COMPUTERS AND BIOMEDICAL RESEARCH 14, 355–380 (1981)

GELLAB: A Computer System for 2D Gel Electrophoresis Analysis. II. Pairing Spots

P. F. LEMKIN AND L. E. LIPKIN

Image Processing Section, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205

Received September 8, 1980

An algorithm and computer program CMPGEL, which is part of the GELLAB 2D electrophoretic gel analysis system, is described for pairing comparable spots in two gels. Pairing of spots between two gels at a time is necessary for the comparison of multiple gel images by the construction of multiple gel data bases. Interrogation of and experimentation with the spot data base may then be performed in order to extract measurements on particular spots for the set of gels. The enormous problem of pairing all spots in one gel with all spots in another gel is reduced to a small number of much simpler problems by partitioning spots in the gels into local regions. These are denoted by proximity to so called "landmark spots" and are termed "landmark regions." Spots within landmark regions are then compared in order to perform the pairing.

1. INTRODUCTION

In a previous paper (*1*) in this series (*1, 2*) we have emphasized the need for computer support of 2D gel electrophoresis analysis and described a component of an implemented system for providing this support, a system we call GELLAB. In the first paper, we also treated the problem of spot extraction within a single gel. In this paper 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, such that the (possibly shifted) location of the same spot in both gels is recorded, with reference to one of them. This is a prerequisite to detecting whether individual polypeptides change with respect to experimental conditions, i.e., this pairing of spots within a set of gels taken two at a time is the means whereby a multiple gel data bse is gradually constructed. The articulations of this comprehensive data base into subsets in accordance with the biological problem is a subject we discuss in the last paper of this series (*2*).

The entire GELLAB procedure for tracking spots over a set of gels is presented in Fig. 1 of Ref. (*1*) as a set of single but coordinated sequential tasks. Gels are first acquired, then spots are segmented using the SG2DRV program. This results in a gel segmentation file (GSF) consisting of a list of spot (x, y, density) triples (*1*). One form of segmenter output is a gel segmentation file (GSF) consisting of a list of spots and their features (cf. Table I). Note in the

0010-4809/81/040355-26$02.00/0

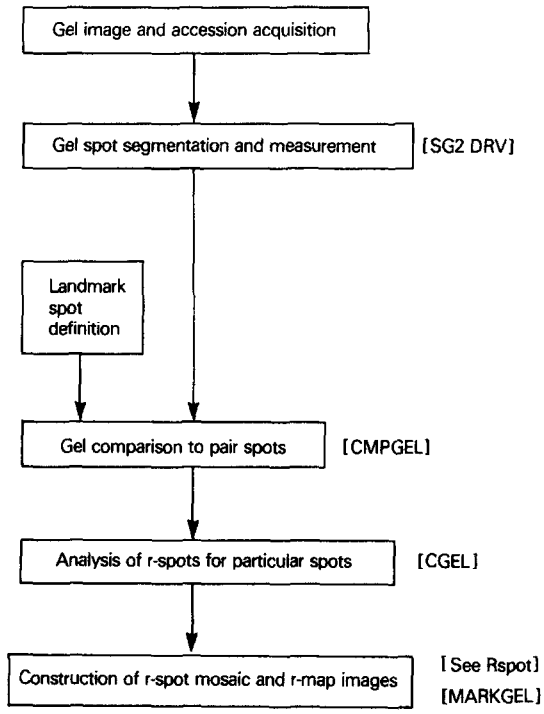


FIG. 1. Block diagram of the 2D-gel analysis GELLAB system. Programs associated with major steps of GELLAB are indicated in square brackets. Gel images are acquired by scanning with a vidicon TV camera interfaced to a picture memory and saved on the computer. Accession information about the set of gels is also used to update an accession file. The gel images are then segmented and measurements made of the spots which are found. Landmark spots, which are either known proteins or well-defined spots spaced fairly evenly throughout the gel, are then manually selected. Using gel image flicker alignment, the landmark spots are aligned for all of the gels with a representative gel (R-gel). This information and the raw segmentation data is then used to pair 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 composite gel data base (CGL). Thus a R-spot set (supposedly) contains the same spot from all the gels in which it occurs.

table that each spot has a spot index (which can be used to refer to the spot), an (x, y) centroid, and a density measurement given in several formats. These include D (total raw density in a spot), D' (D corrected for local background density), $(D'/\text{Total}D')\%$, and volume estimate V based on a Gaussian model of the spot. Any of these can be used in the gel-pairing algorithm, with $(D'/\text{Total}D')\%$ serving as the default density. In terms of the diagram, this paper discusses the second step; it presents an algorithm for spot pairing between two gels using a small set of local landmarks to locally align subregions. The CMPGEL program implements this algorithm and produces a gel comparison file (GCF). Finally, a multiple gel spot data base is constructed and analyzed (*2*).

In the analysis of 2D gels the argument for computer aid does not rest on

TABLE I

EXAMPLE OF GEL SEGMENTATION FILE

```
SG2DRV : Version June 18, 1980 - 1:40PM
Today's date is 07/01/1980, 03:06:20 PM
User: [33,3]
Gel Segmentation File is: P10054.GSF
0054.1/PAT:3/-/-/3-7-79/#5/PHA/3:10,5-20%/
120 HRS/H3/4 HRS/21 HRS/PHA/
E00293/Q011/-NONE-/VIDICON-MAN,28MM,F8,69CM/LESTER*
 57 84 112 135 152 169 183 194 204 217 222 0 0 0 0 0 0 96 490 51 398
Switches: /CTLCORE
Window [96:490,51:398]
Area sizing limits ( 10.00: 2000.00)
Density sizing limits ( .10: 500.00)
Density range sizing limits ( .04: 2.70)
Saving output image in [44,3]Z00293.PIX
Saving central core image in [44,3]C00293.PIX
Mean background matrix in ND (std dev)
 .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .01(+/- .01)
 .00(+/- .00) .00(+/- .00) .01(+/- .01) .02(+/- .02) .00(+/- .00) .00(+/- .00)
 .00(+/- .00) .00(+/- .00) .00(+/- .00) .01(+/- .01) .00(+/- .00) .00(+/- .00)
 .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00)
 .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00) .00(+/- .00)
 CC#   1 M.E.R[ 119: 128,   52:  60] D.R.=[  .25: 1.35] D/A= .879 MnB= .000
  1st MOM[ 123.37, 55.82] A=    51  D= 44.84 D'= 44.84 (D'/totalD')%=  1.41%
  Sx=  2.01 Sy=  2.27 Sxy=  1.67     V= 43.62
 CC#   2 M.E.R[ 127: 133,   52:  54] D.R.=[  .03: 1.24] D/A= .582 MnB= .000
  1st MOM[ 127.74, 53.13] A=    12  D=  6.99 D'=  6.99 (D'/totalD')%=   .22%
  Sx=  1.03 Sy=   .82 Sxy=   .74     V=  7.41
 CC#   3 M.E.R[  97: 105,   53:  61] D.R.=[  .00:  .22] D/A= .076 MnB= .000
  1st MOM[ 102.42, 57.46] A=    48  D=  3.67 D'=  3.67 (D'/totalD')%=   .12%
  Sx=  1.68 Sy=  1.28 Sxy=  1.16     V=  3.29
        .
        .
        .
 CC# 453 M.E.R[ 206: 217,  384: 397] D.R.=[  .00:  .25] D/A= .099 MnB= .000
  1st MOM[ 210.70, 390.27] A=    84  D=  8.33 D'=  8.33 (D'/totalD')%=   .26%
  Sx=  2.23 Sy=  1.73 Sxy=  1.57     V=  6.93
total of  453 accepted D spots accumulated density= 3173.00, area= 16065
total of  453 accepted D' spots accumulated density= 3173.00, area= 16065
total of 7393 omitted spots accumulated density= 675.29, area= 62523
Omitted/Accepted density =   21%
FINISHED! The GSF is P10054.GSF
Real TIME =00:39:06
CPU TIME  =00:14:02,   36.626%
```

Note. Illustration of part of the gel segmentation file for ^{3}H-labeled PHA stimulated lymphocyte gel 54.1 with both parameters and some of the spot feature list data presented.

problem complexity, though the problems are biologically complex, nor on image-processing brilliance, though the algorithms are efficient and appear sufficient for the task. A gel analysis system is necessary because most gels contain a very large number of spots, spots which are at best only locally congruent from gel to gel (*1*), which cannot be counted on to maintain "shape" or optical density and contain little infrastructure on which to build a characterization. There is at best a local congruence between two gels related by some a priori undetermined affine transformation.

In dealing with this material, human factor considerations place a practical limit on the number of spots about which density information can be manually obtained. Manual techniques such as optical flicker comparison between two spots on separate gels is useful for local alignment, especially in cases with

obvious spot differences (*2, 3*). But this method forces the user to deal with the gels sequentially in that only pairwise gel comparisons can be made, making the process time consuming and difficult for the viewer to see a pattern directly over a set of gels. Flicker and analogous 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 indespensable. 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, pictorial, diagrammatic, numerical, etc., that aid the user in seeing patterns difficult to grasp when attention is focused on small regions. Final output of GELLAB includes labeled gel image maps, where statistically interesting spots have been marked as well as numeric spot data lists to support these findings.

Partitioned Search

Aside from the number of spots to be examined, a major problem complicating spot localization is the local distortions in the gel so that neighboring spots in one gel will likely be neighbors in another gel but the intervening distances between them will change to some degree. Several semiautomated methods for aligning corresponding spots in two gels, described as being in various stages of design and use, explicitly deal with this condition.

