COMPUTERS AND BIOMEDICAL RESEARCH 14, 407–446 (1981)

GELLAB: A Computer System for Two-Dimensional Gel Electrophoresis Analysis. III. Multiple Two-Dimensional Gel Analysis

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

This is the last of a series of three papers describing a computer analysis system, GELLAB, for aiding the digitization, segmentation, interimage comparisons, and analysis of two-dimensional (2D) gel electrophoresis images. The system is directed to the identification of biologically significant changes manifest as spot differences despite complications introduced by a wide variety of gel preparative, staining, detection, and digitization variances. The operations applicable to single-gel images (segmentation, etc.) are described in the first paper. The second deals with several of the elements that will appear in the fully developed data base; i.e., constructs developed from images taken two at a time, based on regional correspondences established by landmarks. This paper focuses on methods for handling multiple gels. The keystone of this analysis is the R-gel, the representative gel image. This acts as the major framework to link spots across the set of gels and serves as a practical substitute for an unattainable canonical model. It is the referent around which the total spot data base is organized. The primary data base consists of the totality of the lists of corresponding spots (R-spot sets) and their associated properties and interrelations. Interrogation of, and experimentation with, the data base allows the user to find and extract measurements of biological significance from these congeneric polypeptides. Some of the problems of and some solutions for dealing with a large number of possibly incomplete biologically determined intergel spot density distributions are presented. Solution strategies include constructing mosaics, using the digitized images of areas surrounding selected spots, and creating labeled map images of R-gel correspondences. Listings of relative position and density information of ordered R-spot sets is also of value in this interactive system which permits refining of the initial data base. The multiple-gel data base system is very general in that gel segmentation and spot-pairing algorithms other than those now used in GELLAB may be substituted in the early phases of data base construction. Discussion of CGEL, the GELLAB program for constructing, partitioning, searching, retrieving, and formatting the data base, is illustrated using results obtained from 2D gel data of phytohemagglutinin (PHA) stimulation of lymphocytes. Experience with utilizing GELLAB on a variety of biological problems has suggested potential system developments and new system features which are noted. GELLAB runs on a DECSYSTEM-20 (or DECSYSTEM-10) and was written in SAIL.

1. INTRODUCTION

A 2D PAGE electrophoretic gel is a complex of distinct polypeptides, each one of which is characterized by density and position relative to other polypeptides ("spots"). Unlike a geographic map, proximity of polypeptides on a gel is no particular indication of related genesis or biological function. But

0010-4809/81/050407-40$02.00/0

nevertheless the large number of discrete spots in a gel and the similarity that is preserved among gels from a similar source allows one to follow many proteins in a single determination. 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.

Previous papers in this series (*1*, *2*) have established the need for computer support of 2D gel electrophoresis analysis. The building of such support along largely data-structural lines has been shown to be essential. Many conditions (e.g., intergel spot position variability), all of them multidetermined, have been shown to critically affect the ways in which we can extract biological facts from experiments by employing 2D gel electrophoresis as an analytic tool. We have treated the problems of spot extraction (*1*), and pairwise spot comparison (*2*), and in the process we have indicated that experiments involving dose or time variables require comparisons of spots from multiple gels. We will 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 program, CGEL, for multiple-gel analysis.

In dealing with these and other problems, we have come to feel that computer support of 2D gel analysis needs to be along largely data-structural lines, where associated spots and their characteristics can be grouped by one criterion and readily regrouped as one attempts to "see around" the data from several perspectives. From our early efforts at gel analysis (*3*) 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 (i.e., the program needed to be able to partition and to concatenate lists acquired at different times and from different gels), because without checking *all* or at least most of the spots in the set of two or more gels, no statement of types of spot differences can otherwise be made. This implies 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 a triple, comprising x and y position (centroid) and an adjusted integrated density value. Among gels, the idiosyncratic variations of these triples due to variation in preparation, detection, etc., confound what are the "real" variations produced in the biological/clinical system by dose, time, clinical state, etc. In a sense analogous to the canonical matrix (which, for example, characterizes a conic independent of position, rotation, etc.) we propose the concept of a canonical gel or C-gel, valid for the domain of a given experiment or a defined clinical situation. Such a C-gel, a mathematical, and nonpictorial object, would provide information characterizing position and density distributions for all spots over all gels in the set. Further, it would exclude the data idiosyncratic to preparative and detection conditions unrelated to the biologic issue. A necessary but not sufficient condition for construction of a C-gel would be the spotwise 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 begin the construction of a canonical gel would require $(n - 1)$ factorial comparisons times the number of spots. Since each element of the C-gel would be a function expressing the variation of the spot descriptor triple as a function of the biomedical variable, it is not likely to be a feasible construct. Though practically not easily realized, 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 a pictorial object. It is a real gel chosen from the set 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 almost all if not all spots encountered in any of the members of the set. It is not necessarily a 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 for multiple comparison.

In contrast to C-gel construction, the R-gel is used as the basis against which other gels in the set are compared. As noted previously (*2*), each spot in the R-gel is the potential index to a R-spot set. A R-spot set is the set of spots, one at most from each gel in the set of gels, which corresponds to a given spot in the R-gel. The set of R-spot sets would under ideal conditions include all spots in all gels. Until biochemistry can provide essentially noise-free gels, such a complete and ideal accounting is simply not attainable.

2. General Method of Analysis

2.1. The General System of Analysis

The design philosophy underlying the components of the GELLAB system that deal with multiple gels is the interactive and flexible manipulation of spot data organized by congeneric association. Paired spots and their densities and locations are recorded in a congener-oriented data base, which can be searched from a variety of ways and a variety of representations, numeric, diagrammatic, pictorial, textual, or tabular, of this data base or of its derivatives can be instantly displayed in order that the researcher may quickly grasp patterns and implications. Hypothesis verification is performed by interactive reordering and new representations of segments of the data. This section discusses the general approach in terms of concepts central to it.

Fundamental to our system of analysis of multiple gels is the concept of congeneric polypeptides giving 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. Under varying experimental conditions, the quantitative expression of such production may be muted or exaggerated. But in each gel where it is detected, the spot denoting the congeneric polypeptide occupies the same relative position in the local morphology.

The List of R-spots and R-spot Sets. Formally, we represent these linked constructs, the list of R-spots and the R-spot set, as follows:

1. *The list of R-spots.* All the distinguishable spots in the R-gel, taken together, constitute a list of so-called 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. 1, A,B,C, etc., in Gel R). Spots that compose the list of R-spots must be distinguished from a R-spot set.

2. *A R-spot set.* This is a set of spots, having at most one member from each gel (but definitely including a particular member of R-spots), corresponding to a given spot in the R-gel (cf. Fig. 1, the B series of spots). Each member of a R-spot set is a congener of that particular polypeptide. 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) array of gels.

The linkage and reciprocal dependency between the single gel list of R-spots and a R-spots 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 program 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. 1, series C).

Thus, a R-spot set has a membership of at least two congeners. Spots which

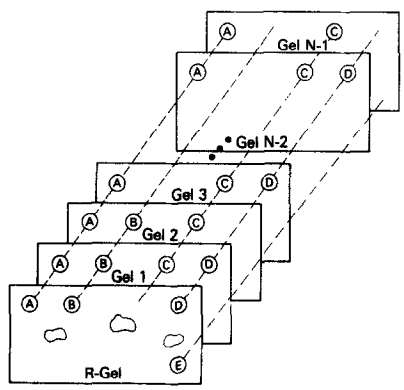

FIG. 1. Spots A, B, and D in the R-gel are R-spots. Spot E, though it is in the R-gel, is not an R-spot. This is because it is unmatched in any of the other gels in the set. The list of R-spots in the gel is {A, B, D}. R-spot set A consists of {Ar, A1, A2, . . . , A n-1}. This set has one member from each gel. Some R-spot sets such as B and D have members only from some of the gels. The presumption 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 would be given by the following: 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 by another polypeptide. This kind of false positive is largely dependent on the pairing algorithm. The set C is not an R-spot set, because though represented in all other gels, there is no member in the R-gel. The presumption is that the members of this set 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. Of course experiments involving more than one R-gel are possible and are the solution to this particular difficulty.

are in the R-gel but which do not correspond to spots in any of the other gels are obviously not R-spots.

A R-spot set represents a presumptive congeneric set of polypeptides. Figure 1 shows the possibility of a congeneric set which is not a R-spot set (spot set C). Since R-gels are real objects and 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. The list of R-spot sets constitute the totality of spots in the data base, i.e., the array organized as the list of R-spot set vectors.

Local Morphology. We have found that for gel analysis a most effective image-processing strategy is to concentrate on sets of local morphologies (both within and across gels) rather than to treat one object at a time as is commonly done in most biological image processing. 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 made by machine and the human confirmation. Although the individual spot may be considered the biochemical primitive, from the viewpoint of image processing (machine or human) the operational primitive form is the local morphology, i.e., the minimum region of extent which provides location and thus identification information.

Recognition and identification (as opposed to detection) is quite difficult because of the absence of fixed size and shape of the individual spot. The problem is analogous to that of an observer, without an ephemeris (a table of computed star positions for every day of a given period) and without a clock, who is asked to identify a single star where all others are artificially blocked out of his field of vision. Just as it is easier to identify some stars for which we have established rules (e.g., the pointers to the pole star) so it is necessary to establish, if only empirically, some relative-position information to yield spot identification data. We are then, in dealing with spot identification, actually concerned with problems of local morphology, in which we are aided by the machine to (1) establish the proper window of regard, (2) maintain the local coordinate system, and (3) perform alternate pictorial and numeric comparisons. It is only seemingly paradoxical that the absence of internal structure that makes it difficult for the human to identify individual spots makes it simpler for machine procedures at the single-spot level. Contrariwise, in order for this automatic computer aid to do us any good, at least for the present, it is first necessary that the local morphology be exploited by the human who can readily isolate a spot given its local context.

The local morphology cannot be considered a fixed entity. It is obvious that the extent of the region which provides the identity-establishing positional information will vary depending on spot representation in both a quantitative and qualitative sense. In this regard the landmark spot, denoting the landmark region, serves as a key to define an operational local morphology. The machine facilitates the discovery and confirmation of corresponding local morphologies, which, once established, are labeled. The system, at user prompting, is capable

of naming an object undistinguished by any feature other than its relative location.

The Data Base. Our previous papers have detailed procedures that have been preparatory in that they deal with operations on individual spots or spot pairs. Having constructed our 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, experimental, clinical, or temporal variables. The richness of the data base does not limit us to any one of these; the facilities which we now describe allow a multiplicity of orderings. A variety of presentations may be chosen, which may be best determined by the nature of the experiment. The biology demands that the analytic process be limited in its "attention" to a set of congeneric spots, one from each gel, a process that transcends the constraints of the individual gel. Our data management system permits this type of analysis to be applied successively to the majority of such spot sets.

The primary data from which the data base is constructed consist of (1) the digitized gel image, (2) the GEL.ID accession file with its experimental/clinical data and calibration information, and (3) the LMS.LM landmark set file consisting of the set of landmark spots, interactively established by the user, relating the chosen R-gel to each of the other individual gels in the experiment.

The primary data base that is constructed consists in its simplest form of the list of R-spot sets. This means that not every spot in every gel is represented. From the previous definitions, it is clear that a spot in the R-gel that has no congener in another gel will not be represented in the data base. Similarly, a set of congeners that have no representative in the R-gel will not be represented in the data base. The data base therefore is not even completely inclusive of all congeneric sets. From the preparative viewpoint, the better the gels in an experiment as well as the more adequate the segmentation and spot matching, the richer and more complete is the list of R-spot sets and each component R-spot set. For the present, for congeners with no representative in R-spots to be in the data base, a new R-gel would need to be established.

The data-structural operations performed consist of a series of computational and representational operations on the list of R-spot sets or sublists or lists of R-spot subsets. The latter subsetting may be automatically accomplished based on a characteristic of a gel (from the accession file), on a statistical property, etc. (Note: The reader may recall from previous papers in this series that each gel has, associated with it, accession file information, total gel density, and number of spots in the gel, which is used to label tables and plots as well as for normalization for some operations.) Alternatively, the user may construct at will working sets from the entire set of gels. A wide variety of representations of the data, both image and numeric, is available with numerous optional modes of display including superimposition on the original image. Important data structures are:

1. The set of working gels used to restrict the CGEL operations to a subset of the gels in the data base. Only gels in the working set are used in the computations.

2. The classification sets which contain the names of the gels in each of up to nine classes. Thus, the user can, depending on the problem he is dealing with, classify gels by temperature, by disease, by metabolic condition, etc.

3. A "search results list" of R-spots which were found by one or more of the various available search options (or explicitly defined) is available to many of the CGEL operators.

In dealing with real data, it is frequently necessary to create a working subset of the original data base. The same data may be used to analyze different aspects of the same experiment. A related requirement is the facility to declare classes of gels and to create further subsets based on class membership.

In any set of gels with associated experimental conditions, it is useful to partition them in various ways in response to different questions. Thus, for example, in the case of a much distorted poorly run gel with many outliers, one might wish to temporarily remove it from the set of gels in order to find statistically significant spots in the remaining members of the set. Later, the temporarily removed gel(s) could be restored to the set and these spots checked. Effectively, this procedure uses the results of the "good" portion of the set of gels to investigate the outliers.

Solution Strategies. The types of properties that characterize spots, the principle of local morphology, and the varied objectives that different users will bring to the analysis indicate a system which does not produce a simple solution. Instead the system offers a solution strategy or set of strategies rather than a direct and single solution. Prominent among the tools available for such strategies is the multiple representation of the same data, with a seemingly prodigal retention of what elsewhere might be considered intermediate results or scratch images. Many of the system procedures are essentially procedures of presentation, again allowing the user to alternate between, say, numeric position data and synthetic images. The segmenter output is a case in point (*1*).

