIXODT [1]. Problem spots were found to be due to (a) very light small spots, (b) fuzzy spots, and (c) rarely, very light spots in the tail of a very large spot.

The use of higher spatial resolution scanning conditions although advantageous for spot resolution, etc., imposes some burden on the factor of field of view. Dynamic range in the density domain using a Vidicon camera (approximately 0–2.0 OD) is less than that for the class of photomultiplier scanners. For analyzing most spots in most autoradiograph gels, this is not a major problem as the average spot is usually less than about 1.5 OD peak density when care is used to avoid saturation of the autoradiographs. Silver stained gels [35–36] constitute more of a problem as more spots tend to saturate in the dynamic range of the Vidicon. By controlling the silver stain development process, the maximum OD of the gel can be controlled to within a workable range for most spots.

A representative gel (*R*-gel) is used as the approximation to the canonical gel (*C*-gel). As a consequence of this approximation some problems arise which include: mis-pairing a spot because it is poorly defined in the *R*-gel; missing spots which are in other gels but not in the *R*-gel; and noise in the *R*-gel masquerading as true *R*-spots. Spots in the GELLAB data base may be manually edited to correct errors discovered by the user.

False positives may appear in a statistical search of the CGL data base. These can result from incorrect inclusion of one or more noise points in a *R*-spot set, which nevertheless meets statistical criteria. Mosaic and *R*-map images are the major tool for handling such false positives by backchecking. Direct visual examination of the *R*-spot numeric data itself is useful in finding outliers. The false negative spot rate may be decreased by finding additional *R*-spot sets of interest by manually scanning the CGL data base for interesting *R*-spot sets with one or two outlier spots which caused problems with the current statistical tests. It is possible to automatically *ignore* or, alternatively, to *find* *R*-spot sets with outliers since they have a large *R*-spot set coefficient of variation as well as significant differences in other features.

5.2. *Future directions*

Interpreting the constellation of *R*-spots as multiclass distributions facilitates finding subtle shifts in spot quantitation. Viewed in this manner, the expected variance of