In one, a gel is transformed locally to the distortions of a second gel (*4*). Three evenly spaced corresponding spots are manually defined for both gels in the region to be transformed. A linear transformation of this region is performed to translate, rotate, and stretch the image locally. After the transformation, spots are matched in this local region using a least-squares fitting procedure counting those pairs less than a specified distance apart in the two gels. This procedure is used to successively partition the gel space with a set of manually defined triangular regions and thus is able to pair a large number of the spots in the gels.

A gel image analysis system (*5, 6*) is being used to develop protein maps automatically and interactively on a wide variety of human biological materials. An interactive mode using a color display allows comparison of spots between gels. Geometric correction is done locally using a linear interpolation in localized regions of a pair of gels. The corrected images can then be used to produce protein maps containing position information as well as the amount of polypeptide for the local regions. These maps can then be compared between gels. Rectangularly defined subregions of a gel can be investigated independently (*6*) and serve as a basis which to discuss a subset of spots.

In another system (*7*), spots from separate gels are brought into alignment by the operator by first manually bringing 15–20 pairs of spots into alignment. Then, the alignment vectors for each of these localities are used to shift one

pattern into register with the other and the images are then replotted. Measurements of the same spots in different gels is done with operator assistance in denoting the spot of interest.

2. A Landmark-Driven Spot-Pairing Algorithm

We present in this paper, an alternative view of the primary pairing algorithm. This involves the construction of a projected image composed from the two members of the gel pair. In the actual matching, the computations are performed in a single plane—the representative gel plane. A central notion is that of the establishment of landmark spots that serve to "anchor" the other spots in its vicinity. Essentially, landmark spots are aligned by the user and then the machine automatically aligns all other spots with the corresponding spots in the other gel. The procedure is simple and, as we will see in the next paper, has been successfully extended to align a multiple set of gels.

A set of landmark spots increases the efficiency of intergel spot matching by providing an empirical basis for the partitioning of a gel image into tractable corresponding subregions.

A landmark spot may be defined in various ways. In one current empirical procedure for choosing landmarks, it is a morphologically distinctive spot such that neighboring spots and the landmark spot form a consistent morphological structure. An alternative definition of the landmark spot might be to double label the gels such that known protein standards have one radioactive isotope label and the sample being analyzed another. Moreover, this structure should be easily recognized across the set of gels. The landmark spots are selected to cover the regions of interest of the gel fairly evenly. This has the advantage of preventing undue biases in the labeling toward some regions of the gel compared to others. From 10 to 25 landmarks are generally established depending on the quality of the gel with fewer required for the better gels. These are called the landmark set. The operator aligns the landmark spots in the two gel images using the FLICKER system, which permits him to move one of the gels while keeping the other constant. Viewing time for each of the images may be independently set and varied until the user is satisfied he has "superimposed" the two images of the same spot. He notifies the machine and then processes the next landmark spot in the same fashion. Once the landmarks are established, alignment of nearby spots is automatic (see (*3*) for discussion of flicker alignment of gels as well as the Appendix).

The landmark region is a 2D space surrounding a landmark spot. It is defined as a fuzzy region having more certainty closer in toward the landmark spot. The half-radius of certainty Ri for landmark i is a distance defined to be half the distance from landmark i to the nearest neighbor landmark. Figure 2 illustrates the half-radius concept. Images of a spot within the half-radius of a landmark set would have a higher probability of being aligned (since the landmarks have "perfect" alignment) than if the spot were outside of this radius. Thus, using this heuristic concept of partitioning the spots by landmark region, it is possible

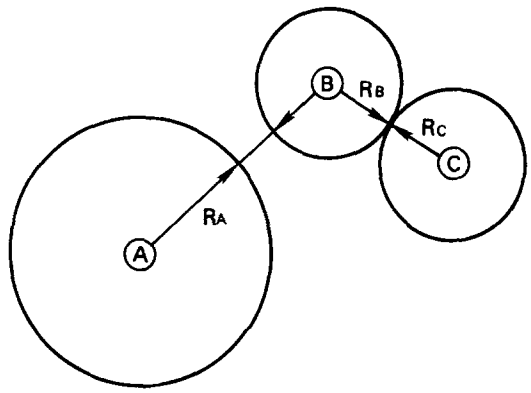


FIG. 2. Definition of landmark half-radius of certainty Ri and nearest-neighbor landmarks. Ri is $\frac{1}{2}$ the minimum distance from landmark i to its nearest adjacent landmark j. In this example, radius $Ra > Rb$ and $Rb = Rc$. The nearest-neighbor of landmark A is B and its next-nearest-neighbor landmark is C.

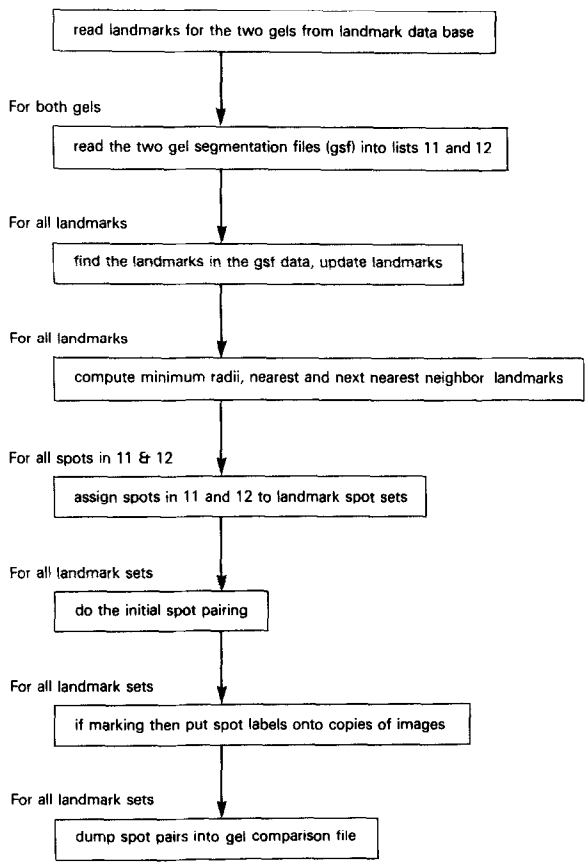


FIG. 3. Block diagram of the 2D-gel comparison procedure.

to pair the spots automatically once the landmarks are established. The landmark spots are then compared with the two GSF spot lists and the best segmented spot 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), that landmark is not used in the pairing process. It is possible to ascertain how reliable the particular pairing actually was by back-checking paired spots in the labeled images.

The probability of finding the same spot in two gels relative to the aligned images of a close by landmark is greater than if the entire gel spot space were to be searched. This partitioned search has the added advantage that landmark regions contain an order of magnitude fewer spots than the total gel space. Thus the combinatorics of performing the spot matching is greatly decreased as well.

Implementation of Landmark-Oriented Spot Pairing between Two Gels

The spot-pairing algorithm is illustrated in flowchart form in Fig. 3. It is implemented as the CMPGEL program in the SAIL programming language (*8*) for a DECSYSTEM-10 or DECSYSTEM-20 computer. The actual pairing is performed in two passes through the landmark sets data, called the primary and secondary pairing procedures. Each procedure operates on one landmark set at a time.

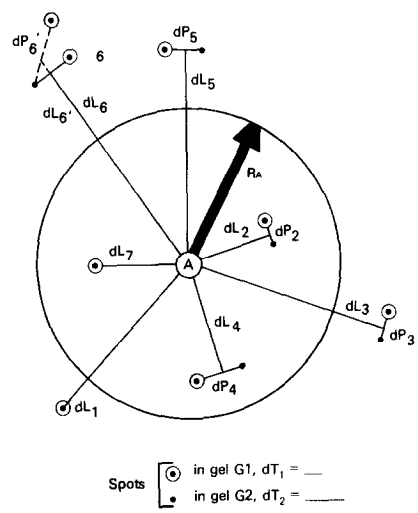


FIG. 4. Spot pair primary labeling assignment definitions. Each potential nearest neighbor spot pair in a landmark set has one of four labels: SP—sure pair, PP—possible pair, AP—ambiguous pair, US—unresolved spot. The labeling cases are defined by the following cases: (1) US—unresolved spot (no dP); (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).