Other tools available to the GELLAB user are R-maps and mosaic images. These images 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 represents a many-to-one mapping whereby corresponding regions from each gel are brought into physical proximity in a single synthetic image. The regions are each centered around the spot of interest for the corresponding gel. The mosaic provides a powerful tool whereby the user may assure himself 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 individual spot as seen from the numeric R-spot set data or local mosaic image and its place in the gel. It is invaluable for rapid evaluation of the validity of spots found to be of interest by GELLAB searches or manual examination of the data lists. Because of the locality of mosaics they are insufficient for establishing a spot's context and thus the R-map fills this void. Numeric data, particularly coordinate and density values presented in rank-ordered tabular form, are useful for evaluating magnitude differences between spots in an R-spot set. The gray-scale numeric representation of each pixel comprising a spot is occasionally useful in determining whether a spot is

actually one or two or whether a single spot was split by the segmenter. The GEL.ID accession file information always travels with a data set or its derivatives. Any portion of it may be used as the associative key with which to regroup data within the data base.

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

In sum this represents a general method of organizing and selectively compressing the data of 2D gels so that the user may more efficiently perceive patterns out of the welter of individual spots. Once patterns are established it is a direct process to quantitate their individual components by merely printing their R-spot sets.

In such an approach it is vital that certain automatic features be emphasized. In particular the automatic establishing of the consequences of an identification or confirmation of spot correspondence is imperative. The better this is done, the less there is to block the bringing of derived data into spatial and/or temporal proximity, the latter being a necessity for human evaluation.

Analyzing Multiple Gels as a Continuum. Each polypeptide visualized as a spot may be thought of as having a distribution of spot densities when sampled in a set of gels. In the case of significant spot density differences, it is expected that this distribution will cluster multimodally according to the biological state of the sample. It is important therefore that biologically nonsignificant variances be controlled and minimized. Adequate numbers of 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 since 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 some biological problems where the changes are above the noise level and resolvable by the system. At no point in the analysis of gels should the computer-generated decisions be the endpoint of the analysis.

2.2. CGEL Spot Data Base Analysis System

An overview of the entire gel analysis procedure is illustrated in Fig. 2. The hardware environment 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. The RTPP is described in Refs. (*6, 15–19*). The DEC-2020, using the TOPS-10 monitor, has 512K words of 36-bit memory, three 160-Mbyte disks and two magtape drives. Additional details are provided in Appendix A, while Fig. 3 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. Using the gel image files, accession file, and

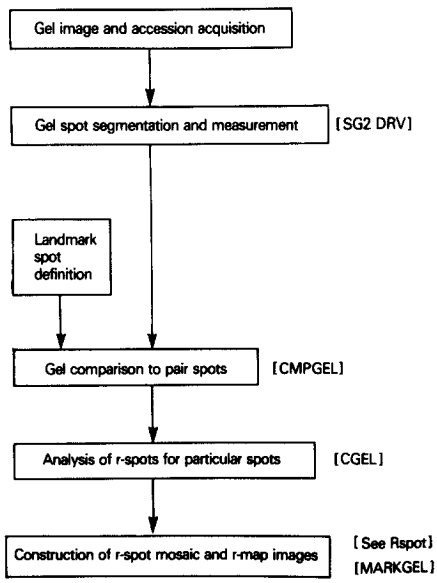

FIG. 2. Block diagram of the 2D-gel analysis GELLAB system. Programs associated with major steps of GELLAB are indicated in "[. . .]." Gel images are acquired by scanning with a vidicon TV camera interfaced to a picture memory and saved on the computer. Accession information about the set of gels is 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 are 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 composite gel data base (CGL). Thus a R-spot set (supposedly) contains congener spots from all the gels in which it occurs.

landmark spot sets file, the spot segmentation, gel spot pairing, and CGL data base construction and analysis are performed on the DEC-2020. A cost accounting of the various steps in the analysis of an average set of 20 gels found typical DEC-2020 times were about 25 min of cpu time/gel with 150K-word program core sizes and about 15 min of RTPP real time/gel.

The consequences of this environment impose some practical limits to the capacity of GELLAB. As illustrated in Fig. 3, gel analysis is primarily a series of data reduction steps mapping image information into a set of (about 1000) spot density distributions. Images are reduced to spot lists. Spots lists are reduced to spot pair lists, and finally, spot pair lists are reduced to a list of R-spot sets. Clearly, comparing 300 to 1000 spots in up to 25 gels would be a monumental task if done manually. The majority of the computation is involved in the initial image data reduction whereas further analysis is dependent on the type of questions to be asked about a particular set of gels.

Procedures used in the later phases of analysis are based on the analytic principles discussed above and carried out by means of an interactive program

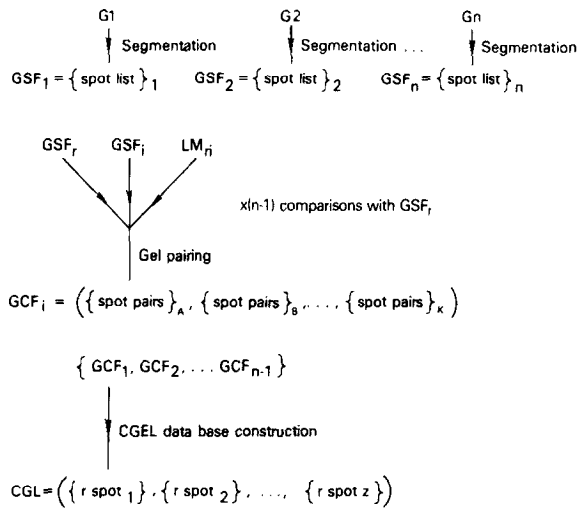

FIG. 3. Data structures used in the gel analysis. The GSFs (gel segmentation files) are produced by the segmentation of the gel images. The GCFs (gel comparison files) are produced by comparing the GSFs using a set of landmark spots. The CGL data base is constructed by merging the GCFs.

called CGEL. CGEL can be used to construct a representative spot data base and then to analyze all or part of this data base. It is an interpreter, taking the set of $n - 1$ gel comparison files (GCF) as input. As will be recalled from the previous paper, the GCFs are produced by the CMPGEL program as each of these gels is paired, one at a time, against the R-gel (*2*).

Generation of the R-spot Set Data Base and Overview. The first step in the construction of the data base is the generation of individual R-spot sets followed by their concatenation into the R-spot set data base (list of sets).

The set of $n - 1$ GCFs are read one spot pair at a time for each gel pair where one of the spots is a R-gel spot. Currently, up to 48 gels may be analyzed together. Each R-spot, referenced by a "key," is formed for this spot (see Appendix B) and the data base is tested to determine whether a R-spot set currently exists for that spot. If it does not exist, then a new R-spot is created and both spots are put into that set. If it does exist, then the other spot in the pair is inserted into that set. In either case, the R-spot set is initially ordered by the rank of the spots densities, darkest first. Alternate R-spot set orderings are then routinely performed as part of the analysis.

Subsets of the CGL data base may be constructed by restricting a R-spot set's consideration according to various parameters. In dealing with real data, it is frequently necessary to create a working (subset) set of the original data base. The same data may be used to analyze different aspects of the same experiment. Further, the working set of gels may be subdivided into up to nine user-defined classes. Then, various statistical test procedures may be invoked to search the set of R-spot sets for statistically interesting R-spots. In

presenting individual operations or commands, we will use data on lymphocyte response to phytohemagglutinin (PHA) for the majority of our illustrations.

The CGEL system will be presented in part by examples of operations and results obtained on one of several projects to which it has been applied in detail. An individual operation or command will be described and results obtained thereby cited immediately. We have chosen the work on lymphocyte response to phytohemagglutinin (PHA) for the majority of our illustrations. Table I lists the top-level CGEL commands.

Table II illustrates a simple sequence of CGEL commands which creates a CGL data base and in the process partitions the CGL data base into "resting" and "PHA-stimulated" classes. Tables IIIa and b illustrate some typical CGL R-spot set data base entries where spots are ordered by density. It should be emphasized that the term *density* will always refer to the value as determined by the current density mode. Spot density may be reported and used in computations as absolute (D'), percentage (relative to the total gel D'), and ratio density (D' relative to the sum of D' for selected spots). The SET DENSITY MODE command is used to change the current density mode. When the R-spot sets are printed, percentage density is denoted by "%", D' by "'", and ratio by "R". The percentage density mode is the default.

Typical characteristics of a R-spot set are shown in the first (IIIa) table. of the 13 gels in the data base only 10 contribute a member to the R-spot [1] set (gels 4.1, 36.1, and 51.1 are not represented). Prominent among the characteristics exhibited is the consistency of spatial position shown by corresponding spots which is explicit in the columns headed *DP, DL, Dx,* and *Dy*. Other characteristics include the certainty of pairing, i.e., some pair labels

TABLE I
CGEL TOP-LEVEL COMMANDS

```
CReate - Create a CGL data base from a set of CMPGEL .GCF files.
CHecking - Apply R-spot statistics checking until next command. This
        ignores R-spot sets out of range.
EDit - Edit spots from the CGL data base.
EXit - Exit CGEL to the TOPS-10 monitor to save image for later use.
Gels - Gels lists the names and total densities of the current gels.
Help - Print this message.
Inquire - Interrogate and search the CGEL data base for particular R-spots.
Plot - Draw (plot .PLX) %density/%density spot plots from the CGL DB.
REOrder - Reorder the CGEL R-spot database in current density mode.
REStore - Restore a CGEL R-spot data base from a .CGL file.
SAve - Save the CGEL data base in a (.CGL) file.
SET Accession file name - change the default GEL.ID accession file name.
SET Classes - Define gel class partition.
SET Density mode - Report results in Absolute, Percent or Ratio units
SET Fields - Set the list of fields desired for gel labeling.
SET Label - Set the "label" code to (S, P, A, U, *) used in searching.
SET RAtio list - of R-spots 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 - Generate an SPSS .SPS summary file of part of the specified
        CGL database.
TAbulate - Print: R-spot sets by Rank or Ratio, Mn-variation (.TBL) files.
TImer - Time commands until next TIMER command (default is timing on).
```

The top-level CGEL program commands listed here is available to the user on an interactive basis. The first part of the command that is required for it to be unique up to the lowercase letters of the command is given in uppercase.

TABLE II

CONSTRUCTION OF A DATA BASE: A SAMPLE CGEL COMMAND SEQUENCE

```
.RUN CGEL
*SET ACCESSION FILE
*gel01.id
*SET FORMAT
*2,3,10,12,13 - set specified accession file fields
*CREATE CGL data base
*c10001.gcf - 12 gel comparison files with R-gel=54.1
*c10002.gcf
        .
        .
        .
*c10073.gcf
*c10074.gcf
*
*(ps) - Sure pairs and Possible pairs only
*SET CLASSES
*automatic classification mode
*yes - change class names
*pha
*resting
*
*SET STATISTICS
*0,512 - range from mean of relative distance from landmark
*0,512 - range of DP allowed
*0,512 - range of DL allowed
*0,300 - mean density of R-spot set
*0,100 - standard deviation of R-spot set density
*10% confidence limit
*yes - R-spot set must have same # working gels
*INQUIRE
*index search for spots found in all gels
*SET RATIO LIST
** - use spot result list just found
*SET DENSITY MODE
*Ratio mode for reporting spot density
*REORDER
*INQUIRE
*rank order - search at 10%
*SPSS
*pha10r.sps - output file
** - use search results list to specify list of spots
*SAVE
*pha5.cgl
*EXIT
.SAVE PHA5.EXE
```

A typical CGEL command sequence used to construct a normalized CGL data base file (PHA5.CGL) and runable core image (PHA5.EXE). A rank-order search is performed to find statistically significant spots. The CGEL commands are given in capitals and the answers to the CGEL prompts are in lowercase. The "." prefix indicates a TOPS-10 monitor command while the "*" indicates a CGEL command. Comments are preceeded by "-".

are SP (sure pair) while others are PP (possible pair). More often a set will contain mostly SP or PP labels. Still other forms of spot characteristics include varying modes of density representation, i.e., absolute, ratio, or percentage of total.

R-spot sets [1], [41], and [119] illustrated in Table IIIa as percentages of total density and in Table IIIb as ratios (times 100%) normalized by a set of R-spots found in all 13 gels. R-spot [41] is a landmark spot with *DP*, *DL*, *Dx*, and *Dy* being 0 by definition. Appendix B discusses some of the details of the data structures used in the CGEL program including R-spot set data structures.

CGEL Commands. For convenience of reference the top-level CGEL operations are listed in Table 1. The user employs this interpretive system to analyze a set of gels as determined by a set of chosen parameters. Particularly when dealing with a new type of gel, the user employs CGEL "experimentally." Procedures are called forth, results are displayed and examined for confirmation or rejection of the tentative hypothesis, other procedures are called, and so forth. The nature of the interaction is highly dependent on the scientific questions asked of the gel data base.

In the following description of these commands, the command names are denoted in capital letters.

The set of CGEL program commands includes the CREATE operation which is used to build the initial R-spot CGEL data base from a set of GCFs. The accession file, describing a set of gels, must first be declared with the SET ACCESSION FILE command. Gels instantiated in the CGEL data base will then have their associated accession file information instantiated as well. The SAVE command saves the R-spot data base by creating an ASCII data file with a ".CGL" file name extension. The RESTORE command provides the user option of restoration or merging of CGEL data base files. The GELS command lists the names, accession information, number of spots segmented, and total gel density, and total density of the ratio-normalized spot set (see below) for each gel entered into the CGEL data base. It could be used, for example, to generate the first part of Table IIIa.

At any stage, the EXIT command enables the user to save the entire core image including the data base. This kind of checkpointing is especially useful in preserving the investment in the data base and experimental selection of parameters.

