Peter F. Lemkin and
Lewis E. Lipkin

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

2-D Electrophoresis gel data base analysis: Aspects of data structures and search strategies in GELLAB*

Search strategies for finding spot differences among multiple two-dimensional (2-D) polyacrylamide electrophoresis gels are discussed in the context of the GELLAB spot data base management system. A 2-D gel experiment should have a well-defined biological experimental and preparation protocol reflecting the hypotheses of the problem. So too should the analysis of its corresponding 2-D gel computer spot data base have a protocol. This protocol is heavily influenced by the nature of the biological experiment as well as 2-D gel preparation considerations including the realities of artifactual and systematic noise. It is further influenced by constraints due to computational considerations. The search strategy is that part of the analysis protocol in which an experimenter iteratively defines tests to find significant spot differences. One goal of designing a well thought out search protocol is to reduce the number of search iterations required. Aspects of some requirements and constraints for useful search strategies are discussed.

1 Introduction

When performing an experiment entailing the use of 2-D polyacrylamide electrophoretic gels (PAGE) as a tool [1-5], care is taken in designing the biological and 2-D gel preparation protocols. Similarly, a carefully defined gel data base analysis protocol is also essential, especially in the case of large multidimensional data bases for effective analysis. Because of the great computational power of GELLAB [6-12] and similar 2-D gel analysis systems [13-21], particular attention needs to be focused on the biological model(s) for spot changes suggested by investigators. Another way of thinking about this process is in terms of defining a search protocol that will result in significant spot differences being discovered. Many of the GELLAB search protocols were derived and implemented as a result of the requirements of biological protocols brought to our attention through the collaboration of many GELLAB users [10, 12, 22, 23, and in preparation: Howard, R. J., Aley, S. B. and Lemkin, P. F.; Wirth, P. J., Lemkin, P. F., Alexander, L. A., Thorgeirsson, S. S. and Lemkin, P.; McGuire, B., Colbert, D. A., Lemkin, P. F., Wirth, P. J., Heilman, C. A. and Thorgeirsson, S. S.]. Fig. 1 illustrates the general analysis process.

Given a set of gel images (derived from autoradiographs, photographs or stained gels themselves), one can construct a composite gel (CGL) data base (DB) using the procedure illustrated in Fig. 2. Because of lack of space we will not go into details here. See [7-12] for a discussion of this construction procedure: this consists of a preliminary discussion of the initial CGL analysis phase using the spot segmentation (*i. e.* spot feature extraction - including spot centroid and background integrated density) program SG2DRV [6-7, 11-12], gel pairing (*i. e.* spots between two gels) program CMPGEL [6, 8, 11] and multiple gel spot data base program CGELP [9-12]. All these programs run on a Digital Equipment Corporation DECsystem-10 (or -20) computer under either the TOPS-10 (or -20) monitor program.

Briefly, a representative gel or Rgel is selected from the set of gels in the experiment for inter-gel alignment purposes. Each of the other gels is aligned with the Rgel at a set of manually defined spots called landmark spots. All of the previously segmented spots in each gel are then automatically paired between the Rgel and other gels. The CGELP program is then used to construct and manipulate the paged CGL DB which groups corresponding spots from different gels together in sets called Rspot sets. Spots missing in the Rgel but present in other gels are extrapolated into the Rgel and denoted as eRspot sets so that all spots found in all gels are included in the composite gel DB. Fig. 3a schematically illustrates a CGL DB with its representative gel.

Correspondence: Dr. P. F. Lemkin, Park Bldg., Room 417, N.I.H., Bethesda, MD 20205, USA