In the primary pairing algorithm (Fig. 4), the spots are first mapped to the Cartesian coordinate system defined by making the landmark spot (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 in the projected image of G2. Because of possible asymmetry of the two sets the reverse comparison is also performed so that each spot in G2 is provisionally paired with its nearest-neighbor spot in G1. The nearest-neighbor distance is called dP (pair distance). The distance dL is the distance from the landmark spot to the mean locus of the two spots in the provisional pair. Two parameter distances are empirically defined: dT1 and dT2. Spots closer than dT1 are relatively well paired. Spots greater than dT2 are very poorly paired and possibly should not be paired. The working values of dT1 and dT2 (5 and 10 pixels, respectively) were determined empirically, by examination of the nearest-neighbor values of

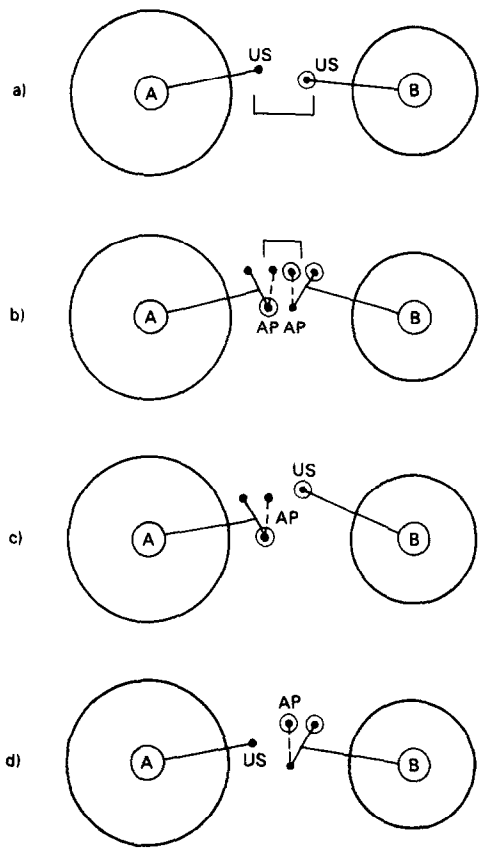


FIG. 5. Secondary spot pairing can be used to further resolve AP and US labels in adjacent landmark sets into SP or PP labels which are then placed in either of the two sets: (a) two unresolved spots, (b) two ambiguous pairs, (c–d) one ambiguous spot and one resolved spot. The new pair is put into whichever landmark set has the smallest dL for the potential pair.

several sets of paired gels under a wide variety of conditions. Figure 4 shows various cases which can occur. Four types of pairing labels can be assigned. There are sure pair "SP," possible pair "PP," ambiguous pair "AP," and unresolved spot "US." The primary spot pair labeling assignments are defined in Fig. 4.

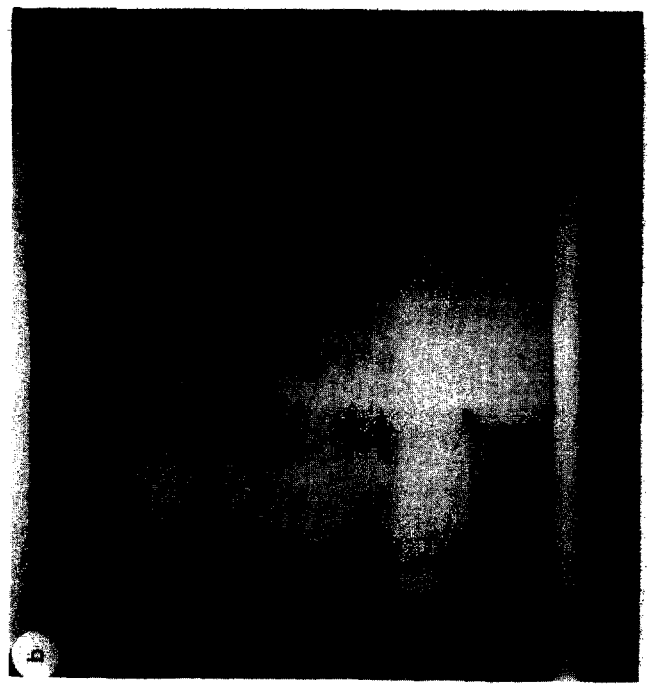
The primary pairing algorithm is a simple first-order model. Consequently some spots on the periphery of the landmark region may be misclassified as an AP or US whereas they be a SP and PP classification in another adjacent landmark region. Figure 5 illustrates these cases. To correct these few misclassifications, the 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 their threshold criteria) is then placed in the landmark set with the smallest dL.

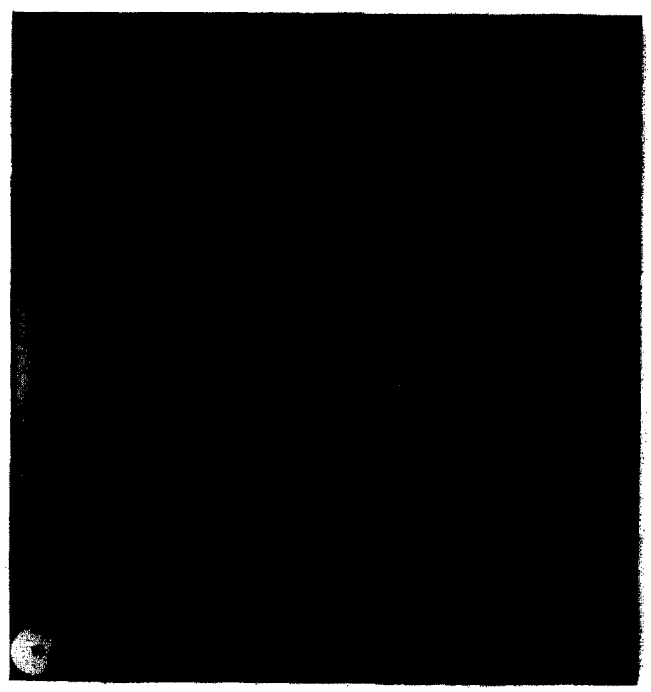
CMPGEL Output

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

3. RESULTS

The spot-pairing algorithm just described has been in use over the past year and has been applied to over 300 gels consisting of both autoradiograph and silver stained gels. Two pairs of gels will be presented here to illustrate the algorithm at the various stages of processing. Further discussion of the results of spot pairing is in the final paper of this series which deals with multiple gel data bases (*2*). Figures 6a,b (7a,b) show a pair of ^{3}H (^{35}S) labeled lymphocyte gels from the same normal patient (*9, 10*). Figure 6 has 250 μm/pixel resolution and Fig. 7 has 170 μm/pixel. In Fig. 6, gel 54.1 had PHA added to stimulate growth while gel 36.1 is a control. In Fig. 7, gel 102.2 had PHA added to stimulate growth while 103.2 was a control. Autoradiograph exposure was therefore correspondingly longer for the non-PHA gels to compensate for a lower rate of synthesis of radiolabeled proteins. Figures 6c,d and 7c,d show the segmented spot images. Gel 36.1 has 207 spots while gel 54.1 has 457. Gel 102.2 has 1081 spots and 103.2 has 1035.





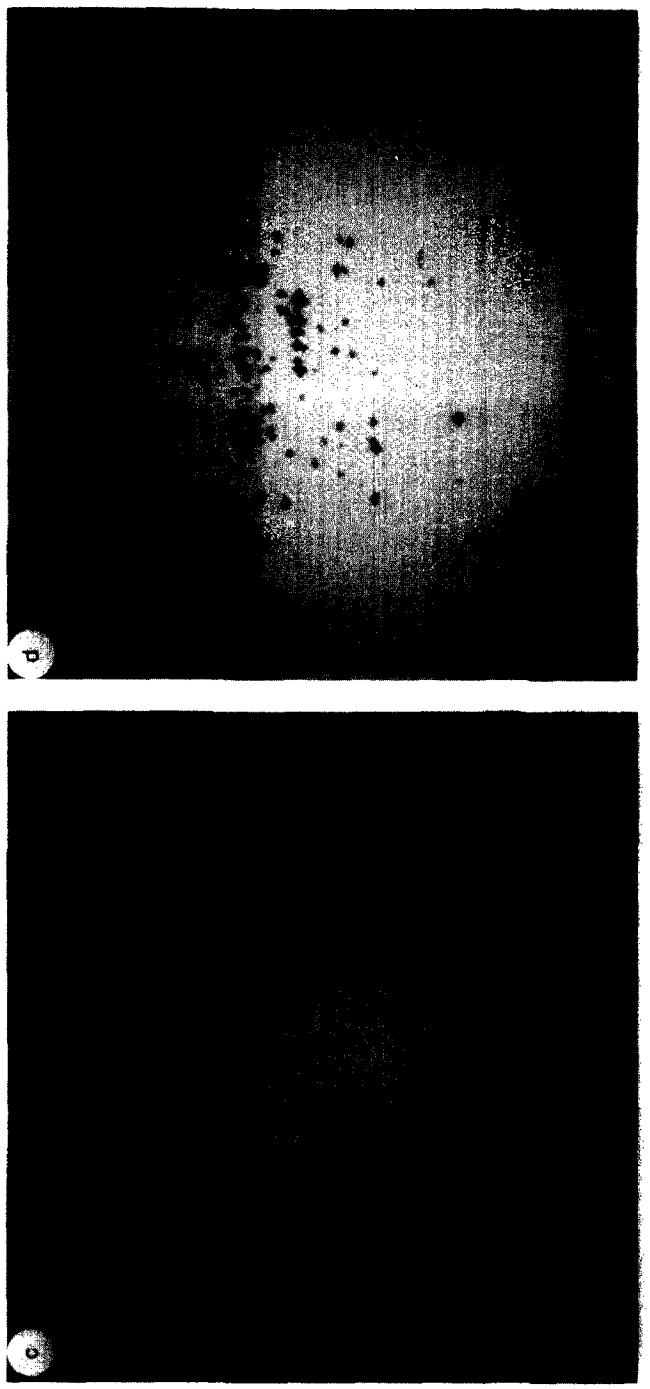
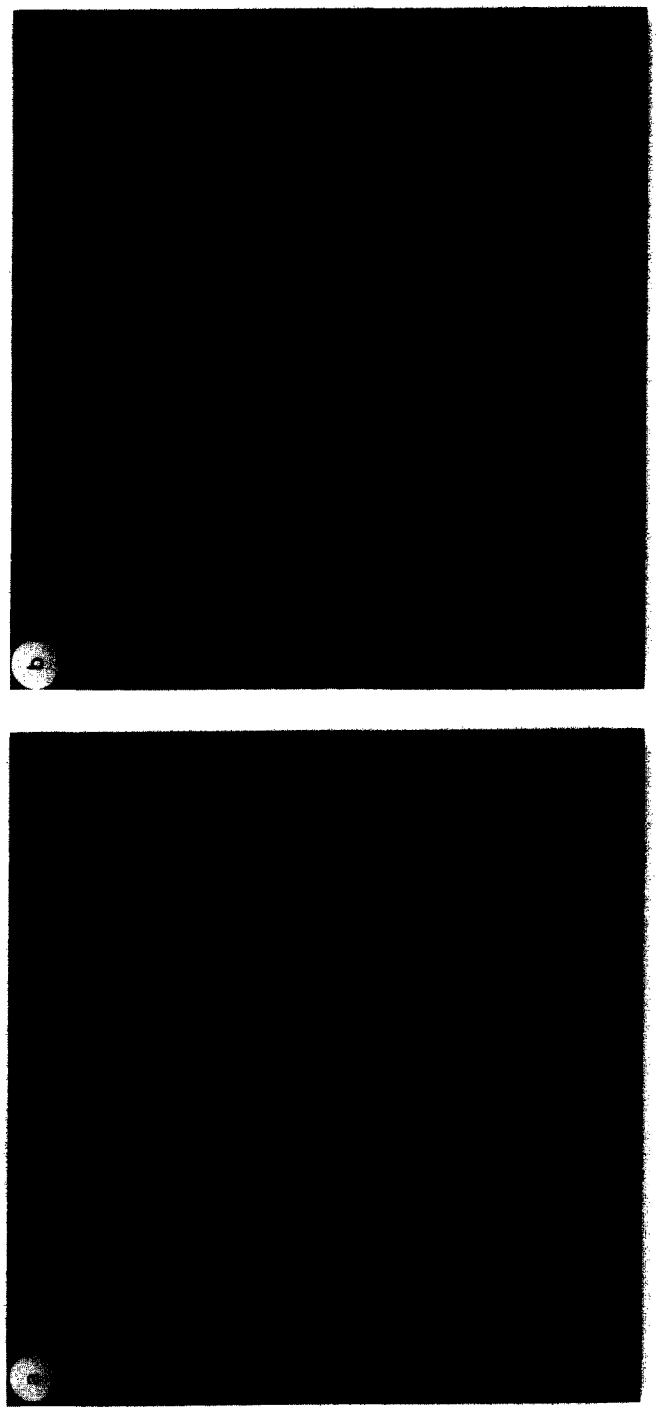