Each gel has various experiment-dependent information documented in the accession file (*1*) as shown in Table IV. The CGEL program extracts selection information from the accession file during data base creation. The fields of each gel record used in the CGEL data base may be declared prior to data base creation or changed (and thus updated) later. The SET FIELDS command requests a list of accession file record field numbers after printing out the following template, where the numbers below the fields correspond to the field specified:

```
ACC #/PATIENT/BIRTHDATE/RACE&SEX/EXP DATE/EXP #/
  1       2          3            4           5          6
   CULTURE REAG/AMPH, GEL/INTRVL BEFR LBLNG/
           7                8           9
   LBLNG  ISOTOPE/DURTN LABEL/DURTN OF EXPSR/
          10                11                12
     STUDY/FILE #/TAPE #/OPT. BACKUP TAPE #/
    13        14     15           16
      CAMERA,LENS,DISTANCE/EXPRMNTR*.
                  17                18
```

TABLE IIIa

EXAMPLES OF CGEL DATA BASE FILE

```
Output file: PHA5.CGL 03/27/1980,  06:01:21 PM
Total density[0054,1]=3253, # spots=425
   Study: PAT:3/PHA/120 HRS/H3/4 HRS/21 HRS/MITOGEN STUDY/
Total density[0001,1]=1776, # spots=265
   Study: PAT:1/PHA/120 HRS/S35/4 HRS/3.2 HRS/MITOGEN STUDY/
Total density[0002,1]=1711, # spots=251
   Study: PAT:1/REST/120 HRS/S35/4 HRS/27 HRS/MITOGEN STUDY/
Total density[0004,1]=12126, # spots=583
   Study: PAT:1/REST/120 HRS/S35/4 HRS/89 HRS/MITOGEN STUDY/
Total density[0005,1]=7320, # spots=513
   Study: PAT:1/PHA/120 HRS/S35/4 HRS/7 HRS/MITOGEN STUDY/
Total density[0033,1]=3004, # spots=341
   Study: PAT:4/REST/120 HRS/S35/4 HRS/164 HRS/MITOGEN STUDY/
Total density[0034,1]=6987, # spots=574
   Study: PAT:4/PHA/120 HRS/H3/4 HRS/29 HRS/MITOGEN STUDY/
Total density[0036,1]=1129, # spots=173
   Study: PAT:3/REST/120 HRS/H3/4 HRS/140 HRS/MITOGEN STUDY/
Total density[0051,1]=6593, # spots=596
   Study: PAT:3/PHA/120 HRS/H3/4 HRS/43 HRS/MITOGEN STUDY/
Total density[0057,2]=5560, # spots=595
   Study: PAT:4/PHA/120 HRS/H3/4 HRS/48 HRS/MITOGEN STUDY/
Total density[0073,1]=6136, # spots=602
   Study: PAT:1/PHA/120 HRS/S35/4 HRS/23 HRS/MITOGEN STUDY/
Total density[0074,1]=3162, # spots=514
   Study: PAT:1/REST/120 HRS/S35/4 HRS/240 HRS/MITOGEN STUDY/
Total density[0069,2]=2320, # spots=329
   Study: PAT:4/REST/120 HRS/H3/4 HRS/312 HRS/MITOGEN STUDY/

R-spot[  1] ACC#0054,1[312] (X,Y)abs=(346,219)Mn D= .60 SD= .32 -  spots=10
 ACC#[Index]C  %Dens  pACC#[Index]  D'   Lbl LM  DP  DL  Dx   Dy   Xabs Yabs
------------- ------- ------------ ---- --- -- ---- -- ---- ---- ---- ----
0002,1[ 207]2   1.01% 0054,1[ 312]  17.2  PP A    4.1 12 ( -6, -7) (246,187)
0074,1[ 361]2    .96% 0054,1[ 312]  30.4  PP A    2.0 14 ( -5,-13) (359,222)
0005,1[ 373]1    .89% 0054,1[ 312]  65.4  SP A    4.1 12 ( -6, -7) (260,192)
0001,1[ 209]1    .89% 0054,1[ 312]  15.8  PP A    3.2 12 ( -6, -8) (214,207)
0034,1[ 399]1    .66% 0054,1[ 312]  46.0  PP A    2.2 14 ( -7,-12) (298,176)
0073,1[ 452]1    .61% 0054,1[ 312]  37.7  PP A     .0 12 ( -5,-11) (373,229)
0033,1[ 278]2    .40% 0054,1[ 312]  12.0  PP A    3.2 13 ( -8,-10) (276,197)
0069,2[ 261]2    .21% 0054,1[ 312]   4.9  SP A    1.0 12 ( -4,-11) (338,223)
0054,1[ 312]1    .21% 0001,1[ 209]   6.7  PP A    3.2 12 ( -5,-11) (346,219)
0057,2[ 465]1    .14% 0054,1[ 312]   7.9  SP A    3.0 12 ( -2,-11) (376,198)

R-spot[ 41] ACC#0054,1[224] (X,Y)abs=(354,188)Mn D= .86 SD= .29 #  spots=12
 ACC#[Index]C  %Dens  pACC#[Index]  D'   Lbl LM  DP  DL  Dx   Dy   Xabs Yabs
------------- ------- ------------ ---- --- -- ---- -- ---- ---- ---- ----
0033,1[ 199]2   1.23% 0054,1[ 224]  36.9  SP D*   .0  0 (  0,  0) (287,158)
0034,1[ 289]1   1.20% 0054,1[ 224]  83.9  SP D*   .0  0 (  0,  0) (308,132)
0069,2[ 185]2   1.19% 0054,1[ 224]  27.6  SP D*   .0  0 (  0,  0) (344,194)
0036,1[  69]2   1.15% 0054,1[ 224]  13.0  SP D*   .0  0 (  0,  0) (289,185)
0001,1[ 157]1   1.04% 0054,1[ 224]  18.5  SP D*   .0  0 (  0,  0) (222,180)
0074,1[ 260]2    .96% 0054,1[ 224]  30.4  SP D*   .0  0 (  0,  0) (367,190)
0057,2[ 363]1    .77% 0054,1[ 224]  42.7  SP D*   .0  0 (  0,  0) (382,171)
0073,1[ 357]1    .65% 0054,1[ 224]  39.8  SP D*   .0  0 (  0,  0) (360,200)
0054,1[ 224]1    .61% 0001,1[ 157]  19.9  SP D*   .0  0 (  0,  0) (354,188)
0004,1[ 288]2    .60% 0054,1[ 224]  72.7  SP D*   .0  0 (  0,  0) (279,170)
0005,1[ 287]1    .53% 0054,1[ 224]  38.8  SP D*   .0  0 (  0,  0) (273,167)
0051,1[ 382]1    .42% 0054,1[ 224]  27.5  SP D*   .0  0 (  0,  0) (395,212)

R-spot[119] ACC#0054,1[248] (X,Y)abs=(370,196)Mn D= .53 SD= .34 #  spots=12
 ACC#[Index]C  %Dens  pACC#[Index]  D'   Lbl LM  DP  DL  Dx   Dy   Xabs Yabs
------------- ------- ------------ ---- --- -- ---- -- ---- ---- ---- ----
0036,1[  76]2   1.29% 0054,1[ 248]  14.6  PP V    3.0 17 (  6,-16) (308,190)
0004,1[ 290]2    .98% 0054,1[ 248] 119.2  PP V    5.4 16 ( 11,-11) (300,173)
0002,1[ 158]2    .80% 0054,1[ 248]  13.6  PP V    6.1 12 (  5,-11) (276,162)
0069,2[ 209]2    .67% 0054,1[ 248]  15.6  PP V    1.0 14 (  6,-12) (361,200)
0033,1[ 215]2    .55% 0054,1[ 248]  16.6  PP V     .0 14 (  6,-13) (306,166)
0034,1[ 309]1    .45% 0054,1[ 248]  31.5  PP V    5.0 19 (  6,-18) (330,140)
0057,2[ 370]1    .42% 0054,1[ 246]  23.5  PP V    2.0 15 (  8,-13) (399,173)
0074,1[ 277]2    .31% 0054,1[ 248]   9.9  PP V     .0 14 (  6,-13) (365,197)
0001,1[ 168]1    .24% 0054,1[ 248]   4.3  PP V    1.0 15 (  6,-14) (243,184)
0005,1[ 290]1    .23% 0054,1[ 248]  17.1  PP V    3.2 15 (  9,-12) (292,168)
0073,1[ 369]1    .19% 0054,1[ 248]  11.8  PP V    2.0 16 (  6,-15) (396,205)
0054,1[ 248]1    .17% 0001,1[ 168]   5.6  PP V    1.0 15 (  6,-13) (370,196)
```

A stimulated CGL data base for lymphocytes, containing the R-gel, is gel 54.1, containing 400 R-spots consisting of SP and PP spot pairs. Correspondences to R-spot [1] are missing in gels 4.1.

TABLE IIIb

EXAMPLES OF RATIO OUTPUT MODES OF THE CGL DATA BASE

R-spot Ratio list: 2 3 28 43 44 48 50 53 86 96 98 103 105 116 118 121 128

R-spot[1] ACC#0054.1[312] (X,Y)abs=(346,219)Mn D= 4.90 SD= 2.72 # spots=10

ACC#[Index]C	RDens	pACC#[Index]	D'	Lbl	LM	DP	DL	Dx	Dy	Xabs	Yabs
0002.1[207]2	6.16R	0054.1[312]	17.2	PP	A	4.1	12	(-6, -7)		(346,167)	
0074.1[361]2	7.95R	0054.1[312]	30.4	PP	A	2.0	14	(-5,-13)		(359,222)	
0005.1[373]1	7.88R	0054.1[312]	65.4	SP	A	4.1	12	(-6, -7)		(260,192)	
0073.1[452]1	7.04R	0054.1[312]	37.7	PP	A	.0	12	(-5,-11)		(373,229)	
0034.1[399]1	5.46R	0054.1[312]	46.0	PP	A	2.2	14	(-7,-12)		(298,176)	
0001.1[209]1	4.97R	0054.1[312]	15.8	PP	A	3.2	12	(-6, -8)		(214,207)	
0033.1[278]2	3.43R	0054.1[312]	12.0	PP	A	3.2	13	(-8,-10)		(276,197)	
0054.1[312]1	1.45R	0001.1[209]	6.7	PP	A	3.2	12	(-5,-11)		(346,219)	
0057.2[465]1	1.43R	0054.1[312]	7.9	SP	A	3.0	12	(-2,-11)		(376,196)	
0069.2[261]2	1.16R	0054.1[312]	4.9	SP	A	1.0	12	(-4,-11)		(338,223)	

R-spot[41] ACC#0054.1[224] (X,Y)abs=(354,188)Mn D= 7.23 SD= 1.84 # spots=12

ACC#[Index]C	RDens	pACC#[Index]	D'	Lbl	LM	DP	DL	Dx	Dy	Xabs	Yabs
0033.1[199]2	10.54R	0054.1[224]	36.9	SP	D*	.0	0	(0,	0)	(287,156)	
0034.1[289]1	9.96R	0054.1[224]	53.9	SP	D*	.0	0	(0,	0)	(308,132)	
0004.1[288]2	8.38R	0054.1[224]	72.7	SP	D*	.0	0	(0,	0)	(279,170)	
0074.1[260]2	7.95R	0054.1[224]	30.4	SP	D*	.0	0	(0,	0)	(367,190)	
0036.1[69]2	7.77R	0054.1[224]	13.0	SP	D*	.0	0	(0,	0)	(289,185)	
0057.2[363]1	7.72R	0054.1[224]	42.7	SP	D*	.0	0	(0,	0)	(382,171)	
0073.1[357]1	7.44R	0054.1[224]	39.8	SP	D*	.0	0	(0,	0)	(380,200)	
0069.2[185]2	6.63R	0054.1[224]	27.6	SP	D*	.0	0	(0,	0)	(344,194)	
0001.1[157]1	5.82R	0054.1[224]	18.5	SP	D*	.0	0	(0,	0)	(222,180)	
0051.1[362]1	5.59R	0054.1[224]	27.5	SP	D*	.0	0	(0,	0)	(395,212)	
0005.1[287]1	4.68R	0054.1[224]	38.8	SP	D*	.0	0	(0,	0)	(273,167)	
0054.1[224]1	4.32R	0001.1[157]	19.9	SP	D*	.0	0	(0,	0)	(354,188)	

R-spot[119] ACC#0054.1[248] (X,Y)abs=(370,196)Mn D= 4.57 SD= 3.46 # spots=12

ACC#[Index]C	RDens	pACC#[Index]	D'	Lbl	LM	DP	DL	Dx	Dy	Xabs	Yabs
0004.1[290]2	13.74R	0054.1[248]	119.2	PP	V	5.4	16	(11,-11)		(300,173)	
0036.1[78]2	8.72R	0054.1[248]	14.6	PP	V	3.0	17	(6,-16)		(308,190)	
0002.1[158]2	6.45R	0054.1[248]	13.6	PP	V	6.1	12	(5,-11)		(276,162)	
0033.1[215]2	4.74R	0054.1[248]	16.6	PP	V	.0	14	(6,-13)		(306,166)	
0057.2[370]1	4.25R	0054.1[248]	23.5	PP	V	2.0	15	(8,-13)		(399,173)	
0069.2[209]2	3.75R	0054.1[248]	15.6	PP	V	1.0	14	(6,-12)		(361,200)	
0034.1[309]1	3.74R	0054.1[248]	31.5	PP	V	5.0	19	(6,-18)		(330,140)	
0074.1[277]2	2.59R	0054.1[248]	9.9	PP	V	.0	14	(6,-13)		(365,197)	
0073.1[369]1	2.20R	0054.1[248]	11.8	PP	V	2.0	16	(6,-15)		(396,205)	
0005.1[290]1	2.06R	0054.1[248]	17.1	PP	V	3.2	15	(9,-12)		(292,168)	
0001.1[168]1	1.35R	0054.1[248]	4.3	PP	V	1.0	15	(6,-14)		(243,184)	
0054.1[248]1	1.22R	0001.1[168]	5.6	PP	V	1.0	15	(6,-13)		(370,196)	

The same spots as in Table IIIa but ordered and listed by ratio density relative to spots present in all 13 gels (in terms of 100%) and reordered by rank. Notice that after normalization and reordering, R-spot [41] now shows significant difference (as well as R-spot [119]) in the density differences between the two classes. R-spot [1] still does not show significant difference between the two classes.