Abbreviations: **AP:** Ambiguous pair spot label; **Bg:** Intercept for least squares density fit; **Cg:** Centroid of landmark set for gel g; **CGELP:** GELLAB CGL DB management program; **Cgel':** Estimate of the canonical gel of all spots; **CGL:** Composite gel (data base); **CMPGEL:** GELLAB spot pairing program between a gel and the Rgel; **Ci:** Subset of gels in the CGL DB belonging to class i; **Cr:** Centroid of landmark set for the Rgel; **CV:** Coefficient of variation (standard deviation/mean); **2-D:** Two-dimensional; **DB:** Data base; **DBMS:** Data base management system; **D'gi:** Normalized density of gel g for Rspot i; **D'gi:** Unnormalized density of gel g for Rspot i; **D''gi:** Piecewise linear least square normalized density of gel g for Rspot i; **EP:** Extrapolated pair spot label; **eRspot:** Extrapolated Rspot (i.e. missing in Rgel); **GCF:** Gel comparison file produced by CMPGEL program; **GELLAB:** A set of programs for computer-aided 2-D gel analysis; **GSF:** Gel segmentation file produced by SG2DRV program; **IDBM:** Internal data base management system; **INTERSECTION:** The set theoretic "intersection" of two sets; **LMS:** Landmark set of spots common to a gel and the Rgel; **Mg:** Slope of density least squares calculation; **MIC:** DECsystem-10 interactive batch file processor; **MW:** Apparent molecular weight (in a gel); **Ng:** Normalization factor for gel g; **PAGE:** Polyacrylamide gel electrophoresis; **PCG:** Paged (with respect to a disk file) CGL DB; **pIe:** Isoelectric point (in a gel); **PIX:** Picture file; **PP:** Possible pair spot label; **Rgel:** Representative gel from a set of gels used for landmarking; **Rmap:** Derived gel image with selected spots labed with Rspot #s; **Rspot set:** Corresponding spots across gels for a single Rgel spot in CGL DB; **SDS:** Sodium dodecyl sulfate; **SG2DRV:** GELLAB spot segmentation program; **SP:** Sure pair spot label; **SPSS:** Statistical Package for the Social Sciences; **SRL:** Search results list (of Rspot numbers); **UNION:** The set theoretic "unión" of two sets; **US:** Unresolved (pair) spot label; **WS:** Working set of gels in a CGL DB; ~: The set theoretic negation symbol

* Presented at the Second International Argonne-Mayo Symposium on Technical Advances in Two-dimensional Electrophoresis and Clinical Applications, Argonne National Laboratory, Argonne, IL 60439, USA, August 29 - September 1, 1982

0173-0835/83/0102-0071$02.50/0

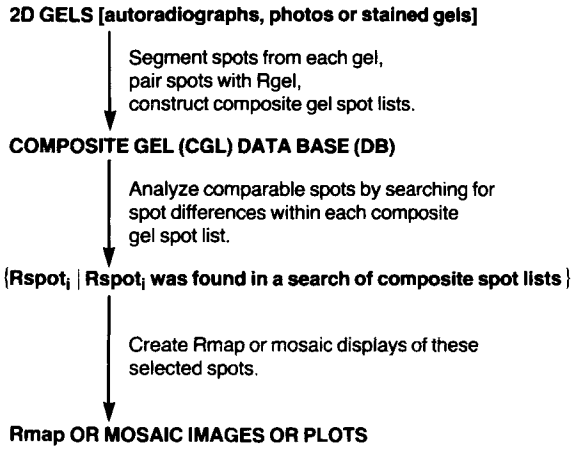


Figure 1. Overview of 2-D gel analysis process. Gels are scanned by the computer after which a composite spot data base, CGL DB, consisting of corresponding spots across gels is constructed. The CGL DB is searched for spot differences which are then visualized on derived Rmap and mosaic images for manual verification for true and false positive differences.

The term paged denotes that the CGL DB is brought into and out of computer main memory from a very large disk file in small chunks called 'pages'. This permits constructing and maintaining a very large multiple gel DB consisting of many gels with many spots. (In the case of GELLAB, a data base consisting of up to 128 gels of over 3000 spots/gel can be maintained. Typically, however, data bases consist of 10 to 50 gels of about 1000 spots/gel). Finally, spot differences are visualized and then manually verified by creating derived images called Rmaps and mosaics. An Rmap is an image of one of the gels in the data base overlayed with selected Rspot labels. A mosaic image is composed of panels of subregions of all of the gel images surrounding a particular Rspot and are ordered by increasing spot total optical density. Every 16 panels are grouped into one image and therefore multiple images are generated for data bases with more than 16 gels. Many examples of these have appeared in previous publications [6, 9–12, 22, 23].