FIG. 6. Original (a,b) and segmented spot images (c,d) scanned at 250 μm/pixel. Images (a,c) are the lymphocyte PHA stimulated gel 54.1 and images b,d) are the resting gel 36.1.



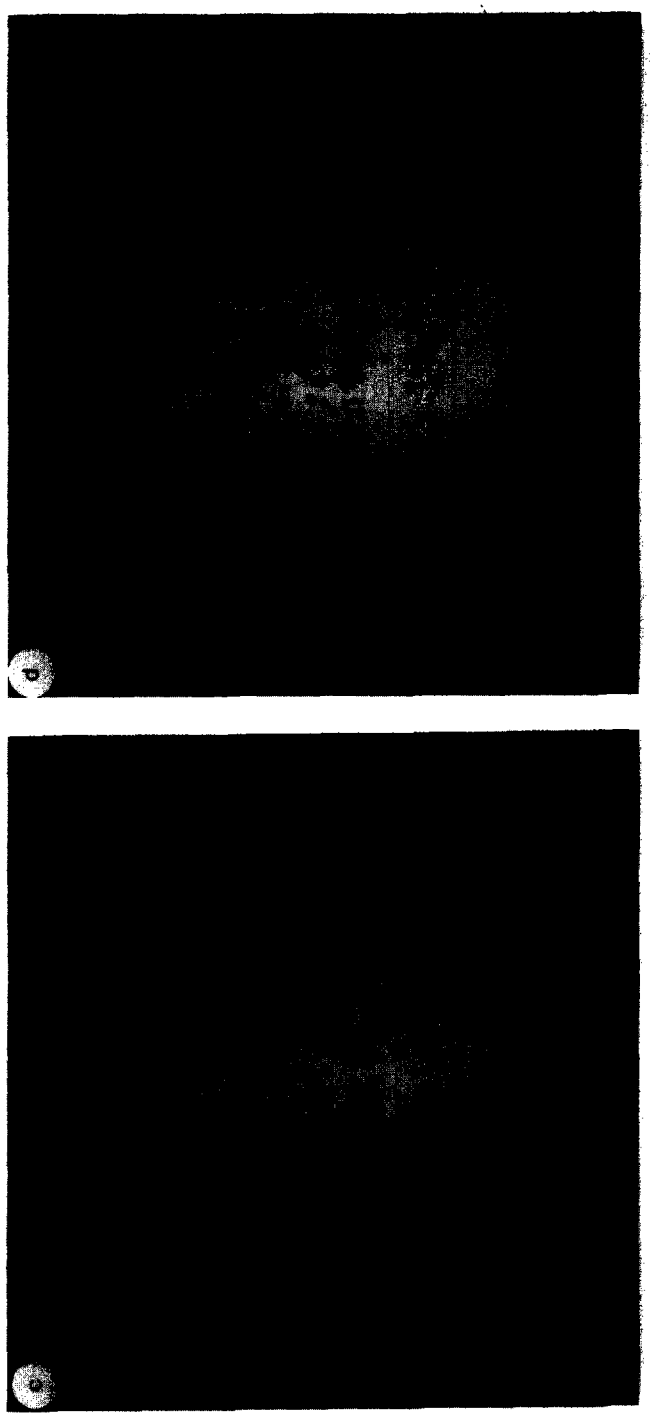
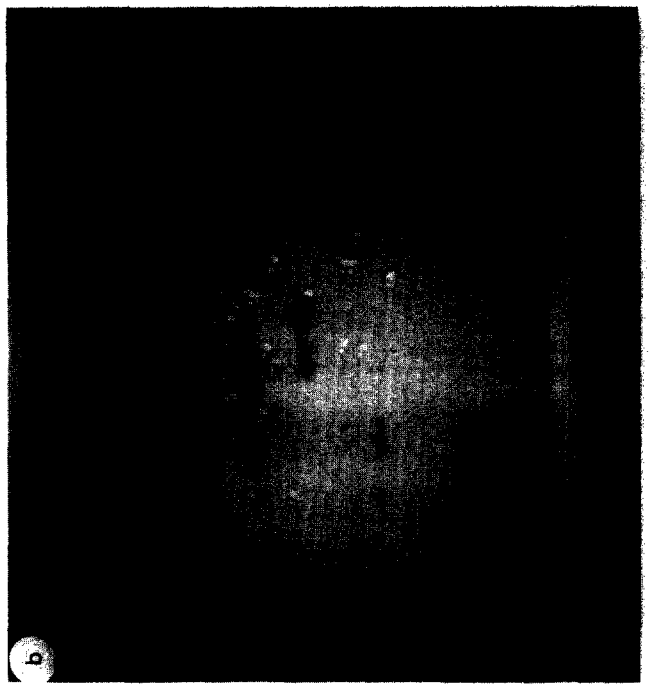
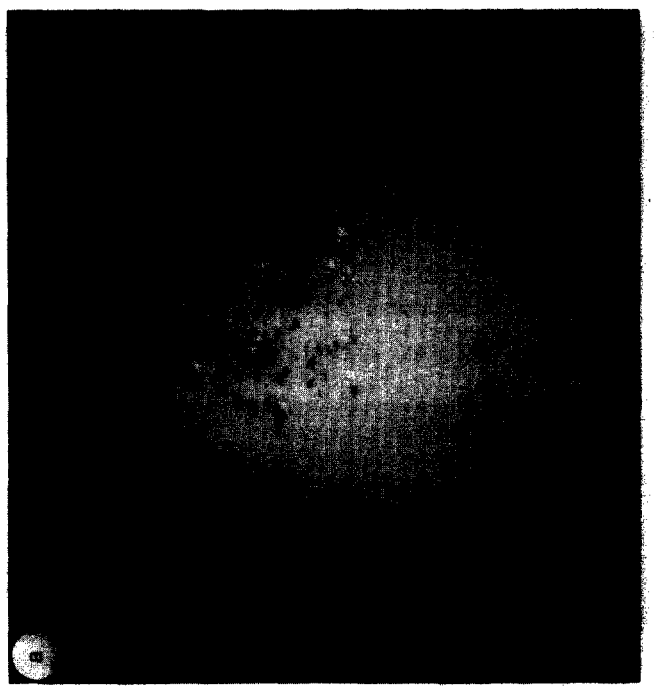


FIG. 7. Original (a,b) and segmented spot images (c,d) scanned at 170 μm/pixel. Images (a,c) are the lymphocyte PHA stimulated gel 102.2 and images (b,d) are the resting gel 103.2.