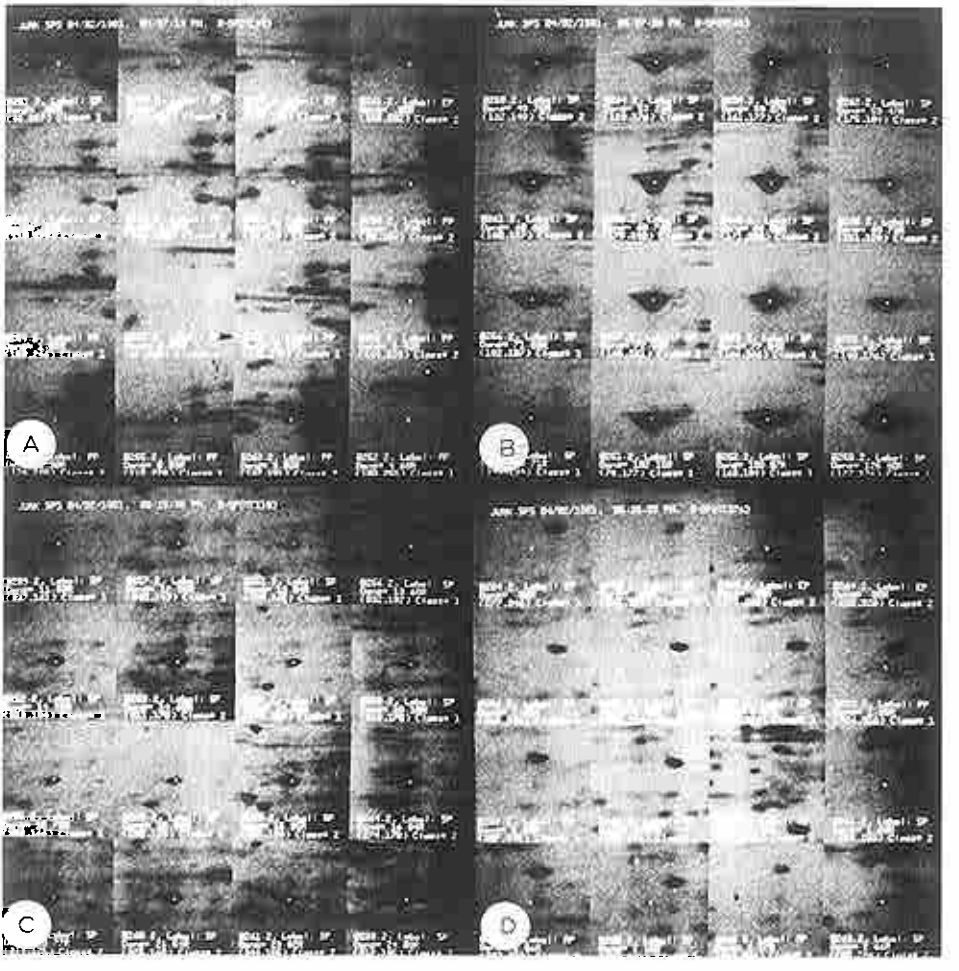


Fig. 14. A mosaic mapping of subregions of a set of gels is performed using the SEERSPOT program on a *R*-spot set of (P388D1 gels). This displays an ordered (lightest to darkest, right to left and top to bottom) illustrate congener spots for a particular *R*-spot. (a) *R*-spot [34] is a poorly defined spot which sometimes appears in a streak and has 5 EPs (extrapolated pair). (b) *R*-Spot [64] has a relative density decrease with time. (c) *R*-Spot [119] has a relative density increase with time. (d) *R*-Spot [376] has a very light spot relative with density increase with time and four EP spots. The *R*-spot region extracted from each corresponding image (magnified by a 2× zoom) is inserted into 128 × 128 pixel subregions of the mosaic image. At the bottom of each panel is an accession number label, gel class and density information. The name of the SPSS file and date is written at the top of the image. The *R*-spot in each panel has a 3 × 3 white '+' drawn in its center. Since the mean density of the images vary – it would be difficult to display and photograph such a mosaic, therefore, the normal mode of operation is to compute the mean background density of each subregion and then to adjust each panel to have the same mean background. The default 2× zoom parameter may be changed to 1×, 4× or 8× by specifying "ZOOM:*n*×".

particular *R*-spot sets can be easily measured and used as a basis for further gel analysis experiments. By having all of the *R*-spot distributions available simultaneously to the data management system, it is now possible to correlate *R*-spot set changes such that sets of *R*-spots changing in the same or in opposition as a function of independent experimental variable can now be determined.

Because of the variety of biological applications and problems in gel preparation, we do not anticipate a fully automated system. However, once a sequence of parameterized operations are identified as habitually used for a class of gels, they can be set up for automated running on a stripped-down system. Such automatic sequencing of operations is now routinely available in GELLAB using the MAKJOB program. It accepts a list of gel accession numbers which are to be analyzed as well as class information. Three batch processing jobs are created to interactively landmark, segment and pair gels, and build an initial CGL data base with several 2-class *F*-test searches. In the P388D1 set of gels, a number of spots were found automatically by the system showing significant statistical differences in the cell line with time. Some of these differences are seen in Figs. 13 and 14, Table 8.

6. CONCLUSION

A set of algorithms in the GELLAB system for the analysis of multiple 2D electrophoretic gels image spot lists using a spot segmentation, spot pairing and congener gel data base analysis has been presented. These algorithms have been successful in analyzing spots under a wide variety of gel conditions. A gel data management system such as CGELP opens the way for asking and answering questions about lists of spot density distributions. Such data reduction applied to a set of gel images has greatly reduced the amount of redundant information retained. Furthermore, by constructing the data base using the inverted file concept, it is possible to rapidly access and update the database. Treating the CGL data base as a set of density distributions leads to the application of various statistical tests for automaticaly determining spot significance which is crucial when investigating a large number of gels with potentially of the order of thousands of spots each.

Significant problems, statistical, operational and others, still remain which must be resolved before reliable reproducible multiple 2D PAGE gel analysis can be routinely performed. That is not to say that useful intermediate results can not be obtained. On the contrary, using backchecking with mosaic and *R*-map images much useful data can be resolved. We are optimistic that many of these problems can be handled by improvements at all levels of gel analysis including better gel preparation, spot extraction and pairing, and the use of better analytical and statistical techniques which take some of these problems into account.

ACKNOWLEDGEMENTS

The constant help afforded by Morton Schultz, Bruce Shapiro, and Earl Smith, our

colleagues in the Image Processing Section has been invaluable. Our collaborators Carl Merril and David Goldman of NIMH and Eric Lester (formerly of NCI, now at University of Chicago Medical School) have provided stimulating ideas and critical evaluation of the methodology as it has developed. At Chicago Eric has done much in our effort to export our system at NCI to a TOPS20 system at another site where we are attempting to eliminate as many problems as possible before making it generally available. Bob Connors of NCI syggested the Kruska-Wallis rank order test for comparing more than two classes of gels simultaneously. Parts of this paper are derived from material which has appeared in part in [1–3,6].