In order to better understand the computational implications of the search strategies to be discussed, some understanding

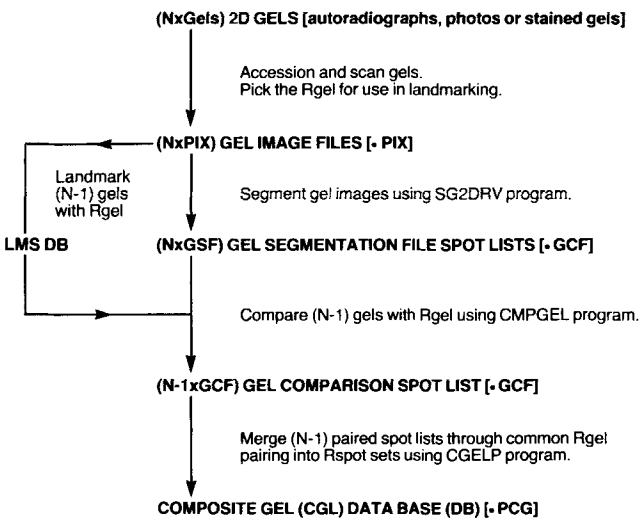


Figure 2. Overview of the process for constructing a composite CGL DB. The various processing stages are described in detail in [6–9] and extensions to GELLAB in [10–12].

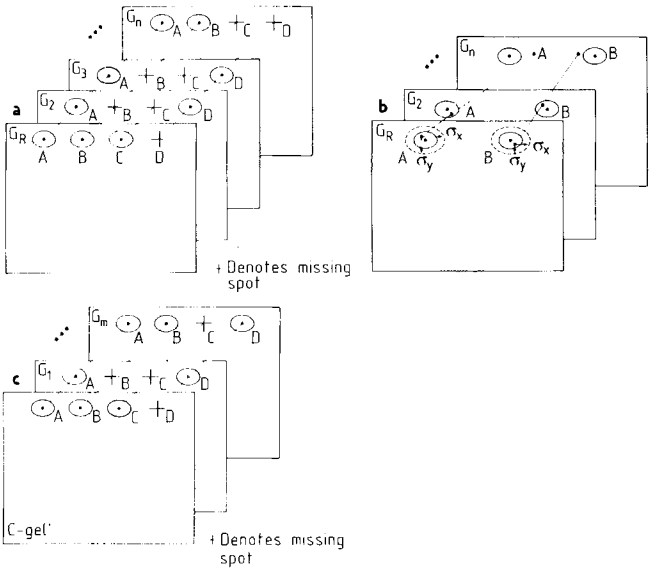


Figure 3. Examples of corresponding spots in GELLAB CGL DB. (a) Example of the occurence of the pairing of spots between gels (Gr,G2 to Gn). Spot A occurs in all gels; spot B in the R-gel and some of the other gels; spot C only in the R-gel (*i. e.*, US spot in the R-gel); Spot D in some of the other gels but not in the R-gel (*i. e.*, US spots in other gels). Spots [A, B and C] would be found as un-extended R-spots; spot D would be an eR-spot. (b) Example of mean spot positions computed by projecting spots from replicate gels onto the domain of the R-gel. When that has been done, means and standard deviations of X centroid, Y centroid, density and area may be computed. Such mean spots estimated from replicate gels are called synthetic C-gel' spots (*i. e.*, estimates of the canonical gel). Such an estimate is also used in establishing eRspots (Rspots missing in the Rgel – hence must be extrapolated there.) (c) Example of the occurence of spots between the synthetic C-gel' and a set of experimental gels (G1 to Gn). Spots A, B and C are R-spots found in C-gel'; spot D is an eR-spot not found in C-gel'. Therefore, spot C would be a spot found only in the experimental gels. (These three figures appear in [12].)

of the CGL DB structure and mechanisms to manipulate it is helpful. The PCGL (Paged CGL) data base is presented as an Internal Data Base Management (IDBM) system. In addition, some of the key data structures used in CGELP are mentioned. Finally, we concentrate on the basis for and derivation of search strategies for a 2-D gel data base analysis.

2 Methods

2.1 GELLAB as an internal data base management system

GELLAB is a complete data management system for 2-D electrophoretic gel analysis within a given laboratory. It is designed to compare gels in a single set of gel experiments. However, it can be used for comparing 2-D gel data among gel sets of different gel experiments in the same or different laboratories using the estimate of the canonical gel or Cgel' concept discussed in [12] (illustrated in Figs. 3b–c). The overall utility of the latter depends to a great extent on the reproducibility of sets of gels over long periods of time and between laboratories.