b


c

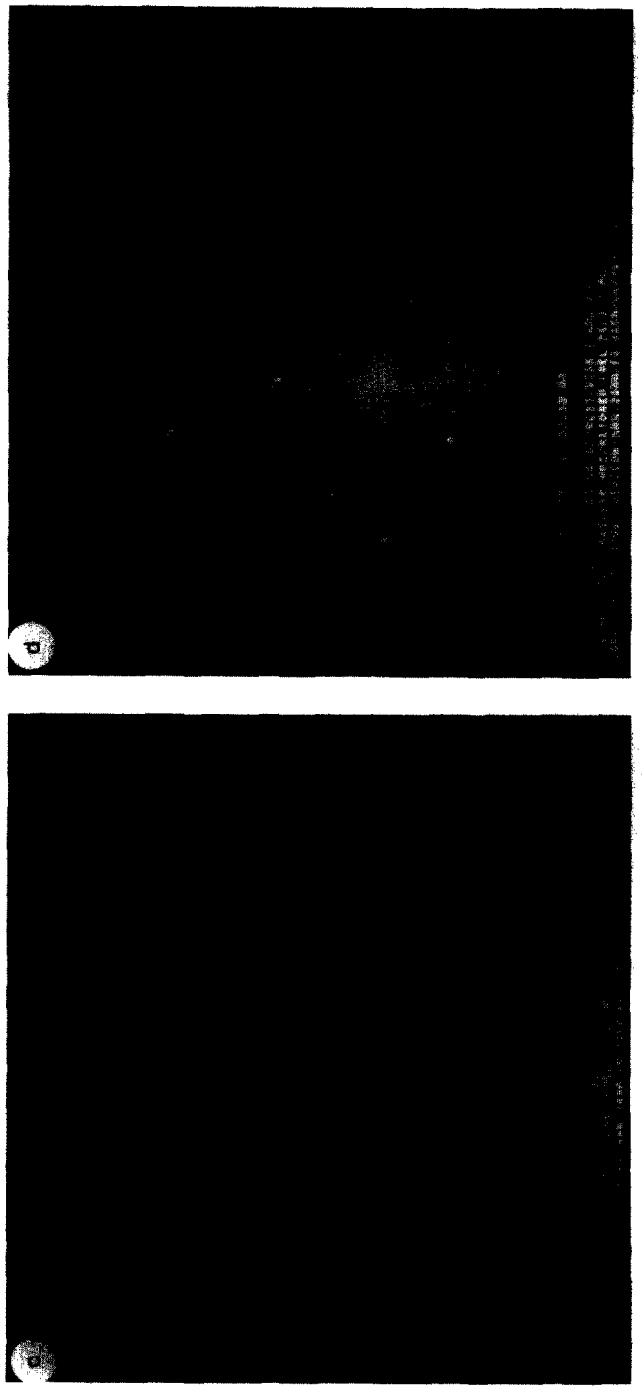
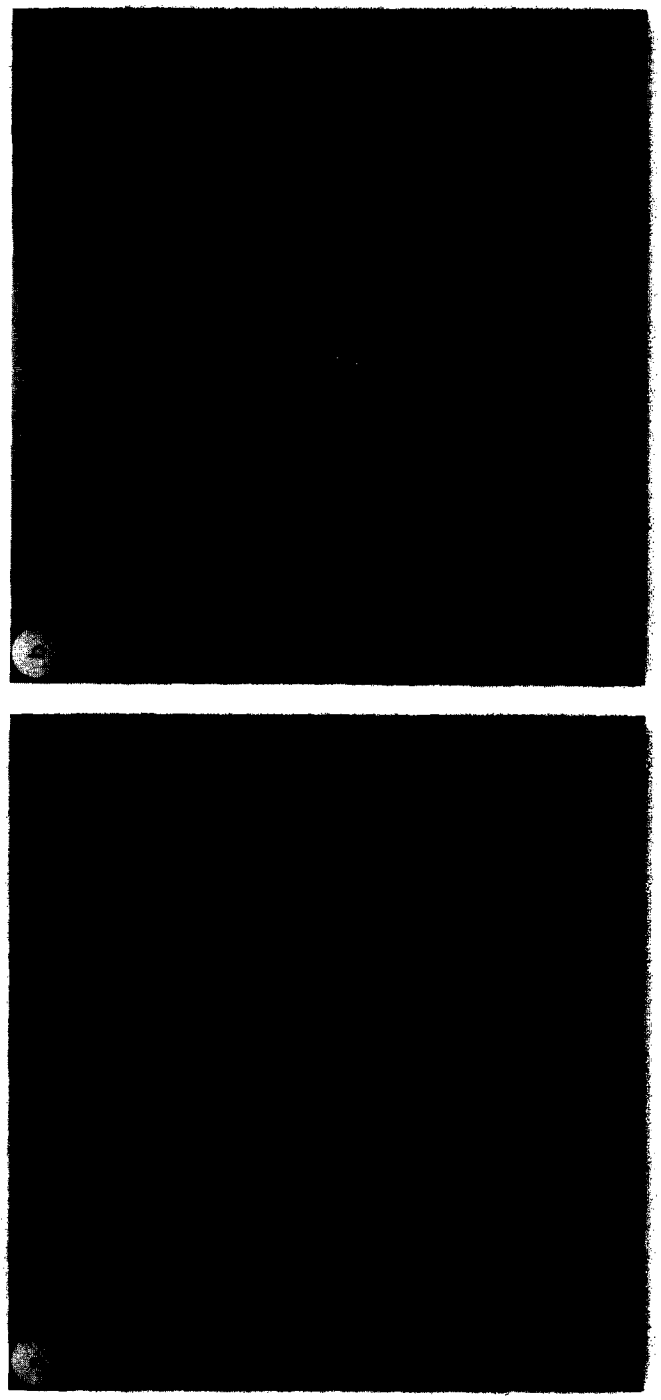


FIG. 8. Landmarks used in the two pairs of gels. The landmark is defined at the small "+" sign with its landmark set name to the right: (a) gel 54.1, (b) gel 36.1, (c) gel 102.2, (d) gel 103.2



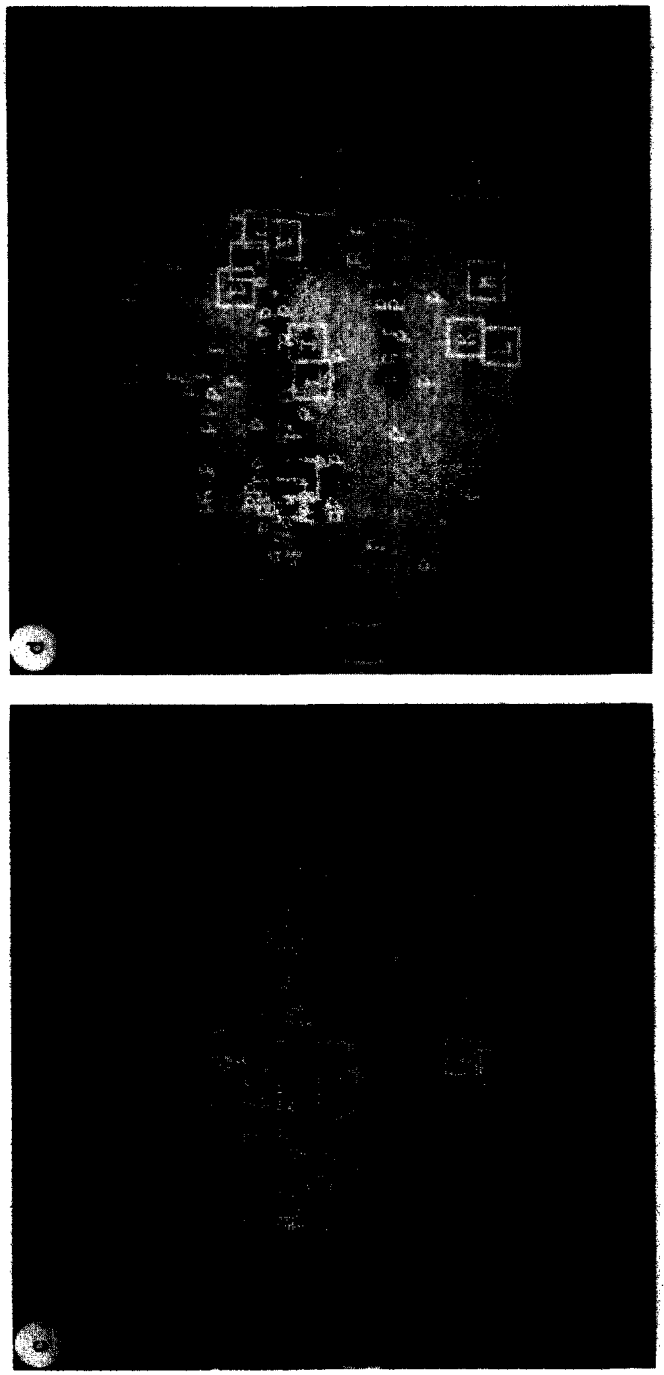


FIG. 9. CMPGEL labeled images with pair labels: S—sure pair, P—possible pair, A—ambiguous pair, '+'—unresolved pair. Gels 54.1 and 36.1 are labeled in (a,b) a 2× magnified central region of these images is in (c,d).