36.1, and 51.1. A spot's %Dens is its density relative to all spots (including AP and US) in a gel. *Dx* and *Dy* are the spots position relative to its associated landmark. The (*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 interpretation of 1 = PHA stimulated and 2 = resting. The pACC#[index] refers to the spot paired with the list entry (ACC#[index]). *D'* is the background corrected absolute density of the spot. The (*X*abs, *Y*abs) is the absolute position of the spot in the gel image. The Lbl is the pairing label SP, PP, AP, or US. Note that the heuristic pairing values *DP* (distance between spots in a pair) and *DL* (distance from a pair to the landmark spot) are similar for most spots as are the (*Dx*, *Dy*) relative distances to the landmark spot. Because of this 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 [41] 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 [119] would

TABLE IV

EXAMPLES OF GEL.ID 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
PHASPT,CA = PHA E-SPOT FILE
        .
        .
        .
0001.1/PAT:1/-/-/10-27-78/#7/PHA/3:10, 5-20%/
120 HRS/S35/4 HRS/3.2 HRS/MITOGEN STUDY/
E00001/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*
 34 58 79 102 121 139 156 169 183 193 200 207 212 0 0 0 3 366 61 327
0001.2/PAT:1/-/-/10-27-78/#7/PHA/3:10, 5-20%/
120 HRS/S35/4 HRS/24 HRS/MITOGEN STUDY/
E00005/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*

0002.1/PAT:1/-/-/10-27-78/#4/RESTING/3:10, 5-20%/
120 HRS/S35/4 HRS/27 HRS/MITOGEN STUDY/
E00009/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*
 34 58 80 103 122 140 158 171 184 193 201 208 212 0 0 0 25 355 55 306
0003.1/PAT:1/-/-/10-23-78/#5/RESTING/3:10, 5-20%/
120 HRS/S35/4 HRS/24 HRS/MITOGEN STUDY/
E00013/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*

0003.2/PAT:1/-/-/10-23-78/#5/RESTING/3:10, 5-20%/
120 HRS/S35/4 HRS/72 HRS/MITOGEN STUDY/
E00017/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*

0004.1/PAT:1/-/-/10-31-78/#6/RESTING/3:10, 5-20%/
120 HRS/S35/4 HRS/89 HRS/MITOGEN STUDY/
E00021/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*
 34 58 80 103 123 141 158 171 184 194 201 208 212 0 0 0 42 422 65 306
0005.1/PAT:1/-/-/10-31-78/#9/PHA/3:10, 5-20%/
120 HRS/S35/4 HRS/7 HRS/MITOGEN STUDY/
E00025/Q011/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*
 47 74 103 126 144 160 176 188 199 205 215 0 0 0 0 0 30 407 43 320
0006.1/PAT:1/-/-/10-23-78/#8/PHA/3:10, 5-20%/
120 HRS/S35/4 HRS/5 HRS/MITOGEN STUDY/
E00029/R230/-NONE-/VIDICON-MAN,28MM,F8,69CM OR 250 MICRONS/LESTER*
```

Data necessary for some CGEL operations are exemplified below. These accession file data are characterized as follows. Each data record is four lines. The first four lines of the file define the record field descriptors which are separated by "/" and terminated with a "*." The fourth line of a record is the set of gray value peaks corresponding to the ND wedge calibration if it exists. The last four numbers of that line are the computing window for that gel [$x1{:}x2$, $y1{:}y2$] if the window exists. No calibration exists for gels 1.2, 3.1, 3.2, and 6.1 in this example. The blood samples were obtained from normal patients in a blood bank.

The SET LABEL command restricts pairing labels to any combination of: S (sure pair), P (possible pair), A (ambiguous pair), U (unresolved spot), and * (landmark spot). The default option is P and S. Thus pairing certainty may be used to partition the CGL data base.

The data base may alternatively be restricted to a subset of the gels by the SET WORKING GELS command. This removes one or more gels from immediate consideration reconfiguring the data base to the remaining gels.

Searches and data base analyses performed on the CGEL data base are done relative to a particular R-gel. It is possible to construct several independent data bases simultaneously in the same CGEL core image with different R-gels

as defined by the CMPGEL gel pairing program (*2*). The current R-gel name determines which one is accessible at any given time and may be changed using the SET RGEL command.

The SET CLASSES command gives wide flexibility in naming gel classes, and in partitioning the set of gels into as many as nine classes either automatically (using accession file data via the SET FORMAT command) or manually. During an interactive session, application of these commands may be done repetitively to redefine the gel partition by class.

The PLOT command generates a log density–log density scatter plot essential for evaluation of gel-to-gel comparability (*4*). The result of its use is a graphics display and/or a plotter file with a ".PLT" extension. Included in the output is the number of spots in common to both gels as well as a statistical measure of the Euclidean distance of each point from the line of identity, i.e.,

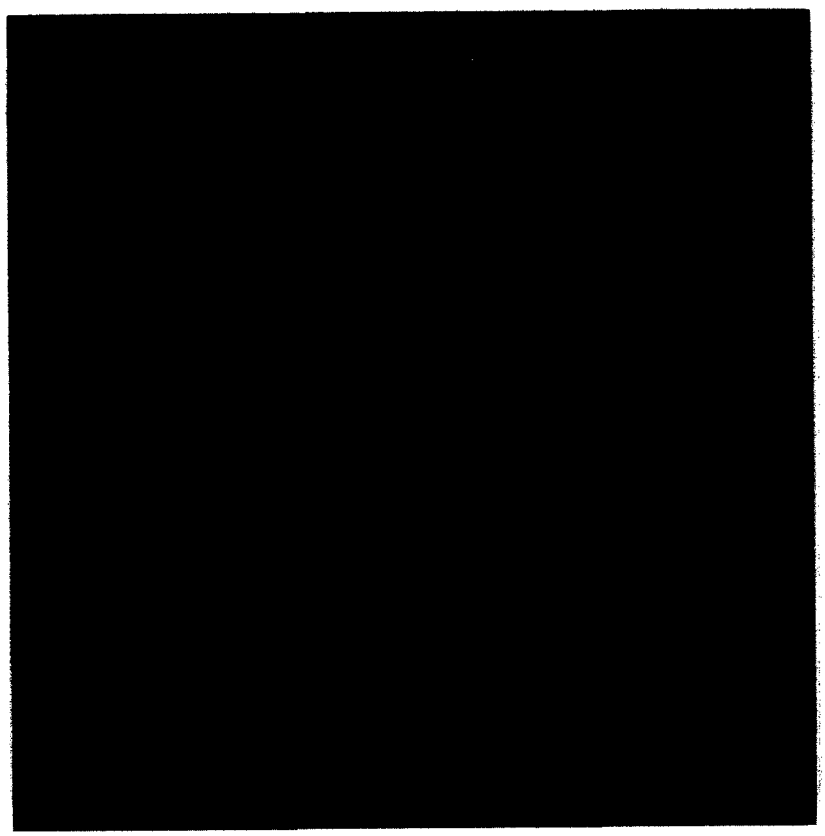

FIG. 4. A typical scatter plot for the same gel but with different autoradiographic exposure times. Gel 32.3 is a 13-hr exposure and gel 32.4 is a 24-hr exposure. The R-spot data base was normalized by the sum of the densities of the landmark spots. Note that most data clusters around the 45° line with some outlier errors attributable to the variance in the autoradiograph generation—scanning–segmentation (and possibly pairing alogirthm) process. The mean variation for these gels is 0.005. The log–log scale ranges from 0.1 to 20.0 with decades denoted by the largest scale markers.

the mean variation defined as the square root of the sum of the squares of the distances from the 45° line divided by the number of spots (cf. Fig. 4).

Three types of tables may be generated and both printed on the user's teletype as well as being saved in a file (with a ".TBL" file extension) using the TABULATE command. The first type of table is an upper diagonal mean variation matrix (with associated upper diagonal spot pair count matrix) for all gels in the CGL data base. This is illustrated in Table V for a set of PHA-stimulated lymphocyte gels. A small number of selected R-spots (indicated either manually or as the result of a search to be described) may be plotted in a

TABLE V

MEAN VARIATION TABLE FOR SET OF PHA GELS

```
File: P5NMNV.TBL 06/25/1980,  12:17:24 AM
[0054,1] study: PAT:3 /PHA /120 HRS /H3 /4 HRS /21 HRS /MITOGENS/
[0001,1] study: PAT:1 /PHA /120 HRS /S35 /4 HRS /3.2 HRS /MITOGENS/
[0002,1] study: PAT:1 /REST /120 HRS /S35 /4 HRS /27 HRS /MITOGENS/
[0004,1] study: PAT:1 /REST /120 HRS /S35 /4 HRS /89 HRS /MITOGENS/
[0005,1] study: PAT:1 /PHA /120 HRS /S35 /4 HRS /7 HRS /MITOGENS/
[0033,1] study: PAT:4 /REST /120 HRS /S35 /4 HRS /164 HRS /MITOGENS/
[0034,1] study: PAT:4 /PHA /120 HRS /H3 /4 HRS /29 HRS /MITOGENS/
[0036,1] study: PAT:3 /REST /120 HRS /H3 /4 HRS /140 HRS /MITOGENS/
[0051,1] study: PAT:3 /PHA /120 HRS /H3 /4 HRS /43 HRS /MITOGENS/
[0057,2] study: PAT:4 /PHA /120 HRS /H3 /4 HRS /48 HRS /MITOGENS/
[0073,1] study: PAT:1 /PHA /120 HRS /S35 /4 HRS /23 HRS /MITOGENS/
[0074,1] study: PAT:1 /REST /120 HRS /S35 /4 HRS /240 HRS /MITOGENS/
[0069,2] study: PAT:4 /REST /120 HRS /H3 /4 HRS /312 HRS /MITOGENS/
Mean Variation for gels in data base, Labels: (PS)
R-spots Ratio list: 2 3 29 43 44 48 50 53 86 96 98 103 105 116 118 121 128

      |54,1|01,1|02,1|04,1|05,1|33,1|34,1|36,1|51,1|57,2|73,1|74,1|69,2
------------------------------------------------------------------------
0054,1|     |.271|.362|.207|.158|.163|.126|.336|.163|.097|.164|.157|.131
0001,1|     |    |.427|.459|.213|.281|.213|.447|.360|.318|.216|.197|.360
0002,1|     |    |    |.432|.395|.370|.291|.470|.421|.386|.299|.361|.417
0004,1|     |    |    |    |.220|.335|.274|.510|.269|.224|.288|.299|.304
0005,1|     |    |    |    |    |.214|.152|.369|.231|.190|.126|.167|.240
0033,1|     |    |    |    |    |    |.150|.249|.270|.210|.205|.162|.208
0034,1|     |    |    |    |    |    |    |.348|.167|.159|.139|.137|.188
0036,1|     |    |    |    |    |    |    |    |.468|.438|.372|.283|.384
0051,1|     |    |    |    |    |    |    |    |    |.179|.196|.217|.211
0057,2|     |    |    |    |    |    |    |    |    |    |.193|.196|.166
0073,1|     |    |    |    |    |    |    |    |    |    |    |.123|.210
0074,1|     |    |    |    |    |    |    |    |    |    |    |    |.176
0069,2|     |    |    |    |    |    |    |    |    |    |    |    |

      |54,1|01,1|02,1|04,1|05,1|33,1|34,1|36,1|51,1|57,2|73,1|74,1|69,2
------------------------------------------------------------------------
0054,1|     | 140| 152| 219| 236| 182| 228| 107| 164| 290| 188| 198| 180
0001,1|     |    |  67|  99| 103|  81|  99|  50|  94| 111| 102| 103|  94
0002,1|     |    |    | 118| 115|  80| 105|  51|  85| 109|  92|  94|  91
0004,1|     |    |    |    | 164| 108| 156|  67| 124| 175| 132| 127| 121
0005,1|     |    |    |    |    | 115| 150|  69| 128| 184| 137| 150| 122
0033,1|     |    |    |    |    |    | 132|  81| 103| 141| 111| 116| 104
0034,1|     |    |    |    |    |    |    |  77| 141| 185| 149| 153| 127
0036,1|     |    |    |    |    |    |    |    |  61|  82|  67|  63|  62
0051,1|     |    |    |    |    |    |    |    |    | 146| 137| 140| 112
0057,2|     |    |    |    |    |    |    |    |    |    | 156| 159| 145
0073,1|     |    |    |    |    |    |    |    |    |    |    | 151| 125
0074,1|     |    |    |    |    |    |    |    |    |    |    |    | 130
0069,2|     |    |    |    |    |    |    |    |    |    |    |    |
```

The mean variation table was computed for the set of lymphocyte gels used in the PHA stimulation experiment. The mean variation is computed for each pair of gels taken two at a time. The bottom table is the number of R-spot pairs which were common to both gels.

rank-order table with density as the ordinate. The normalization may be performed implicitly (e.g., using the data base normalization to be discussed) or explicitly using a list of R-spots requested from the user). Eight spots fit across the teletype output page, so that for more than eight spots, multiple pages are used as required. Table VI illustrates a rank-order table for some selected R-spots from the PHA data.

Data Base Normalization. A given amount of polypeptide concentrated in a small spot will sometimes yield a lower total density value than if the same quantity of material were spread over a wider area. This is because the various

TABLE VI

RANK-ORDER DENSITY TABLE FOR SELECTED SPOTS

```
File: PHA5Nw.TBL 06/25/1980,  03:15:18 PM
RANK-ORDER table: <ACC#>&<LMset>&<Class #>
R-spots Ratio list: 2 3 28 43 44 48 50 53 86 96 98 103 105 116 118 121 128