It is useful to have a consistent CGL DB spot numbering (which we term an Rspot number) on a particular type of material which is used in several different data bases. This

can be implemented as follows: the Rgel can be selected to be a previously computed Cgel′. Then all landmarking and spot pairing is done using this synthetic gel. When the CGEL DB is to be searched, the Cgel′ is removed from the working set of gels [9–12] so as not to enter into any of the searches. However, its presence imposes a specific Rspot numbering on the set of gels. Figs. 3b–c illustrate the Cgel′ concept. The issue of whether to use Cgel′ numbered data bases will not be discussed here. Rather, we will concern ourselves with the problem of searching a large multiple gel spot DB in an efficient manner.

In the literature, formalisms for coherently manipulating large quantities of data are called data base management systems or DBMS. Baker [24] defines a DBMS as a software facility used by one or more programs to assess the same DB (possibly in different ways) as well as to protect it from unauthorized access or changes. When the program becomes too complex, or the quantity of data too large to keep completely in the computer main memory, then a central data control facility greatly facilitates system design and implementation. This is not to say that the entire process could not be handled by a program kludge (*i. e.* an *ad hoc* creation, which is not recommended as good programming practice). As pointed out by Baker, there are two classes of DBMS: internal and external. CGELP, the DBMS of GELLAB, is an internal DBMS. Internal DBMS are characterized by a technique and not by a separate program (although this technique may be implemented as a set of procedures included in application software as is the case with CGELP).

In such a system, data resides in 1) disk files, 2) procedure areas of core memory and 3) global areas of core memory. All of these areas are under the control of the programmer. One problem in using global variables is a loss of programmer control in very large programs. This is caused by side-effects in using global variables by a large number of procedures and not adhering to a consistent use of these variables. An IDBM passes data directly from one procedure to another in a global manner through the intermediary of global variables, thus encouraging the consistent use of these global variables. In CGELP this is implemented by the use of a consistent set of global variables and a set of disk data base paging procedures.

The disk DB file in GELLAB logically consists of two parts, the first part being the state of global variables (which functionally may be thought of as the CGELP user state). It is updated on the disk only when directed to be so done by the user. It is read from the disk only when first instantiating a particular data base with CGELP in a particular user interactive session. The second part of the DB file is the actual Rspot sets (a set of corresponding spots in multiple gels) feature data. Each Rspot set is allocated a fixed amount of sequential disk space such that for a given number of gels there is sufficient space to store a spot for each gel and also up to 1.5 ambiguous pair (AP label) spots. (Ambiguous spots are additional spots or spot fragments associated with corresponding spots – see [8 or 11] for a more detailed discussion).

If in constructing or expanding a particular CGL DB it is found that there is insufficient space in which to store a particular Rspot set, then CGELP will automatically expand this space for all Rspot sets by shuffling the data to an extension of the file. Each Rspot set is reallocated as a larger contiguous subregion in the file. The shuffling of the file is done prior to continuation of the process which caused the problem. Because this is a time-consuming process that is hopefully never invoked, we have empirically selected a rather large AP factor of 1.5 spots/gel which insures that this expansion procedure will almost never occur under usual operation. If a large number of gels is added at a later time to an existing CGL DB, then the automatic expansion algorithm might be invoked in order to make room for the additional gels.

Since the offset of the start of each Rspot set can be quickly computed, we can then 'page' any Rspot set in or out of core memory. In practice, because consecutively numbered Rspots sets will generally be needed during a search through the CGL DB, a number of these Rspots sets are paged at one time (*i. e.* as many as fit into a 10 000 word core memory buffer). The number of Rspots paged at one time is a function of the number of gels in the data base. This is a reasonably efficient procedure since the cost of moving the disk head is much greater in terms of time than actually performing the data transfer.

2.1.1 Data structures in CGELP

Some of the special software data structures which support high level record keeping and the search aspects of the CGL DB are: sets, inverted lists, records, strings and associative table lookup. Because of lack of space we will not go into the details of these data structures and their use. Sets are implemented as bit arrays and are used for manipulating subsets of gels and Rspots. In the current 32-bit/word or larger generation of computers, it is possible to store 32 potential elements of the set in one word. A '0' bit value indicates no membership and a '1' bit value indicates membership. In many cases, intersection, union and set difference between 32 possible pre-assigned subset elements may be performed in one computer hardware instruction. For example, set intersection is a logical AND instruction, union is an inclusive OR and subtraction is the logical AND of one set with the 1's complement of the subtrahend. To implement a CGELP set of 3072 possible Rspots requires only 96 32-bit words! Thus many sets can be kept in core memory for efficient and rapid processing.