TABLE II

EXAMPLE OF LMS-LM LANDMARK SPOT FILE

```
                              .
                              .
                              .
/ CMPGEL: VER  7/26/79 - 3:04PM
/ INTO SYS@:LM0012.DA FROM GSF FILES: P10054.DA AND P10036.DA
/ SURE PAIR THRESHOLD=  5, POSSIBLE PAIR THRESHOLD=  10
 SEP 21, 1978, TIME 11:52:21 AM
 LANDMARK  A  G1[353, 229], G2[285, 235]
 LANDMARK  B  G1[330, 181], G2[268, 175]
 LANDMARK  C  G1[340, 191], G2[278, 179]
 LANDMARK  D  G1[356, 185], G2[290, 183]
 LANDMARK  E  G1[353, 195], G2[286, 192]
 LANDMARK  F  G1[252, 199], G2[197, 200]
 LANDMARK  G  G1[223, 148], G2[160, 145]
 LANDMARK  H  G1[200, 217], G2[139, 224]
 LANDMARK  I  G1[290, 206], G2[234, 206]
 LANDMARK  J  G1[305, 205], G2[249, 202]
 LANDMARK  K  G1[306, 258], G2[251, 264]
 LANDMARK  L  G1[302, 271], G2[245, 277]
 LANDMARK  M  G1[333, 266], G2[273, 272]
 LANDMARK  N  G1[365, 301], G2[302, 302]
 LANDMARK  O  G1[389, 313], G2[326, 319]
 LANDMARK  P  G1[394, 271], G2[331, 275]
 LANDMARK  Q  G1[408, 250], G2[345, 254]
 LANDMARK  R  G1[395, 232], G2[332, 236]
 LANDMARK  S  G1[376, 259], G2[313, 265]
 LANDMARK  T  G1[363, 266], G2[300, 272]
 LANDMARK  U  G1[375, 211], G2[313, 208]
 LANDMARK  V  G1[365, 208], G2[303, 205]
 LANDMARK  W  G1[243, 194], G2[188, 196]
 LANDMARK  X  G1[224, 272], G2[169, 274]
 ELAPSED TIME:   731. SEC, OR  12.18 MIN.
                              .
                              .
                              .
```

Note. An example of a typical entry in the landmark data base file. The set of spots in this entry is accessed by gel name content. That is, by the gel name pairs (54.1,36.1) or (36.1,54.1). G1 corresponds to gel 54.1 and G2 to 36.1. The G1 $[x,y]$ 2-tuples are the positions of the spots in gel i for the specified landmark. The elapsed time is the time spent by the user interactively landmarking spots and varies from about 3 to 30 min for very difficult gels with the average time being between 5 to 7 min.

Figure 8 shows the landmark spots for the two pairs of gels with each of these spots marked with a small plus sign and the landmark name to its right. Table II illustrates a typical landmark set entry for the 54.1/36.1 pair of gels.

After the spot labels are 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 image as "S," "P," and "A" while US appears as a small "+" (because using the larger "U"symbol would crowd the image where it noisy). Figure 9 shows the labeled pairs for the lymphocyte gel pairs at 250 μm/pixel, where landmark spots in these images have a box around them. Tables IIIa, b show part of the output of a typical GCF for gels 54.1 and 36.1. Table IIIa illustrates the landmark registration and parameters while IIIb shows some representative paired spots as well as pairing statistics.

Secondary pairing for gels 54.1/36.1 increased the (SP + PP)/total spots percentage from 61.5 to 66.1%. For gels 102.2/103.2 it increased from 60.1 to 61.3%. In general a 1.5 to 3% increase in (SP + PP) labeling is observed. This

seems to indicate that most pairing is performed (as expected) during the primary pairing phase of the algorithm. When two very different gels (such as 54.1 with 457 spots and 36.1 with 207 spots) are compared it is to be expected that a fairly high number of US and AP spot labels would result. For more similar gels, the (SP + PP) ratio is in the range of 70 to 85% depending on the artifactual noise of the gels.

4. Discussion

Global Gel Matching

The global matching of two gels would involve a D'arcy Thompson-type transformation (*11*) of one of the gels to bring it into congruence with the other. In this transformation the set of points on the grid are displaced by a continuously differentiable 2D distortion function. In the original gel the corresponding physical distortion is caused, by, among other factors, inhomogeneties in the ampholine/acrylamide mixtures when the gels are run. One approach to gel analysis is to remove the distortion (as is currently done with satellite images) and then perform the point-by-point comparison (*12*). This correction implies some knowledge of the inverse distortion function that is by and large lacking for 2D gels. Moreover, the amount of computation required to perform the full D'arcy Thompson image transformation is considerably greater than for simply pairing spots as performed by our algorithm (which may be thought of as sublandmark registration). Therefore, since the actual gel analysis requirement is to pair spots for comparison, it is more efficient to simply pair local corresponding spots rather than to transform the gel images themselves and then compare the spots in the images.

The pairing algorithm has an additional advantage—i.e., it is also applicable to automatically defined landmarks, should they become available. One interesting technique that has been proposed (*13*) to circumvent the gel distortion problem is double labeling. If two gels are congruent, then spot pairing is greatly simplified. One gel sample is labeled with ^{3}H and the other with ^{14}C-labeled amino acides when the samples are grown in deficient media. The two samples are added together just prior to running the gel and then two-step autoradiography is performed for the ^{14}C- and ^{3}H-labeled gel. This technique has the advantage of forcing perfect alignment between the two autoradiographs although the screen film will have some of the ^{14}C exposure as well. This technique unfortunately can only be used with material which can be radioactively double labeled. Fluorescent labeling of known spots could play an analogous role for stained gel images.

Impact of Landmark Selection on Later Stages of Processing

The partition of the plane into variable-sized landmark regions is based essentially on the local spread of landmarks. A priori one would think that pairing in regions of high landmark concentration would be more likely to be

TABLE IIIa

EXAMPLES OF GEL COMPARISON FILE PARAMETERS

```
CMPGEL[50,32]: Version June 26, 1980 - 1:50PM
Today's date is 07/01/1980,  03:44:40 PM
User: [33,3]
Gel Comparison File is: C10036.GCF from P10054.GSF and P10036.GSF
0054.1/PAT:3/-/-/3-7-79/#5/PHA/3:10,5-20%/
120 HRS/H3/4 HRS/21 HRS/PHA/
E00293/Q011/-NONE-/VIDICON-MAN,28MM,F8,69CM/LESTER*
 57 84 112 135 152 169 183 194 204 217 222 0 0 0 0 0 0 96 490 51 398
0036.1/PAT:3/?/?/1-17-79/#1/RESTING/3:10, 5-20%/
120 HRS/H3/4 HRS/140 HRS/PHA/
E00185/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM/LESTER*
 34 62 86 109 128 146 163 176 188 197 204 215 0 0 0 0 45 405 86 367
Distance sizing limits (dP1 =  5.00, dP2 =  10.00):
Switches:  /MARK
LMSEL.LM from gel ACC#'s 0054.1 and 0036.1
The (Representitive)R-gel is: 0054.1
 LANDMARK #A  G1[353, 229], G2[285, 235]
 LANDMARK #B  G1[330, 181], G2[268, 175]
          .
          .
          .
G1[A, 349][ 351, 230],E.Diff= 2.2, G2[A, 167][ 286, 236],E.Diff= 1.4-OK
G1[B, 226][ 329, 182],E.Diff= 1.4, G2[B,  69][ 267, 175],E.Diff= 1.0-OK
          .
          .
          .
R[A]=  12 to nearest LMs[V,V], next nearest LMs[U,U]
R[B]=   5 to nearest LMs[C,C], next nearest LMs[A,A]
          .
          .
          .
Marked gel comparison files are: U00185.PIX and V00185.PIX on [44,3]
G1 HAS  453, G2 HAS  207 SPOTS
TOTAL DENSITY G1= 3173.00, G2= 1155.77
OMITTED TOTAL DENSITY G1= 675.29, G2= 2293.90
```

Note. An example of the initial gel comparison file (GCF) output of CMPGEL program applied to gels 54.1 (PHA lymphocyte gel) and 36.1 (resting lymphocyte gel) in the effect of PHA experiment. The manually selected landmark spots are first listed followed by the number of spots in each gel. The best spot estimates of the landmarks are then given. The Euclidian distance between the segmented and manually defined landmark spots are listed. The half-radii and next-nearest-neighbor landmark spots are then listed. Total spot densities are listed for each gel for both included and noise (omitted) spots.

correct. However, a small radius of confidence may have undesirable effects on pairing. Furthermore, spots that would otherwise be matched as sure pairs might be entered into the probable pair category. Thus selecting landmarks too close to each other probably introduces an incorrect bias in partitioning spots due to digitization-type errors and a higher probability of interacting with more landmark sets. It is possible for a spot pair to be found in the next to 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 as a result of increasing the concentration of landmarks is obscure and does not seem easily treated.