   3.3 |            0004,1G2
   3.2 |  0004,1A2           0004,1I2
   3.1 |
   3.0 |
   2.9 |
   2.8 |                     0036,1I2
   2.7 |
   2.6 |                                0033,1N2
   2.5 |                                0004,1N2
   2.4 |
   2.3 |                                0069,2N2
   2.2 |
   2.1 |            0069,2G2           0074,1N2
       |                                0054,1N1
   2.0 |
   1.9 |
   1.8 |
   1.7 |                                          0004,1S2
   1.6 |                                0057,2N1
   1.5 |
   1.4 |            0005,1G1 0057,2I1 0036,1N2
   1.3 |
   1.2 |  0074,1A2                      0034,1N1
   1.1 |                     0074,1I2
   1.0 |                     0002,1I2 0073,1N1
    .9 |  0005,1A1 0002,1G2 0054,1I1           0074,1S2
       |           0073,1G1
    .8 |  0034,1A1 0057,2G1
    .7 |           0054,1G1 0005,1I1 0005,1N1
    .6 |                     0051,1I1 0002,1N2 0073,1S1
       |                                0051,1N1
    .5 |  0073,1A1
    .4 |  0001,1A1 0001,1G1 0001,1I1           0057,2S1
       |  0051,1A1                             0054,1S1
    .3 |  0057,2A1           0034,1I1          0005,1S1
       |  0054,1A1           0073,1I1
    .2 |                                       0034,1S1
    .1 |                                0001,1N1 0001,1S1
-------------------------------------------------------
R-spot |         8       65       79       103       112
Class #  1=PHA, 2=REST
```

A rank-order density table may be constructed for selected spots. Five spots were selected from the PHA-stimulated lymphocyte gel data base. R-spots [79] and [103] were statistically significant at 10% in the rank-order test search. 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.

stains and autoradiograph preparations follow one or another form of the usual "S-shaped" gamma curve. This curve saturates at the high-exposure end and has a "toe" of minimum exposure in the beginning of the curve. In addition, the Schwartzschild–Villager effect causes high-density areas to be underestimated. Even if the image were linear and noise-free, the digitizing device introduces other sources of noise. Therefore, a spot's density value should not be taken as an absolute, but should probably be obtained over a set of duplicate scans and duplicate gels getting a measure of its variance (as suggested by (*4*)).

Intergel density variation makes some normalization scheme necessary. The density data initially transmitted to CGEL are already normalized with respect to total gel density expressed as a percentage. But this is not always satisfactory, hence two other normalization modes are available. First, one may normalize the CGEL data base by a subset of well-defined spots common to all gels or selected for some particular reason. Once this subset of spots is specified (one way is by the search procedures to be described), they may be used to specify the spot ratio list using the SET RATIO LIST command. The total D' (absolute spot density corrected for gel background) of the sum of these spots for each gel is computed and saved.

Using the REORDER command, the data base may be reordered, R-spot set by R-spot set, based on the current density data mode. This is often most useful when the ratio spot list has been specified and the density mode set to "ratio." Each R-spot set will be reordered with the highest density first. If the density data mode is changed without later reordering the data base, the data base will reflect the new density mode interpretation of the data but with the previous density data modes ordering. For example, setting the date mode to absolute density after the data base is first constructed will have the R-spot sets ordered by percentage density.

CGEL Data Base Searching and Investigation. Extraction of information from the completed CGL data base requires interrogation using the INQUIRE command. This command includes a group of subcommands which are detailed below. They permit both printing a few R-spot sets and as well as searching the data base. Table VII lists the various subcommands of the INQUIRE command.

One of various 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 is a composite tabulation of R-spots selected by the current search.

The FIND subcommand searches the R-spot data base for R-spot sets where the R-gel spot is a landmark spot. The resulting list of spots is reported as in Table VIIIa and saved in the search results list.

The INDEX search subcommand finds R-spots meeting all of the statistical limits for mean relative distance of a R-spot set from the landmark spot, mean *DP*, mean *DL*, mean R-spot set density, standard deviation of R-spot set density, and minimum R-spot set size. The *DP* (distance between spots in a

TABLE VII

CGEL INQUIRE COMMANDS

```
Print R-spot set i or # (search results list of R-spots specified through
        search process or explicitly).

List landmark set j, where j is 'A' through 'Z'.

Find the R-spot indices of all R-gel LM sets.

Index search for R-spots meeting statistical limits.

Search for R-spot sets meeting statistical limits.

T-test search for R-spot sets with a given confidence limits
        between classes.

Rank order search for R-spot sets with a given significance
        between two classes (Wilcoxson-Mann-Whitney test).

Kruskol-Wallis rank order search for R-spot sets with a given
        significance between all (up to 9) classes for a minimum
        of 5 gels/class.
```

The CGEL "INQUIRE" command subcommands are available for conducting a search throughout the CGL data base. R-spots may be printed and a subset of R-spots found based on a search using various statistical criteria. The "Index" and "Search" commands find spots meeting feature range criteria of (a) relative distance from a landmark, (b) *DL*—distance to a landmark, (c) *DP*—distance between spots in a pair, (d) mean density in a R-spot set, and (e) standard deviation of this density. The *t* test assumes a bimodal distribution of spot density in a R-spot set which is not necessarily the case. The rank-order tests are distribution free. The "search results list" of R-spots resulting from any of these statistical searches may be used in other parts of the CGEL system.

pair) and *DL* (distance from the landmark spot to the pair) features were discussed (*2*).

The SEARCH subcommand performs the same test as INDEX but prints the actual R-spot set instead of just the first line of each R-spot set meeting the statistical sizing criteria. The latter two tests can be useful for finding R-spot sets which: (1) are complete in having all spots present, or (2) have primarily dark or primarily light spots, or (3) consist of spots with high or low variance. In addition to printing these spots, their indices are saved in the search results list. These sizing limits are changed using the SET STATISTICS top level command which has the following type of dialogue with answers italic:

Relative distance limits are [.00, 512.00]: *0,30*
DL limits are [.00, 512.00]: *0,25*
DP limits are [.00, 512.00]: *0,15*
MN density limits are [.00, 100.00]: *0,300*
S.D. density limits are [.00, 100.00]: *0,50*
Class difference t-Test confidence or Rank order significance limit (1%, 5%, 10%, 20%) is 10%: *1%*
Check if size of R-spot set = # of working set gels (Y/N)? (N): *N*

The T-TEST subcommand may be used, with the specified class difference confidence limit, to find R-spots statistically different in the search through the

TABLE VIII

EXAMPLE OF R-SPOT SET SEARCHES

```
a. Landmark set constraint search (FIND subcommand)

        LM[ A ]=R-spot[   3]
        LM[ B ]=R-spot[  28]
        LM[ C ]=R-spot[  37]
        LM[ D ]=R-spot[  41]
                .
                .
                .
        LM[ V ]=R-spot[ 121]
        LM[ W ]=R-spot[ 128]
        LM[ X ]=R-spot[ 139]



b. T-test constraint search (T-TEST subcommand)

R-spot[ 18] ACC#0054.1[128] (X,Y)abs=(328,147)Mn D= 2.12 SD= 1.26 # spots=11
[  18](m1,m2)= 1.33, 3.08, Lim1[  .93: 1.72], Lim2[ 1.96: 4.20]

R-spot[ 47] ACC#0054.1[241] (X,Y)abs=(258,195)Mn D= 2.89 SD= 2.60 # spots=12
[  47](m1,m2)= 1.31, 5.11, Lim1[  .52: 2.11], Lim2[ 2.89: 7.33]
        .
        .
        .



c. Rank order constraint search (RANK ORDER subcommand)

R-spot[ 44] ACC#0054.1[243] (X,Y)abs=(352,196)Mn D= 3.54 SD= 1.52 # gels=13
n1=   6 n2=   7 n=  13 R=  21 R'=  63 Ralpha=  24