REFERENCES

1. Lemkin, P. and Lipkin, L. (1981) Comp. Biomed. Res. 14, 272.
2. Lemkin, P. and Lipkin, L. (1981) Comp. Biomed. Res. 14, 355.
3. Lemkin, P. and Lipkin, L. (1981) Comp. Biomed. Res.
4. Lipkin, L.E. and Lemkin, P.F. (1980) Clin. Chem. 26, 1403.
5. Lester, E.P., Lemkin, P.F. and Lipkin, L.E. (1981) Anal. Chem. 53, 390A.
6. Lemkin, P.F. and Lipkin, L.E. (1981) In: R. Allen, Arnaud (Eds.) W. Electrophoresis '81, De Gruyter, New York.
7. O'Farrell, P.H. (1975) J. Biol. Chem. 250, 4007.
8. Anderson, N.G. and Anderson, N.L. (1979) Behring Inst. Symposium 1977, Mitt. 63, 169.
9. Lemkin, P., Merril, C., Lipkin, L., Van Keuren, M., Oertel, W., Shapiro, B., Wade, M., Schultz, M. and Smith, E. (1979) Comp. Biomed. Res. 12, 517.
10. Lutin, W.A., Kyle, C.F. and Freeman, J.A. (1978) In: Electrophoresis '78, N. Catsimpoolas (Ed.), Elsevier/North-Holland, Inc., pp. 93–106.
11. Garrels, J.I. (1979) J. Biol. Chem. 254, 7961.
12. Bossinger, J., Miller, M.J., Kiem-Phing, V., Geiduschek, P. and Xuong, N.H. (1979) J. Biol. Chem. 254, 7986.
13. Lemkin, P. (1978) NCI/IP Technical Report #21b, Nat. Tech. Info. Serv. PB278789 (listing PB278790).
14. Lemkin, P. and Lipkin, L. (1980) Comp. Prog. Biomed. 11, 21.
15. Carman, G., Lemkin, P., Lipkin, L., Shapiro, B., Schultz, M. and Kaiser, P. (1974) J. Histochem. Cytochem. 22, 732.
16. Lemkin, P., Carman, G., Lipkin, L., Shapiro, B., Schultz, M., Kaiser, P. (1974) J. Histochem. Cytochem. 22, 725.
17. Lemkin, P., Carman, G., Lipkin, L., Shapiro, B. and Schultz, M. (1977) NCI/IP Technical Report #7a, Nat. Tech. Info. Serv. PB269600/AS.
18. Reiser, J.F. (1976) SAIL, Stanford University Artificial Intelligence Laboratory memo AIM-289, August, 1976. Also available from U.S. Dept. Commerce. Nat. Tech. Inform. Serv. No. Ad-A045-102, Springfield, Va.
19. Merril, C., Switzer, R.C. and Van Keuren, M.L. (1979) Proc. Natl. Acad. Sci. USA, 76, 4335.
20. Merril, C.R., Goldman, D., Sedman, S.A. and Ebert, M.H., (1981) Science 211, 1437–1438.
21. Vo, K-P, Miller, M.J., Geiduschek, E.P., Nielsen, C. and Xuong, N.H. (1981) Anal. Biochem. 112, 258.
22. Rosenfeld, A. (1969) Picture Processing by Computer, Academic Press, New York.
23. Rosenfeld, A. and Kak, A. (1977) Digital Picture Processing, Academic Press, New York.
24. Anderson, N.G., Anderson, N.L. and Tollaksen, S.L. (1979) Clin. Chem. 25, 1199.
25. Taylor, J., Anderson, N.L., Coulter, B.P., Scandora, A.E. and Anderson, N.G. (1979) In: Proceedings of Electrophoresis '79, Radola, B.J. (Ed.),W. de Gruyter, New York.
26. Bookstein, F.L. (1978) The Measurement of Biological Shape and Shape Change, Springer-Verlag, New York.
27. McConkey, E.H. (1979) Anal. Biochem. 96, 39.
28. Natrella, M.G. (1966) Experimental Statistics, NBS Handbook 91, U.S. Govt. Printing Office, Wash., D.C.

29. Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H. (1975) SPSS – statistical package for the Social Sciences, McGraw-Hill, New York.
30. Knott, G.D. (1979) Comp. Prog. Biomed. 10, 271.
31. Lester, E.P., Lemkin, P., Lipkin, L.E. and Cooper, H.L. (1981) J. Immunol. 126, 1428–1434.
32. Lester, E.P., Lemkin, P., Cooper, H.L. and Lipkin, L.E. (1980) Clin. Chem. 26, 1392.
33. Lemkin, P., Lipkin, L., Merril, C. and Shiffrin, S. (1980) Envir. Health. Perspect. 34, 75.
34. Lipkin, L. (1980) Envir. Health. Perspect. 34, 91.
35. Merril, C., Switzer, R.C. and Van Keuren, M.L. (1979) Proc. Natl. Acad. Sci. USA, 76, 4335.
36. Goldman, D., Merril, C.R. and Ebert, M.H. (1980) Clin. Chem. 26, 1317–1322.
37. Schwartz, A.A. and Soha, J.M. (1977) Appl. Opt. 16, 1779–1781.
38. Lipkin, L., Lemkin, P., Shapiro, B. and Sklansky, J. (1979) Comp. Biomed. Res. 12, 279–289.
39. Lemkin, P., GELLAB User Manual, in preparation.

APPENDIX

GELLAB A user's guide to this system will be available from our laboratory. Copies of the software may be available on application to the authors, but this will only be exportable to groups having access to SAIL (on a 'DECSYSTEM-10 or -20).