Considerations of correctness and completeness of the primary pairing algorithm are not simple although the algorithm in itself is quite straightforward. The performance should not be gauged exclusively on the results when gels of widely different spot numbers are compared. On the other

TABLE IIIb

EXAMPLE OF GEL COMPARISON FILE–PAIRING DATA

```
LM[A] G1 HAS      15, G2 HAS       9 SPOTS
#A G1:358[-30,  2]&G2:165[-24, -4] PP,DP= 8.5,DL=30,D1=31.8,D2= 4.8nd1.0Mnd
#A G1:327[-20,-13]&G2:164[-16, -6] PP,DP= 8.1,DL=24,D1= 5.4,D2=12.0nd .5Mnd
#A G1:333[ -5,-11]&G2:160[ -9,-10] AP,DP= 4.1,DL=13,D1= 6.3,D2= 5.3nd .3Mnd
#A G1:334[  6, -9]&G2:156[  4,-14] PP,DP= 5.4,DL=15,D1=10.9,D2= 3.7nd .3Mnd
#A G1:338[-11, -9]&G2:160[ -9,-10] PP,DP= 2.2,DL=14,D1= 7.1,D2= 5.3nd .3Mnd
#A G1:349[  0,  0]&G2:167[  0,  0] SP,DP=  .0,DL= 0,D1=58.4,D2=60.4nd1.2Mnd
#A G1:350[ 11, -3]&G2:  0[  0,  0] US,DL=11.4,DL=11,D1=  .8,D2=  .0nd1.1Mnd
#A G1:352[-21,  1]&G2:173[-24,  1] PP,DP= 3.0,DL=24,D1=58.2,D2= 3.8nd1.2Mnd
#A G1:353[-16, -2]&G2:172[-15,  0] PP,DP= 2.2,DL=16,D1= 4.9,D2=22.1nd .8Mnd
#A G1:354[ -9, -1]&G2:170[ -8, -2] SP,DP= 1.4,DL= 9,D1=30.6,D2= 8.0nd1.0Mnd
#A G1:359[ 14,  1]&G2:  0[  0,  0] US,DL=14.0,DL=14,D1= 2.9,D2=  .0nd1.1Mnd
#A G1:362[ 21,  2]&G2:  0[  0,  0] US,DL=21.1,DL=21,D1= 2.8,D2=  .0nd1.1Mnd
#A G1:365[ 10,  5]&G2:  0[  0,  0] US,DL=11.2,DL=11,D1= 4.5,D2=  .0nd1.1Mnd
#A G1:367[-11,  6]&G2:172[-15,  0] AP,DP= 7.2,DL=15,D1= 1.3,D2=22.1nd .8Mnd
#A G1:371[  5, 11]&G2:  0[  0,  0] US,DL=12.1,DL=12,D1= 2.5,D2=  .0nd1.1Mnd
#A G1:374[-17, 12]&G2:  0[  0,  0] US,DL=20.8,DL=21,D1=  .4,D2=  .0nd .8Mnd
#A G1:333[ 60,-17]&G2:155[-72, -9] AP,DP= 4.5,DL=17,D1= 6.3,D2= 3.5nd .1Mnd

LM[B] G1 HAS 50, G2 HAS 27 SPOTS
#B G1:219[-32, -7]&G2: 54[-32,-14] PP,DP= 7.0,DL=35,D1=  .3,D2= 2.9nd .2Mnd
        .
        .
        .
PAIRING STATISTICS
------------------
        After initial pairing:
US 31
SP 46
PP 222
AP 137
(SP+PP)/(US+AP+SP+PP)= 61.5%
        After secondary pairing:
US 15
SP 46
PP 242
AP 133
(SP+PP)/(US+AP+SP+PP)= 66.1%
Real TIME =00:07:02
CPU TIME  =00:02:02,   28.903%
```

Note. The final labeled landmark sets are listed in the remainder of the GCF with pairing statistics for gels 54.1 and 36.1. The bracketed 2-tuple is the relative cartesian distance from the spot to the landmark. The segmentation spot index preceeds the "[." The spot labels are (SP,PP,AP,US). DP is the distance 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 gel 1 (gel 2). Mnd is the maximum ND value seen for either spot in the pair.

hand comparisons of closely similar gels are too "easy" to be regarded as a valid test. The proper test object characteristics are still incompletely defined.

Similarly, the results of the secondary pairing must be viewed in the light of relative spot numbers and distributions within the gels to be compared. The seemingly small number of additional matches resulting from secondary pairing may well indicate that the major problem lies elsewhere, e.g., in gel inequalities, etc. Properly designed tests on suitably constructed test objects are an item for future work. The current secondary pairing algorithm is a first approximation to solving this problem and does not handle some of the cases which might require looking in other landmark sets for possible re-pairing of spots.

The current pairing algorithm defines a sure pair (SP) as being within the landmark radius Ri for landmark i. We have found that most of the possible pairs (PP) are actual pairs and should be pooled with the sure pairs as well

matched spots. Ambiguous pairs and unresolved spots may result from comparing two widely different or noisy gels. Currently, nothing is done with these (AP and US) spots. This opens up a possible extension to the algorithm incorporating additional procedures to further process the AP and US spots. Spot conglomerates sometimes appear as single spots and other times as several spots, e.g., actin complex.

Occasionally, a landmark spot will not be segmented correctly or will not be defined well by the operator. When this happens, the CMPGEL program rejects that landmark since it will have a large distance to its corresponding GSF spot. This has the effect of placing the set of spots which *would* have been

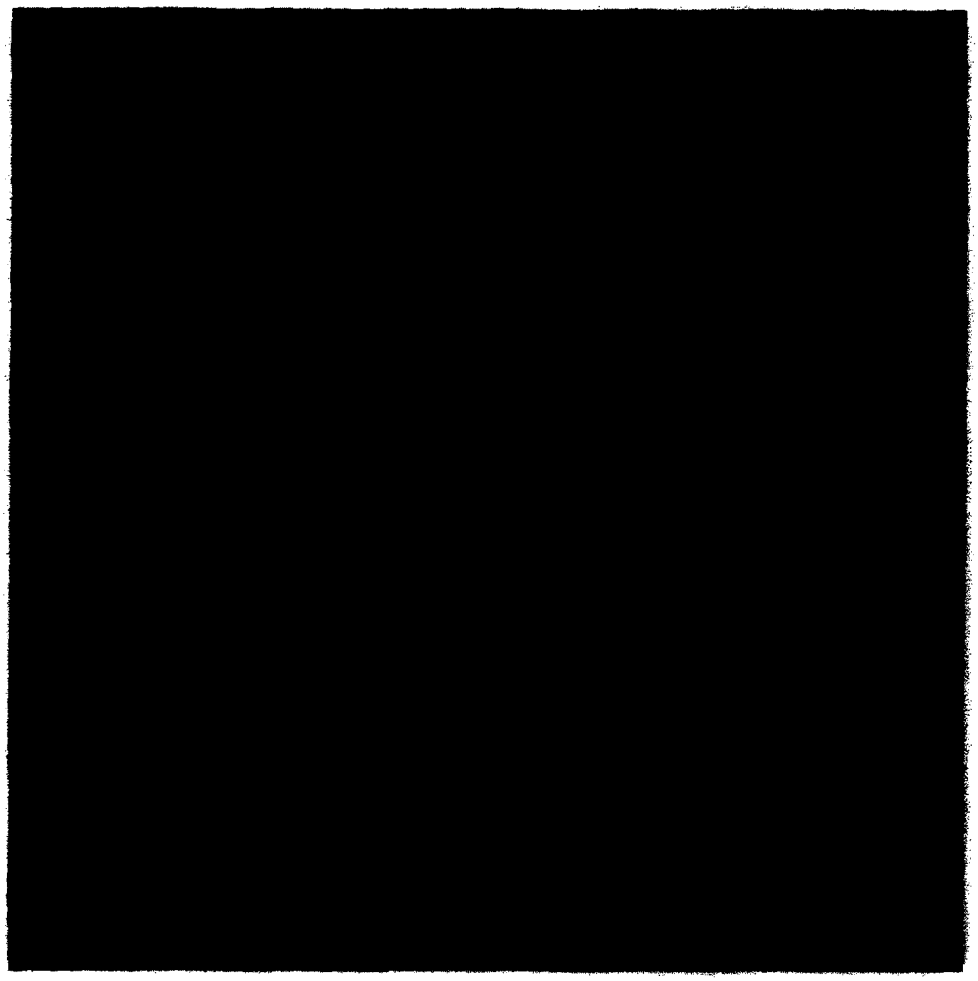


FIG. 10. Density vs density scatter plot (on a log-log scale) of lymphocyte gels 102.2 (PHA) and 103.2 (no PHA). The density values are normalized as a percentage of total D' of the SP + PP paired spots. The scale ranges from 0.1 through 100.0 with decades denoted by extended scale markers.

in that landmark set into adjacent landmark sets. The net result is that some spot pairs are either not paired, are mispaired, or are paired with a higher DP and DL than if placed in the proper landmark set.

We have found that highly populated spot regions should have somewhat more landmarks but landmarks should not be "on top of" one another. Other criteria in landmark selection include using fewer landmarks if the regions have little distortion and line up fairly broadly and well. A landmark should be well defined morphologically being part of a consistent pattern in all of the gels to be compared.

At present depending on gel quality (the single most important factor) and the set of landmarks selected (which must be in all gels to be compared), manually landmarking a pair of gels takes from 3 to 30 min depending on the comparibility of the gels with an average time being about 5 to 7 min.