R-spot[ 73] ACC#0054.1[302] (X,Y)abs=(198,216)Mn D= 4.98 SD= 2.41 # gels=12
n1=   5 n2=   7 n=  12 R=  15 R'=  50 Ralpha=  16
```

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

CGL data base according to the two-tailed *t* test (*5*). The search also assumes that the gels have been partitioned into classes using the SET CLASSES command. The program then requests the names of the two classes to be compared in the search. The spots found are put into the search results list and the R-spot set header and values computed for the *t* test are printed. Table VIIIb has an example of some of this output. The ($m1, m2$) are the means of the spots in the two subsets of spots in the particular R-spot set and the Lim1 and Lim2 are the corresponding limits computed using the standard two-tailed *t*-test calculations.

A rank-order test is used in the search for R-spot sets with a given significance between two classes (Wilcoxson–Mann–Whitney test, WMW) and is invoked by the RANK subcommand (*5*). It, as the *t* test, is a two-class test and must have the gels partitioned into two or more classes and the significance limit set. It prints the R-spot set header line and a line of information on the

WMW test statistics. It too puts spots found into the search results list. Table VIIIc illustrates this output. The parameters $n1$, $n2$ are the number of gels in each of the two classes and the $R(R')$ are the rank sums of the smaller (larger) of these two classes. Ralpha is the table value of the rank sum for $n1$, $n2$.

The Kruskol–Wallis rank-order test is used in a search for R-spot sets with a given significance among all (up to nine) classes for a minimum of five gels per class is also available (*5*). This test is invoked by the KRUSKOL subcommand and results in a new search results list as well as printing R-spot set headers and test results for significant spots.

The PRINT subcommand requests either a single R-spot number, a letter landmark spot name, or the "*" symbol. It then prints the R-spot set in the format illustrated by Tables IIIa,b for the specified R-spot set or the sets of R-spots from the search results list (if "*" was specified).

This completes the list of INQUIRE subcommands. We now return to the CGEL level for describing commands.

Use of the Search Results List. As mentioned previously, the search results list is a composite tabulation of spots selected by various invoked CGEL commands and is available either for further processing or for output. The SPSS command accepts either an explicit list of R-spot indices or the entire search results list and produces a data file for this subset of spots only. The search results list is indicated in such requests for a list of R-spot set names by the "*" symbol. The SPSS command then generates a numeric coded file of these R-spot sets such that the file could be read by the SPSS program (*6*), MLAB (*7*), or other statistical analysis packages. The file has an ".SPS" file name extension. Other GELLAB programs (MARKGEL and SEERSPOT—see below) use the SPSS file as part of their input (cf. Table IX).

Similarly, the TABULATE command can use the search results list to specify which spots to use in the rank-order table generation (as well as being able to specify the list manually). The ratio list may also be defined using the SET RATIO LIST command on the search results list.

Use of Statistical Checking in R-spot Set Operations. It is possible to invoke statistical limits checking at any point where a R-spot is being processed. If the set does not meet any one of the limits set by SET STATISTICS then the R-spot set will not be considered for the operation. This checking is performed automatically as part of the INDEX search and SEARCH subcommands. It may be turned on *for all operations* using the CGEL CHECKING command. Requesting CHECKING again will turn it off (i.e., a "toggle").

This completes the roster of major CGEL operations. Two auxiliary programs employing SPSS files as inputs are now described which produce derived images. These images facilitate the backchecking of any R-spot set in both a global (the R-map) and a local but multiple-gel (mosaic) context.

2.3. MARKGEL R-Map Image Generation Algorithm

The MARKGEL program takes an SPSS file produced by CGEL and

TABLE IX

EXAMPLE OF AN SPSS DATA FILE

File: P5N05S.SPS 06/25/1980, 12:17:03 AM

RSPOT#	ACC#	INDEX	%TOTD	TOTD	LABEL[0:3]	LMSET[1:26]	DP	DL	D1X	D1Y	XABS	YABS
44	0051.1	411	8.30	547.18	1	5 .0	0	0	0	386	220	
44	0057.2	384	7.70	428.16	1	5 .0	0	0	0	380	177	
44	0073.1	381	7.46	457.48	1	5 .0	0	0	0	377	208	
44	0034.1	312	6.97	486.70	1	5 .0	0	0	0	303	142	
44	0054.1	243	5.64	183.64	1	5 .0	0	0	0	352	196	
44	0005.1	307	5.07	371.44	1	5 .0	0	0	0	264	172	
44	0033.1	214	3.94	118.37	1	5 .0	0	0	0	283	166	
44	0001.1	173	3.87	68.75	1	5 .0	0	0	0	219	186	
44	0036.1	92	3.52	39.78	1	5 .0	0	0	0	285	193	
44	0074.1	265	3.30	104.21	1	5 .0	0	0	0	364	199	
44	0069.2	208	2.71	62.94	1	5 .0	0	0	0	341	201	
44	0002.1	169	1.47	25.17	1	5 .0	0	0	0	249	164	
44	0004.1	333	.68	82.45	1	5 .0	0	0	0	276	180	
47	0036.1	89	7.23	81.59	2	6 1.4	9	7	-6	203	194	
47	0069.2	214	7.03	163.19	2	6 1.0	9	8	-4	256	203	
47	0033.1	226	6.57	197.28	2	6 .0	9	8	-5	209	169	
47	0004.1	305	3.90	472.37	2	6 2.2	9	7	-3	200	174	
47	0051.1	393	3.70	244.09	2	6 1.0	10	8	-6	303	217	
47	0034.1	323	1.96	136.81	2	6 2.2	11	9	-7	216	146	
47	0005.1	306	1.10	80.29	2	6 2.0	9	6	-5	179	170	
47	0073.1	304	1.05	64.21	2	6 2.0	9	6	-5	298	209	
47	0002.1	171	.81	13.80	2	6 .0	9	8	-5	166	166	
47	0057.2	382	.63	35.18	2	6 1.0	10	9	-5	285	176	
47	0054.1	241	.63	20.48	2	6 7.2	16	8	-5	258	195	
47	0001.1	156	.13	2.24	2	6 7.2	16	12	-11	141	177	
73	0057.2	460	14.88	827.18	1	8 .0	0	0	0	231	198	
73	0051.1	468	10.84	714.82	1	8 .0	0	0	0	235	241	
73	0034.1	410	9.05	632.62	1	8 .0	0	0	0	145	181	
73	0054.1	302	8.73	283.94	1	8 .0	0	0	0	198	218	
73	0001.1	199	7.33	130.24	1	8 .0	0	0	0	57	200	
73	0073.1	447	7.14	437.99	1	8 .0	0	0	0	240	226	
73	0005.1	357	5.41	396.15	1	8 .0	0	0	0	131	188	
73	0033.1	276	4.11	123.52	1	8 .0	0	0	0	152	196	
73	0004.1	356	3.79	459.79	1	8 .0	0	0	0	149	186	
73	0036.1	133	3.70	41.80	1	8 .0	0	0	0	136	225	
73	0074.1	351	2.75	86.84	1	8 .0	0	0	0	210	219	
73	0069.2	251	2.33	54.03	1	8 .0	0	0	0	159	220	

This is an example of part of an SPSS file created for spots in the search results list found in a 10% rank order search for the normalized PHA-stimulated lymphocytes CGL data base.

generates an image file Mi.PIX (where "i" is the GEL.ID picture file number corresponding to the gel accession number) for all of the R-spots specified for that specified gel. The R-map starts with a copy of the specified gel image. Each R-spot to be labeled is then superimposed in the image by a white "+" in the center of the spot followed by a white R-spot number. The name of the SPSS file and current date is written at the top of the image. Figure 5 shows a typical R-map image of the R-gel in one set of PHA-stimulated lymphocyte gels. The spots selected were the result of applying the rank-order test in the search with a 10% confidence level.

2.4. SEERSPOT Mosaic R-Spot Image Generation

The SEERSPOT program uses an SPSS file produced by CGEL and generates one to three mosaic images from the set of original gel images consisting of regions containing the spot for the set of gels. Spots are selected

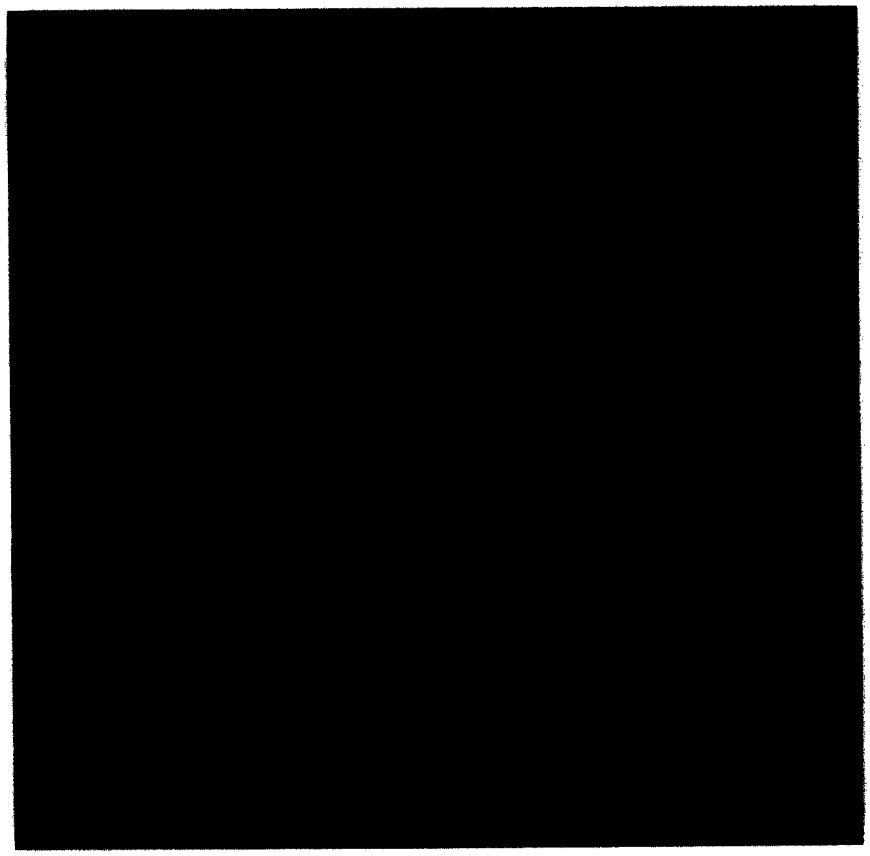

FIG. 5. A R-map image of a 10% rank-order search on normalized CGL data base was produced using the MARKGEL program for PHA-stimulated lymphocytes. The R-map was performed on the R-gel. The user may optionally add the gray value 50 to the image to enable the white labels to be seen clearly by specifying CORRECTBACKGROUND. There are three labeling options: NONUMBER—do not draw label numbers, just the "+" on the center of the spot; the second option is USELANDMARKS—if a R-spot is a landmark spot, use the letter rather than a number for it; the third and default option is to always label with a number.

for making mosaics using the results of the searches and R-maps. There may currently be up to 48 gels in up to three mosaic images. Figures 6a–d show some typical mosaic images generated from the PHA-stimulated lymphocyte gel data base. Figure 7 illustrates the image-mapping operation performed on the spot image subregions for the selected R-spot set.

The PHA gel data base used to illustrate this paper has 13 gels. However, R-spots 41, 73, and 119 in Figs. 6a, b, and d had only 12 gels present. We have performed estimations of where the spot would be in missing gels with good results. This first-order approximation is performed by adding the mean relative distance to the landmark (Dx,Dy) in the R-spot set to the absolute coordinates of the landmark for the missing gel.

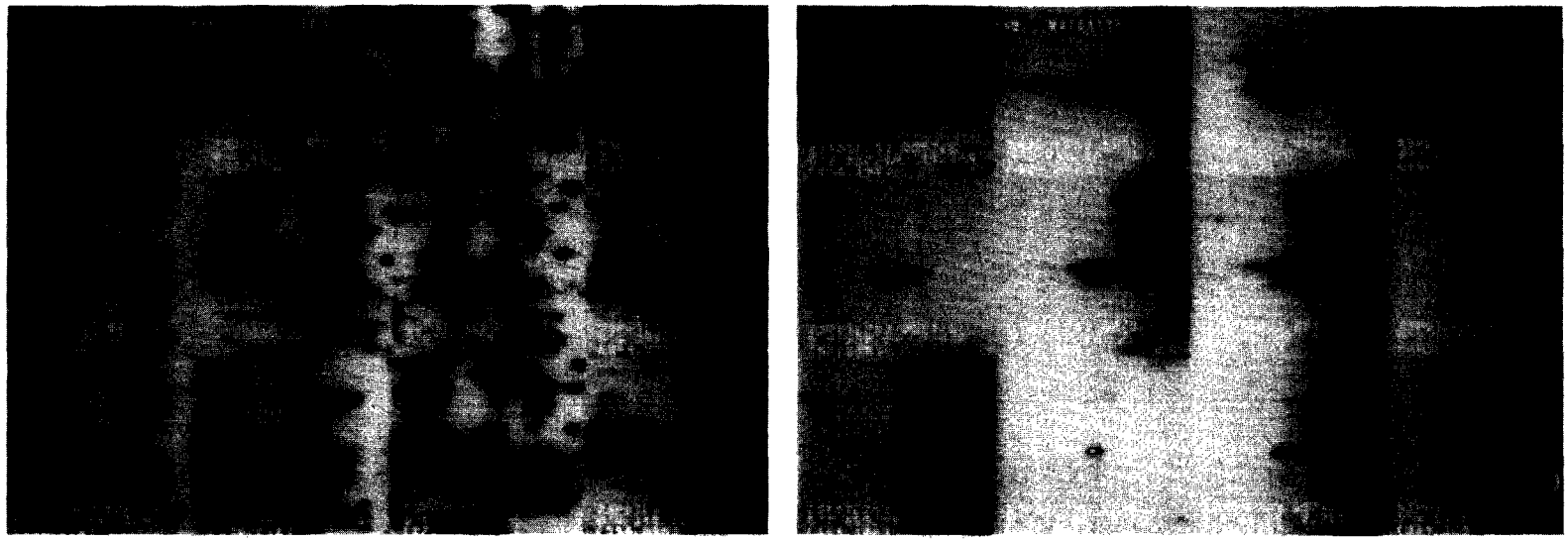

FIG. 6. Mosaic mapping is performed using the SEERSPOT program on a R-spot set of (PHA-stimulated lymphocyte gels) for displaying an ordered list of gels for a particular R-spot. Although, the spots are ordered in the mosaic by rank order in the R-spot set, they could be ordered by other criteria. (a) R-Spot [41] has a relative decrease with PHA, (b) Mosaic R-Spot [73] has a relative increase with PHA, (c) Mosaic R-Spot [103] has a relative decrease with PHA, (d) Mosaic R-Spot [119] has a relative decrease with PHA. The image has the R-spot region extracted from each corresponding image (magnified by a 2X zoom) inserted into 128 × 128 pixel subregions of the output image. At the bottom of each of these subregions is an accession number label and density information. The name of the SPSS file and current date are written at the top of the image. The R-spot itself has a 3 × 3 white "+" drawn in its center. Since the mean density of the images varies it would be difficult to display and photograph such a mosaic. Therefore, the normal mode of operation is to compute the density histogram of each subregion and then to adjust the subregion image by the difference between the peak value of the histogram and gray value 50 (in a range of 0 white to 255 black). It is assumed that the peak corresponds to background. All pixel gray values are correspondingly adjusted. This is done prior to image labeling. The subregions are ordered left to right top to bottom in the images according to the ordering of spots in the R-spot set. The default 2X zoom parameter may optionally be changed to 1X, 4X, or 8X by specifying "ZOOM : nX."

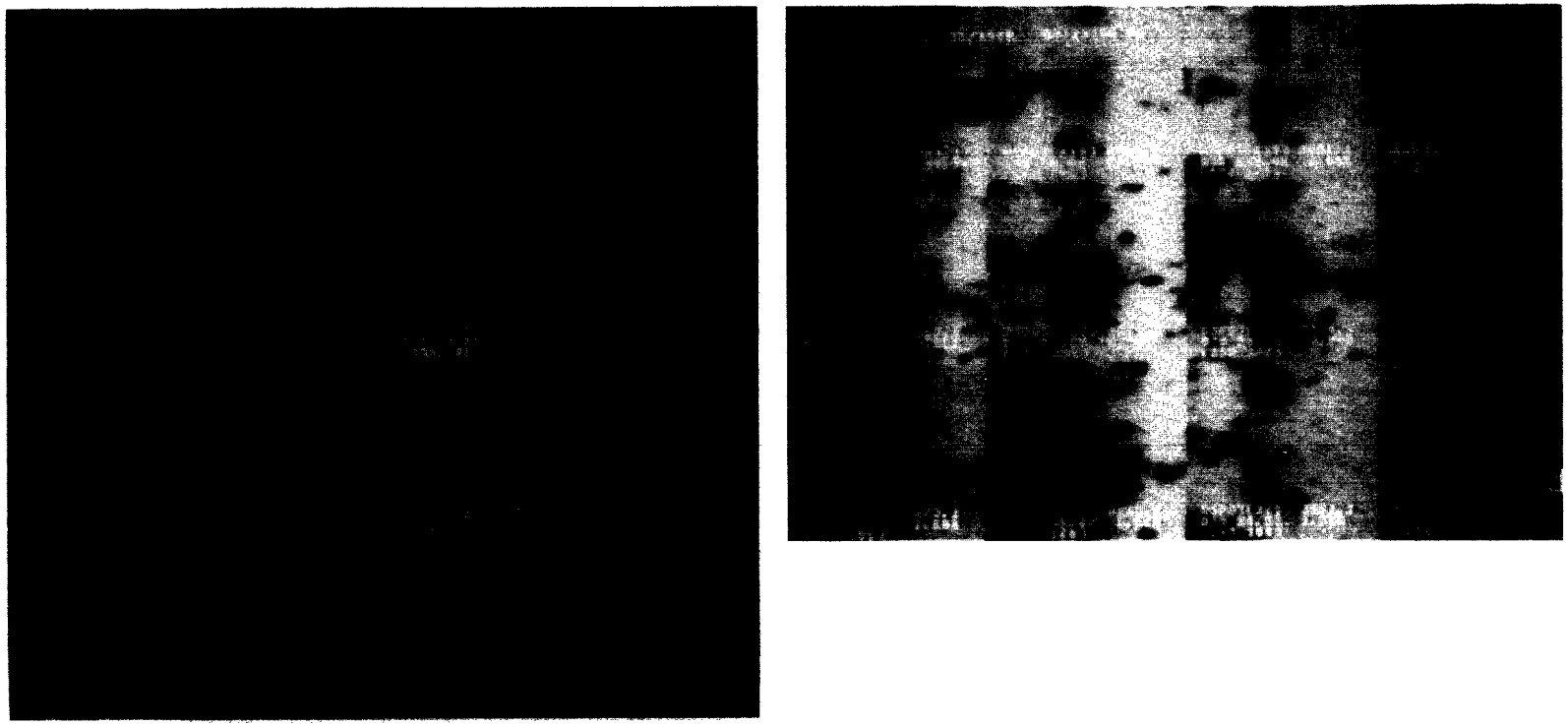

FIG. 6—*Continued*

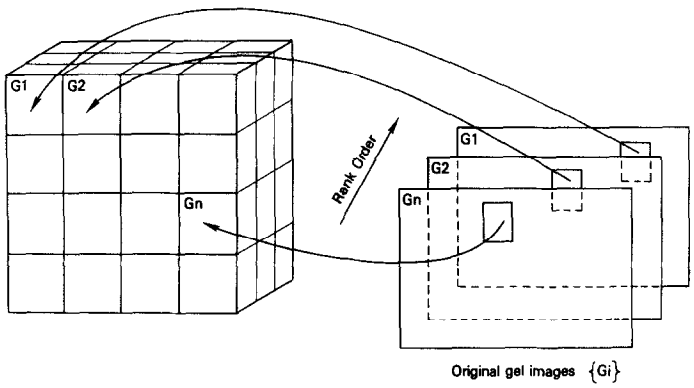

FIG. 7. Mosaic image mapping operation performed on the selected R-spot set. Spot subimages are ordered top to bottom and left to right by the spot order in the R-spot set. Subregions centered on the spot of interest of each of the gels are extracted and inserted in the mosaic image(s).

2.5. *E-Map Data Base For Selected Spots*

Exemplar spots which are found to be of interest in a particular gel may be saved in a system data base called the *exemplar spot file*. In general, an E-spot is a spot from a R-spot set—but it need not be obtained from a CGEL data base. Exemplar spots are represented by a 4-tuple: (a) the accession number of the gel where it was discovered to exist prominantly, (b) its (x,y) position within that image, (c) a two-letter spot code unique to a gel, and (d) a five-character search name identifier of the experimenter who found the spot. Formally, an E-spot is defined:

⟨2 letter spot code⟩⟨acc#⟩⟨search group⟩⟨x,y coord⟩.

For example:

CQ0002.2PHAOA330,170.

This notation facilitates the bookkeeping involved with recording interesting spots found to be present or missing in the various gels. A gel comparison table may be constructed with the names of the gels in the data base indexing the columns and an infinitely extensible set of exemplar spots indexing the rows. The gel comparison table entries currently used are given below. A question mark may be appended to denote difficult cases. Table Xa lists the labels used in the E-map. A spot exemplar map image (i.e., the exemplar spot locations are marked on a copy of the gel image) of a given gel can be generated and then used in searching other images.

The exemplar spot file is set up as an extensible file, SPT.SP, consisting of M exemplar spots in row entries and an N-entry column vector for up to N gels. Each entry in the row vector corresponds to one of the N gels. Thus the file constitutes a $N \times M$-exemplar spot sparse array (where null entries are denoted by "."). An example is illustrated in Table Xb.

TABLE X

E-MAP LABELING CONVENTION

a. Gel labeling convention.