The ability to pair most of the spots in a set of gels enables examination of larger gel data bases and detect subtle shifts and correlations in the spot data. Figure 10 shows a scatter plot of normalized paired (SP and PP labels only) lymphyocyte gels. Most of the spot pairs are close to the 45° line. Some of the outliers are real and some due to noise in the entire gel-image processing

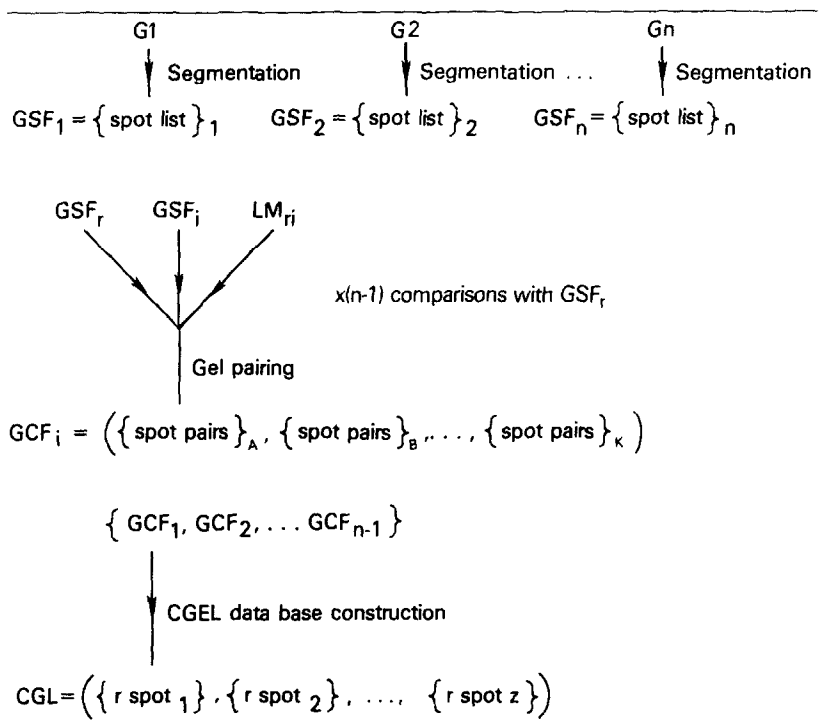


FIG. 11. Data structures used in the gel analysis. The GSFs (gel segmentation files) are produced by the segmentation of the gel images. The GCFs (gel comparison files) are produced by comparing the GSFs using a set of landmark spots. The CGL data base is constructed by merging the GCFs.

system. The next paper (*2*) discusses techniques for further resolving noisy data using multiple gels and means for facilitating the checking of outliers.

5. Conclusion

An algorithm for the spot pairing of 2D electrophoretic gel image spot lists with generation of a gel comparison file has been presented. This algorithm is successful in pairing spots under a wide variety of gel conditions. It is useful as a second stage processor (after image segmentation) in the analysis of 2D gels. Because there is no need for user interaction during gel pairing after the initial landmark selection has been performed, a series of gels may be batch processed to increase efficiency.

After matching, each spot pair produced is ranked with a set of heuristic pairing features. It is often desirable to compare 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 successive data reduction performed at each stage of multiple gel processing 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 (CGL). The third paper discussed this last procedure and its ramifications for 2D gel analysis.

Appendix

A.1. Landmarking a Set of n Gels to a Single Gel

Landmark spot sets are manually defined using the Real Time Picture Processor, RTPP (*14–18*), alignment program by interactively flickering two gels and noting corresponding alignments for a specified set of landmark spots. The initial landmark set is defined for the single gel (denoted the reference or R-gel) in file LM0001.DA. The operator interacts with the TV system using a spark pen tablet to point to a spot while a corresponding white cursor moves over the TV image of the flickering gels. Pressing the pen causes the current position to be noted. Since this landmark set is to be used to define the landmarks for the reference gel, it may be landmarked with any gel including itself. This file and the list of the other n-1 gels to be processed are input to the MAKCMP (cf. Appendix A.2) program which generates a batch job for the RTPP. This batch job will then generate, with the users aid in landmarking, the set of n-1 landmark spot files (LM0002.DA, . . ., LMn.DA).

Alternatively, when landmarking a set of gels to a given reference gel, it is possible to speed up the landmarking process by advancing the cursor to the next landmark under program control. The user then only has to flicker align the two gels at the indicated spot. This feature was found to save on the order of two to three times the time normally required. In addition, it guarantees the

consistency of the landmarks selected because the computer sets the cursor to the next landmark in the R-gel. Thus the initial landmark set will be the same for all gels. After the set of landmark files are defined, they are transferred to the main computer.

The new landmark spot sets are then appended to the end of landmark spot file LMS.LM which is used by CMPGEL along with the particular accession file GEL.ID (cf. (*1*)). This permits the construction of an extensible landmark spot data base which may be referenced by its contents. That is, a set of landmark spots is accessed in the LMS.LM file by a pair of accession number names. A sample of the LMS.LM landmark data base file is illustrated in Table II.

A.2. Semiautomation of Landmarking

When landmarking several gels against a single standard gel, it is useful to use the same set of landmarks in all gels. This can be done manually using the spark pen but is very tedious. A better mechanism is to landmark one gel with the standard gel to define an initial landmark set. Then, this initial landmark set is used in subsequent landmarking to preset the position of the cursor prior to aligning and marking the next landmark between the new gel and the R-gel. The procedure is performed in three steps:

(1) Generate the initial landmark set using the standard gel.

(2) Run the MAKCMP program which will generate a batch job after being supplied with the following parameters:

(a) Name of the initial landmark set file.

(b) Name of the batch file to be generated.

(c) List of gel pairs to be landmarked where the first element of each 2-tuple is the standard gel.

(3) Run the interactive batch job just generated. This job has the standard landmarks supplied for each gel to be landmarked with the standard gel. Landmarking is then performed by enforced correspondence of landmark spots in the standard gel. These are successively indicated on the display by a cursor.

Use of this semiautomatic algorithm was found to decrease the landmarking time required by a factor of 2 to 3. In addition, it reduced fatigue and errors introduced by the user in selecting the wrong spots as the landmark spots in successive gels.

ACKNOWLEDGMENTS

The constant help afforded by Morton Schultz, Bruce Shapiro, and Earl Smith, our colleagues in the Image Processing Unit, 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. We are particularly grateful to our collaborator Eric Lester for allowing us the use of his intermediate results on PHA stimulation of lymphocytes.

REFERENCES

1. LEMKIN, P., AND LIPKIN, L. GELLAB: A computer system for 2D gel electrophoresis analysis. I. Segmentation and system preliminaries. *Comput. Biomed. Res.* **14,** 272 (1981).
2. LEMKIN, P., AND LIPKIN, L. GELLAB: A computer system for 2D gel electrophoresis analysis. III. Multiple gel analysis. *Comput. Biomed. Res.* **14,** in press.
3. LEMKIN, P., MERRIL, C., LIPKIN, L., VAN KEUREN, M., OERTEL, W., SHAPIRO, B., WADE, M., SCHULTZ, M., AND SMITH, E. Software aids for the analysis of 2D gel electrophoresis images. *Comput. Biomed. Res.* **12,** 517 (1979).
4. BOSSINGER, J., MILLER, M. J., KIEM-PHING, V., GEIDUSCHEK, P., AND XUONG, N. Quantitative analysis of two-dimensional electrophetograms. *J. Biol. Chem.* **254,** 7986 (1979).
5. ANDERSON, N. G., AND ANDERSON, N. L. Molecular anatomy. *In* "Proceedings, Behring Inst. Symposium 1977," p. 169. Mitt. Behring Inst. Sympos. No. **63,** 1979.
6. ANDERSON, N. G., ANDERSON, N. L., AND TOLLAKSEN, S. L. Proteins of human urine. I. Concentration and analysis by two-dimensional electrophoresis. *Clin. Chem.* **25,** 1199 (1979).
7. GARRELS, J. I. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254,** 7961 (1979).
8. REISER, J. F. 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., 1976.
9. LESTER, E. P., LEMKIN, P., LIPKIN, L. E., AND COOPER, H. L. Two-dimensional electrophoretic analysis of protein synthesis in resting and growing lymphocytes, *in Vitro, J. Immun.* **126,** 1428 (1981).
10. LESTER, E. P., LEMKIN, P., COOPER, H. L., AND LIPKIN, L. E. Computer-assisted analysis of two-dimensional electrophoresis of human peripheral blood lymphocytes. *Clin. Chem.* **126,** 1392 (1980).
11. BOOKSTEIN, F. L. "*The Measurement of Biological Shape and Shape Change.*" Springer-Verlag, New York, 1978.
12. ROSENFELD, A. "*Picture Processing by Computer.*" Academic Press, New York, 1969.
13. MCCONKEY, E. H. Double-labeled autoradiography for comparison of complex protein mixtures after gel electrophoresis. *Anal. Biochem.* **96,** 39 (1979).
14. LEMKIN, P. "Buffer Memory Monitor System for Interactive Image Processing." NCI/IP Technical Report No. 21b, Nat. Tech. Inform. Serv. PB278789 (listing PB278790), 1978.
15. LEMKIN, P., AND LIPKIN, L. BMON2—A distributed monitor system for biological image processing. *Comput. Programs Biomed.* **11,** 21 (1980).
16. CARMAN, G., LEMKIN, P., LIPKIN, L., SHAPIRO, B., SCHULTZ, M., AND KAISER, P. A real time picture processor for use in biological cell identification. II. Hardware implementation. *J. Histochem. Cytochem.* **22,** 732 (1974).
17. LEMKIN, P., CARMAN, G., LIPKIN, L., SHAPIRO, B. SCHULTZ, M., AND KAISER, P. A real time picture processor for use in biological cell identification. I. System design. *J. Histochem. Cytochem.* **22,** 725 (1974).
18. LEMKIN, P., CARMAN, G., LIPKIN, L., SHAPIRO, B., AND SCHULTZ, M. Real time processor—description and specification, NCI/IP Technical Report No. 7a, Nat. Tech. Info. Serv. PB269600/AS 1977.