```
. = not tested and thus has no entry
+ = present
- = missing
? = not sure whether + or -
P or (+?) = probably present
M or (-?) = probably missing
R = right shift
D = downward shift
U = upward shift
D = downward shift
L = left shift
B = blacker in density
W = whiter in density
I = two or more spots (i.e. multiple spots)
<#>= actual density value of the spot
```

b. Example of an E-map file

```
GEL CONDITION:                    B
                                  0 0 0 0 0
                                  1 2 4 4 4          ...
                                  2 0 2 2 2
                                  3 2 4 5 6
                                  . . . . .
C-SPOT                            1 1 1 2 1
------                            ---------
A-0123,1ALZQQ323,170              + + P ? +
B-0123,1ALZQQ321,253              + + + - I          ...
A-0202,1ALZZZ123,305              - - M D W
                 .
                 .
                 .
```

The exemplar spot file may be used to record spots of interest in one or more gels and their relationship to other gels in a set of gels. Particular instances of interesting spots are recorded in an E-map. The E-map labeling convention is given in (a) and an example of an E-map file is given in (b).

3. SELECTED RESULTS FROM THE PHA DATA

GELLAB is being applied to PHA stimulation of lymphocytes (*8, 9*) the effect of asbestos on P388D1 macrophages (*10, 11*), Lesch–Nyhan syndrome (*12*), and Alzheimer's disease (*12*). In this section we present some preliminary results of the PHA effect. More biologically oriented and detailed reports are in press (*8, 9*).

The 2D patterns of pulse-labeled polypeptides of human peripheral blood lymphocytes stimulated with phytohemagglutinin were compared with those of unstimulated lymphocytes, searching for changes in the relative synthetic rates of particular polypeptides seen on the 2D gels. Initially, some 24 spots were quantitated using manual densitometry. Manual densitometry presents its own problems especially those concerned with reproducibility of spot delineation. Thus in some instances, manual densitometry is more reliable (e.g., the isolated

spot) and in others it is not. These gels were analyzed using GELLAB. The 24 spots which had been subjected to manual analysis (in which spot boundary definitions were subjective) yielded density values that essentially were parallel to the automated GELLAB measurements. Deviations from the manual were few and attributable in part to segmentation problems. Included in the set of 24 spots were some extremely light examples which caused some problems, occasionally being difficult to segment. Segmentation errors of course propagate throughout the gel comparison program and thence to the CGEL level. Using the mosaic facility, these outlier spots were easily detected.

Our experience with GELLAB suggests that $\frac{1}{4}$ to $\frac{1}{3}$ of the polypeptides visualized in these 2D gels show altered relative synthetic rates when resting vs PHA-stimulated lymphocytes were compared. Figures 6a–d show a few examples of some of the spots detected by GELLAB found showing significant changes as a consequence of PHA stimulation. Mosaic images 5a,c,d show examples of PHA-induced decrease in specific polypeptides while 5b is an example of GELLAB-detected increase. The lack of obvious context in each of the mosaics is solved by reference to Fig. 5, a R-map which provides the necessary context.

In general, and as expected, GELLAB did not perform any better than other methods when applied to such gel regions as the alkaline, where noise is high and densities overlap frequently. On the other hand, the tools needed to pursue suspected spot correspondences and to evaluate results iteratively from such regions are available in this system. In the results with PHA stimulation for example 5 of the 24 carefully studied landmark spots were from the difficult alkaline regions and comparisons obtained by GELLAB paralleled the manual.

4. Discussion

With the two previous papers in this series in mind, the GELLAB system for multiple-gel analysis has been defined to the point where we can examine system tactics and system problems in the overall biological context. It is necessary for such a view to carefully consider the kinds of questions which biological problems pose. These will determine the nature, depth, and range of the analyses to be performed.

1. Is only one or a very few 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 specificity of the product and the homogeneity of the generating cell line are very high. Here the gels are used as detectors and serve simply to confirm or deny the existence of a fragment. Simple flicker analysis may be all that is required under favorable conditions. Densitometery for this situation is a secondary consideration if one at all. An example is the case of a single cell line under identical tissue culture conditions with specific genetic mutation with only a few protein changes expected. A case in point is a single gene difference in an *Escherichia coli* mutant (*3*).

2. Are there changes in any of several spots as a result of time? Here, as we found in some of our macrophage experiments (*10, 11*), although the purity of the cell line may be assumed, cells in different phases of the cell cycle may and almost certainly do produce different subsets of fragments. These variations are in quantity as well so that densitometry is also required. Moreover, the complexity of the analysis is increased so that almost invariably n gels rather than a pair of gels must be compared. The answer to such questions as this requires 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 as a result of an applied stimulus? The less known about the outcome, i.e., the more exploratory the search, the more complex and extensive must be the gel analysis. When the cell line is only apparently homogeneous (as was the case of PHA-stimulated lymphocytes (*13*)) and where the effect of the stimulus is both complex and a function of time, the gel analyses become correspondingly more extensive, laborious, and complex. In such situations, where many new products may result, there seems no alternative to automatic spot pairing using a computer.

4. Is there a "fingerprint" of morphologically homogeneous but biologically and functionally different cell groups (e.g., 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 assuming minimum problems in the reproducibility of the gels. Particular interest must be focused not only on differences but on subgroup similarities.

5. Are known polypeptide fragments present in normal or abnormal quantities in a body fluid? Here are the quintessential problems of clinical chemistry but multiplied by the number of spots present in the gel. At first, the answer to questions of this type might seem simpler than to 2, 3, or 4 above. The comparison of a single gel's contents to some internal or external system standard certainly involves future developments in the area which has been called by Anderson "molecular anatomy" (*14*). Because of the need for extensive bookkeeping in multiple-gel analysis, it is likely that some of the types of data structures we will present here will be an aid in this development. These include keeping track of gels from different experiments, maintaining the ever-increasing catalog of spot characteristics, and in monitoring the successive necessary improvements in preparative technology leading to better gel reproducibility and comparability.

Gels may be thought of as complex objects similar to a geographic map with individual polypeptides appearing in distinct morphologic conglomerates. These provide valuable leads as well as a framework on which to build experience. They are not, however, in any way certain reference points. Unlike the geographic map, adjacency of polypeptides in the gel is no particular indication of related genesis or biological function. However, certain

characteristic patterns are obtained with carbamylation and other biochemical treatments. Comparing biological specimens by comparing their corresponding gel maps is one means of determining protein differences. Given a number of gels, polypeptide concentration values may be modeled as a density distribution for each set of corresponding spots.

4.1. System Characteristics

Gel scanning, segmentation, and pairing are each finite resolution digital processes. Each introduce some independent error. The computer analysis of a continuous process (for all practical purposes, a continuous gel) is performed in a digital space at both finite spatial and finite density digitization. Because a Vidicon TV camera has a nonlinear modulation transfer function, errors in its approximation can lead to additional error. These errors, small in general, constitute the lowest variance in the process.

Even when multiple gels of split samples are run there is additional variance beyond that due to gel scanning alone. Multiple samples of the same tissue cultures resulting in multiple gels provide an additional source of variation. Sampling of a biological process at various stages of its progression in synchronized or partially synchronized cultures is another source of error.

Overall GELLAB system variance was explored using the PHA lymphocyte data base. The reproducibility of repeated scans of the same gel at resolutions of 250 μm/pixel was the first test. The mean variation of a gel between two scans of the same gel can be very low (about 0.005). Those spots which differed markedly were checked by direct visual examination of the segmented gel image and in some cases, the central core image was checked at the pixel level for spot definition using PIXODT (*1*). Increasing the spatial resolution to 170 μm/pixel further reduced this already small mean variation.

In another test of scan reproducibility, a special CGEL data base was constructed using a PHA lymphocyte gel and its control (non-PHA). These gels were scanned repeatedly at 170 μm/pixel with four scans of each gel. The vast majority of the data were consistent and exhibited a very low variance. Here, the cause of most deviations lay in the failure of the segmenter, operating on very dark conglomerates, to separate touching spots. Occasionally a significant deviation was found in some R-spot sets. It was apparent that a spot was sometimes not split from an overlapping spot and that scanner noise was a contributing factor. More than 90% of the spots previously identified by manual analysis as showing altered densities in response to PHA were detected by GELLAB as PHA-altered spots.

The system if it is to have utility, must be capable of dealing with variances which mask or obscure the biologically determined systematic variation among congeners which is really a major point of interest. We believe that we have demonstrated that GELLAB, with the use of backchecking, has this ability when applied to gels of reasonable uniformity and those that have been produced with good quality control.

4.2. System Limitations and Compensations

The use of higher-resolution scanning conditions (170 μm/pixel) although advantageous for spot resolution, etc., imposes some burden on the factor of field of view. Occasionally at this resolution every spot may not be included in the 512 × 512 pixel image. Reducing the magnification somewhat would solve this problem.

Dynamic range in the density domain using a Vidicon camera (approximately 0–1.8 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. The average spot is usually less than about 1.0 OD peak density. The usual care to avoid saturation of the autoradiographs is necessary. Silver-stained gels (*15, 16*) constitute more of a problem. More spots tend to saturate in the dynamic range of the Vidicon. By controlling the silver stain development process, the maximum OD of the silver 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) because of difficulties in constructing the C-gel. Some problems as a consequence of this approximation include: missing spots which are in other gels but not in the R-gel; mispairing a spot because it is poorly defined in the R-gel; and noise in the R-gel masquerading as true R-spots. Spots found in the R-gel may be edited. By editing, we mean that if a spot is incorrectly segmented such that it is missing or its centroid is incorrect, a new spot centroid may be manually defined and the old deleted (or replaced). This editing may be done using an interactive graphics program accessing the R-gel Gel Segmentation File (GSF).

A landmark spot should be well defined morphologically as part of a consistent pattern in all of the gels being compared. From 10 to 25 landmarks are generally denoted depending on the quality of the gel, with fewer landmarks required for the better gels. Fewer landmarks are needed if the regions have little distortion and strong local similarity. And highly populated spot regions need to be more densely landmarked than sparsely populated regions. However, the landmarks should not be "on top of" one another. When landmarks (*2*) are selected too close to one another, an incorrect bias is introduced. This is evident in the incorrect partitioning of spots influenced by digitization-type errors. There is also a higher probability of interacting with more landmark sets. We conjecture that the likelihood increases (though still very low) for a spot pair to be found in the *next* to next-nearest-neighbor landmark set rather than in the landmark or next nearest landmark sets. Thus the CMPGEL program (*2*) might be less robust under these conditions.

Independent of the basic biological variation there are additional intergel variances. These include exposure, sample concentration, gel loading, and film and staining development characteristics. The absolute density of a spot will thus vary from gel to gel. This is true even for gels generated at the same time from a split sample. By normalizing each spot by the *total* spot density, this

variation can be discounted. This only works well for good comparable relatively streak-free gels, where almost all spots are detected clearly and in all gels. However, it is also often the case that many false spots will be detected in the alkaline region or in other noisy areas in the gel. These additional "false spots" will not in general contaminate the CGEL data base because they do not pair with spots in other gels. However, they incorrectly augment the total gel density measurements.

Alternatively, a small subset of apparently relatively stable spots found in all gels in the data base may be selected for use in normalization. A standard statistical analysis of variance could be used to find spots in this set which are relatively stable. The sum of the densities of these spots would then be used to normalize all spots in the gels. Furthermore, the set of spots to be compared should not be indicative of the biological change being tested, since using them for normalization will result in other spots having their relative density shifted accordingly. This technique can be extended further by viewing the set of spots for normalization and then eliminating poorly defined or very light spots which results in better normalization estimates.

In any set of gels with associated experimental conditions, it is useful to partition them in various ways in response to different questions. Thus, for example, in the case of a much distorted poorly run gel with many outliers, one might wish to temporarily remove it from the set of gels in order to find statistically significant spots in the remaining members of the set. Later, the temporarily removed gel(s) could be restored to the set and these spots checked. Effectively, this procedure uses the results of the uncorrupted portion of the set of gels to investigate the outliers. It is possible to temporarily remove one or more gel(s) from the CGL data base by not including it (them) in the working set. All computation is then performed without these outlier gels even though they are present in the data base.

False positives may appear in a search results list of R-spots. This is often the result of incorrect inclusion of one or more noise points in a R-spot set, which nevertheless meets statistical criteria. The major tool for handling such false positives is backchecking using mosaic and R-map images. Also effective is direct visual examination of the R-spot numeric data itself. The false negative spot rate may be decreased by finding additional spots of interest by manually scanning the CGL data base R-spot set list for interesting R-spot sets but with outliers which caused problems with the current statistical tests.

When observing R-spot set distributions one occasionally finds outlier spots radically different from the rest of the spots in the set. Some of these changes are real, i.e., of biological origin, while others are artifacts of either the gel preparation or the image analysis. Currently, we assume that all data are valid unless found to be invalid through backchecking. This means that outliers are counted in the R-spot sets for statistical purposes.

However, it is possible to *ignore* or, alternatively, to *find* R-spot sets with outliers since they will have a large R-spot set standard deviation as well as significant differences in other features determined by the SET STATISTICS command and INDEX search.

4.3. Future Directions

At present, manual intervention in the initial stages of GELLAB is required in five key places. These include: (1) scanning the images, (2) defining the accession file information, (3) defining the computing window, (4) calibrating the ND wedge, and (5) landmarking (defining a set of landmarks common to the R-gel and another gel). If increased magnification were used and the ND wedge placed in the same approximate position at the bottom of the image, then both the computing window and ND wedge calibration could be more or less easily automated. By using a R-gel standard map, it might be possible to automate the landmarking to a greater extent.

Being able to view the constellation of R-spots as multiclass distributions facilitates finding subtle shifts in spot quantitation. Viewed in this manner, the expected variance of particular spots can be easily measured and thus 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 spot changes such that spots changing in the same or opposite manner as a function of independent experimental variable can now be determined.

The mosaic operation has developed into such a powerful tool that a fuller extension is suggested. A useful operation would be to be able to request X–Y location data on any spot on the display of a gel image pointed to by the user. Finding the corresponding R-spot set element indicated, the user could look at the actual entire R-spot set data or request that a mosaic image be generated and displayed. This would permit random interrogation of all of the gels for any spot selected. A variation of this algorithm would be to request a mosaic of *only* the segmented spot for each of the gels or the region minus the spot. This would be useful as a check on how well the spot was segmented in each of the gels. Validation of any spot's segmentation would be useful—especially when the spot occurs as part of a conglomerate of spots. Working backward, the equivalent spot could be cut out by the computer and displayed for the operator to view when backchecking results.

The DECSYSTEM-20 is a medium-size machine with generous core allowance. The limits on core resident CGL data bases (if they are composed of a large number of spots from large numbers of gels) are still too stringent. The core memory limitation can be circumvented by paging (i.e., transferring in and out) the data base from a much larger disk file. The physical and logical structures of this file are critical, as would be expected, in order to minimize transfer times for the various CGEL operations. This project is currently under way.

As a result of being able to handle many more gels and spots using the CGEL system, new types of problems arose. Some of these are listed below and may be incorporated into the GELLAB system in the future. Some of the following problems emerged during the development of GELLAB. As solutions are found they will be incorporated into the system.

1. Finding shifts in MW or pIe of a spot or spots. How is the difference

between a true shift and variance in gel system determined? First the variance in the system must be determined. The problem is that this variance is different for different spots (*17*).

2. Correlation of spots or groups of spots changing together or in opposition.

3. Handling spot conglomerates of spots where spots sometimes are separate and other times touch adjacent spots.

4. Accounting for saturated spots though successive stages of the system and possibly obtaining alternate measurements.

5. Handling outlier spots (which may be artifactual or real).

6. Handling R-spot set statistics when: (a) spots are missing from some gels because they are not resolved (incorrectly segmented) or mispaired, (b) spots "appear to be absent" from one class of gels.

7. Merging the results of two CGL data bases of the same set of gels for complementary R-gels (once as a control and once as a patient).

8. Determination of false positive and false negative rates on statistically significant spots.

Because of the variety of applications, we do not ever anticipate a fully automated system. We do suppose that 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.

5. Conclusion

A set of algorithms in the GELLAB system for the analysis of multiple 2D electrophoretic gels image spot lists using a composite gel file as a data base has been presented. These algorithms have been successful in analyzing spots under a wide variety of gel conditions and open 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 information retained. Furthermore, by constructing the data base using the inverted-file concept, it is possible to rapidly access and update the data base for most operations. Treating the composite data base as a set of distributions leads to the application of various statistical tests for determining spot significance in an automatic sequence. This is crucial when investigating a large number of gels with potentially of the order of 1000 spots each.

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

APPENDIX A: STEPS IN GELLAB ANALYSIS FOR MULTIPLE-GEL COMPARISONS

In order to convey some impression of its mechanisms and the interplay of programs and data, we present in outline the steps required for processing a set of gels.

In Step [1], gels are assigned unique accession numbers which are used to reference the gels in the system in the future. An accession number "XXXX.*E*" is a sequentially assigned four-digit number XXXX (with leading 0's) with its exposure *E* being the *E*th exposure of that gel accessioned into the system.

In step [2], gels are scanned using the Vidicon TV camera/RTPP picture memory hardware to digitize the images to 512 × 512 pixel 8-bit gray-scale data stored on 9-track 800 BPI magnetic tape. The RTPP is described in Refs. (*3, 18–22*). If negatives were scanned, they are complemented at this time before being stored on magnetic tape. The set of images (G1, G2, . . . , G*n*) constitute the gel image data base (cf. Fig. 3).

The accession file GEL.ID, is illustrated in Table IV. This is updated in step [3] with relevant patient and gel information as well as the name of the actual picture file for each gel accessioned into the system. This information may be used later to partition the data base by classes based on any keyword in a subset of this information. It is also used in various GELLAB programs as an indirect reference to the gel image file by accession number.

Since the wedge calibration program currently operates only on gel images residing on the RTPP disks, the selected gel images are transferred from magnetic tape to the set of RTPP picture disks in step [4].

The magnetic tape of gel images is copied in step [5] to the DEC-20 picture disk for later use by the segmenter and other programs. It is currently necessary to convert the quad images (four 256 × 256 image segments constitute an RTPP 512 × 512 image) into a single DEC-20 512 × 512 image using the CVGELP program for each of the gels.

Using the TOTDENSITY program on the RTPP in step [6], the user defines the computing window (the region in the gel image where the spots are located—omitting writing and ND wedge information in the gel image). At the same time, the ND wedge is calibrated by computing a gray-scale histogram of a 20-pixel-wide computing window sample of the ND wedge and matching peaks in the smoothed histogram with the actual ND values of the wedge. The user is required to position this sample window. The wedge gray value peaks information for the set of gels is updated into the GEL.ID file which is then transferred to the DEC-20.

Each of the gels in the set to be analyzed is segmented one at a time using the SG2DRV (*1*) spot segmentation program in step [7a] running on the DEC-20. This process is usually run as a batch job. The set of gel segmentation files produced are (GSF1, GSF2, . . . , GSF*n*) (cf. Fig. 3). Corresponding gel

segmentation images are also produced for visual backchecking on how well and which spots were segmented.

While the gel images are being segmented the user must manually define the landmark spot sets on the RTPP in step [7b]. The initial landmark set is defined for the R-gel in file LM1.DA. This file and a list of the other n-1 gels to be compared with the R-gel are input as parameters to the MAKCMP program which generates a batch job for the RTPP. This interactive batch job will generate the set of n-1 landmark spot files {LM2.DA, . . . , LMn.DA} by directing the landmark alignment program with cursor information as to what the next spot in the R-gel the operator is to landmark in the corresponding gel. Thus the landmark set will be the same for all n-1 gels. The RTPP landmark alignment program permits the operator to flicker align the two gels and then mark the aligned spot either with a TV cursor overlaying the gel images as directed by a graphics tablet or, in the case of a previously defined R-gel landmark, at the current cursor position (specified by the program). After the set of landmarks have been defined, they are transferred to the DEC-20.

In step [8], the new landmark spot set files are appended to the end of the landmark spot data base file LMS.LM. This file is used along with the particular accession file GEL.ID. This landmark spot data base 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 (such as (54.1, 73.1), where, in the example given in this paper, gel 54.1 is the R-gel).

The set of n-1 gels to be paired with the R-gel are then processed by the CMPGEL (*2*) program one at a time in step [9] on the DEC-20. The algorithm uses the landmark spots to partition the two GSFs into landmark spot region sets of spots. These are then paired using a heuristic nearest-neighbor algorithm. This is usually run as a batch job. The LMS.LM file will be searched for the required landmark set. The resulting set of gel comparison files is $\{GCF_1, GCF_2, \ldots, GCF_{n-1}\}$ (cf. Fig. 3).

Finally, the CGL data base is constructed and analyzed in step [10] using the CGEL program (cf. Fig. 3).

Appendix B: SAIL Data Structures for GELLAB—Some Details of Interest

The CGEL program is implemented in the SAIL programming language (*23*) for a DEC-20 (or DEC-10) computer. It uses the RECORD facility to create the multiple-field spot data structure records. The CGL data base is stored as a list of R-spot sets indexed by the R-gel spot. Each R-spot set consists of a linked list of spot records. For example, in Table IIIa the R-spot set [1] spot corresponding to R-gel 0054.1 and segmenter GSF index 0290 would have the 10-character CGEL index key "0054.10290." The spot key just mentioned is implemented using the LEAP associative store facility such that R-spot indices in the range of [1: 1100] can be recalled from the 10-character key. This "inverted-file" accessing method is useful when building (or later accessing)

the data base since it is important to determine whether a spot pair (one of which is from the R-gel) is *already* in a R-spot set. Extensive use is made of the macroexpansion and string-processing facilities of SAIL as well. Other important data structures are:

1. The set of working gels used to restrict the CGEL operations to a subset of the gels in the data base. Only gels in the working set are used in the computations.

2. The classification sets which contain the names of the gels in each of up to nine classes. These structures and those for (1) above are implemented using the SAIL LEAP SET facility.

3. The search results list containing a list of R-spot indices found by one of the many search procedures (or defined manually).

The "search results list" of R-spots which were found by various searches (or explicitly defined) is available to many of the CGEL operators for processing. Each gel has, associated with it, accession file text information, total gel density, and number of spots in the gel which is used to label tables and plots as well as for normalization for some operations.

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 and 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. Bob Connors of NCI suggested the Kruska–Wallis rank-order test for comparing more than two classes of gels simultaneously. We are particularly grateful to Eric Lester for allowing us the use of his intermediate results on PHA stimulation of lymphocytes even while he is subjecting his results to backchecking. Thanks are also due Bernice Lipkin for many useful suggestions on the organization of these three manuscripts.

Addendum: Since these three manuscripts were submitted, substantial changes have occurred in the system. These include elimination of the magnetic tape intermediary an integration of the RTPP picture display/acquisition hardware with the DEC-20. A disk data base version of CGEL has been written, called CGELP, with greatly enhanced capabilities. These include the paged data base capacity of up to 128 gels of up to 3000 R-spot sets. These changes as well as additional CGELP commands are documented in (*24*).

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,** 272–297.
2. LEMKIN, P., AND LIPKIN, L. GELLAB: A computer system for 2D gel electrophoresis analysis. II. Spot pairing. *Comput. Biomed. Res.* **14,** 355–380.
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. GARRELS, J. I. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254,** 7961 (1979).
5. NATRELLA, M. G. "Experimental Statistics," NBS Handbook 91, U.S. Government Printing Office, Washington, D.C., 1966.

6. NIE, H. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K., AND BENT, D. H. "SPSS-Statistical Package for the Social Sciences," McGraw–Hill, New York, 1975.
7. KNOTT, G. D. MLAB—A mathematical modelling tool. *Comput. Programs Biomed* **10,** 271 (1979).
8. LESTER, E. P., LEMKIN, P., LIPKIN, L. E., AND COOPER, H. L. A two-dimensional electrophoretic analysis of protein synthesis in resting and growing lymphocytes in Vitro. *J. Immun.* **126,** 1428 (1981).
9. 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.* **26,** 1392 (1980).
10. LEMKIN, P., LIPKIN, L., MERRIL, C., AND SHIFFRIN, S. Protein abnormalities in macrophages bearing asbestos. *Environ. Health. Perspect.* **34,** 75 (1980).
11. LIPKIN, L. Cellular effects of asbestos and other fibers: Correlations with in vivo induction of pleural sacroma. *Environ. Health. Perspect.* **34,** 91 (1980).
12. MERRIL, C., GOLDMAN, D., personal communication.
13. LIPKIN, L., LEMKIN, P. Data base techniques for multiple PAGE (2D gel) analysis, *Clin. Chem.* **26,** 1403 (1980).
14. ANDERSON, N. G., AND ANDERSON, N. L. Molecular anatomy. *Behring Inst. Mitt.* **63,** 169 (1979).
15. MERRIL, C., SWITZER, R. C., AND VAN KEUREN, M. L. Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. *Proc. Natl. Acad. Sci. USA* **76,** 4335 (1979).
16. GOLDMAN, D., MERRIL, C. R., AND EBERT, M. H. Two-dimensional gel electrophoresis of human cerebrospinal fluid proteins. I. *Clin Chem.* **26,** 1317 (1981).
17. HURLEY, P. M., CATSIMPOOLAS, N., AND WOGAN, G. N. Bidimensional electrophesis of nuclear chromosomal proteins from aflatoxin induced liver cancer. *In* "Electrophoresis '78" (N. Catsimpoolas, Ed.), pp. 283–296. Elsevier North-Holland, New York, 1978.
18. LEMKIN, P. "Buffer Memory Monitor System for Interactive Image Processing," NCI/IP Technical Report #21b, Nat. Tech. Info. Serv. PB278789 (listing PB278790), 1978.
19. LEMKIN, P., AND LIPKIN, L. BMON2—A distributed monitor system for biological image processing. *Comput. Programs Biomed.* **11,** 21 (1980).
20. CARMAN, G., LEMKIN, P., LIPKIN, L., SHAPIRO, B., SCHULTZ, M., AND KAISER, P. A real time picture processor for use in biological cell indentification. II. Hardware implementation. *J. Histochem. Cytochem.* **22,** 732 (1974).
21. 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).
22. LEMKIN, P., CARMAN, G., LIPKIN, L., SHAPIRO, B., SCHULTZ, M. "Real Time Picture Processor—Description and Specification." NCI/IP Technical Report #7a, Nat. Tech. Info. Serv. PB269600/AS, 1977.
23. 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-AO45-102, Springfield, Va., 1976.
24. LEMKIN, P. F., LIPKIN, L. E. Database techniques for 2D electrophoretic gel analysis, *Computing in Biological Science,* Elsevier/North-Holland (M. Geisow and A. Barrett, Eds.), in press.