Inverted lists are used for mapping (x, y) positions in Rgel coordinate space to Rspot set numbers during both CGL DB creation and in order to map non-Rgel unresolved spots to extrapolated or eRspot sets using a landmark transformation [12]. This latter algorithm permits finding the Rspot set closest to a specified x, y coordinate pair.

Records are collections of data, called fields, associated with an object which for CGELP is a particular spot in a given gel for an Rspot set. Records are implemented as sub-arrays of the paged CGL DB and contain bit and sub-word fields. Within a particular Rspot set, records are ordered by a 'successor' linked list pointer field of each record. Thus the Rspot set can be maintained as a rank ordered (by spot total integrated density) list.

Strings and string operations (*e. g.* substrings, substring searching, number conversion to and from strings, composition of strings, *etc.*) are used extensively for table and formating presentation, auxiliary gel information, and some data processing.

Associative table lookup (*i. e.* hash functions) are used for mapping gel Accession numbers (which are five digit numbers XXXX.E) into one of 128 internal gel numbers for a given CGL DB. This is used extensively in checking to see whether a gel is in a set of gels or not (*e. g.* the working set of gels, gel subset, or gel class). For a large number of gels, it is much more efficient than checking a linear list of gels. This is critical since every spot (*i. e.* gel) in every Rspot set must be checked to see whether it is 1) in the working set of gels (see discussion on prefilter to follow), and 2) in a particular gel class if a multiple class test is being performed.

2.2 Search strategies under GELLAB

In light of the implementation considerations previously discussed, we will discuss some practical search strategies developed under GELLAB and their rationale. We first review the entire 2-D gel DB computer analysis process at a high level and then at successively more detailed levels as in Fig. 1.

2.2.1 Removing marginal gels from the CGL DB

Marginal gels may be discovered in the CGL DB early in the analysis and temporarily removed from the working set of gels (WS) during statistical searches. This is necessary so as not to incorrectly bias the searches. Later, prior to making mosaic or Rmap-derived images, the marginal gels may be added back into the WS for visualization. This permits checking marginal gels on the basis of results found under more robust constraints. Care must be taken if this removal procedure is used so as not to unduly bias which gels are removed (*i. e.* a particular experimental class of gels may not run well in the PAGE process).

A marginal gel is one which exhibits one or more of the following characteristics: 1) Visually it looks bad with very dark background, gross distortions of spots, very heavy clustering, artifacts, *etc.* 2) It is poorly segmented as evidenced by: 2.a) poorly segmented image (see [7] or [12]), or 2.b) very small or very large total number of segmented spots in relation to other gels in the DB. (This number is present in the CGL DB) or 2.c) it is very light or very dark in relation to other gels in the DB. 3) It has a very large mean square landmark deviation (using the VALIDLANDMARKS command in CGELP as well as being computed during landmarking).

The root mean square landmark deviation is a crude global estimate of gel 'goodness' and is estimated as follows: 1) Compute the centroid Cg of the landmark set of spots for gel g (actually existing as pairs of corresponding spots in both the Rgel and gel g). 2) Map the landmark spots in gel g to the domain of the Rgel by the vector transformation: Offset = Cr-Cg. 3) Compute the root mean square sum of the deviations for the set of landmark spots between the Rgel and the transformed gel g landmark spot coordinates.

Table 1 illustrates the root mean square landmark deviation values for a typical gel data base (72 h P388D1 cell line [25]). Note that two of the gels have high values. High values are indicative of either faulty landmarking or greatly distorted gels (with respect to the Rgel). In the case of faulty landmarking, the relevant gels are then re-landmarked and the data base rebuilt. This consistency check is also available at the time the manual interactive landmarking is performed so that corrections can be made immediately.

2.2.2 Normalization of a CGL DB

It is essential that the images comprising a CGL DB be properly normalized [10–12] prior to performing a quantitative spot difference search. Two of the five possible methods used in CGELP are presented here. For high quality gels with many similar spots, the ratio method is preferred as an estimate of total gel density. In those cases of very different gels where it cannot be practically applied, the modified least squares method is used.

2.2.2.1 Ratio method

1) Set the Rspot feature limits to select robust spots found in all gels in the WS of gels (*i. e.* bad gels and gels not of interest are removed from the WS). Typically, the feature limits are constrained for [area, OD range (*i. e.* maximum OD – minimum OD within a spot), CV area, minimum total integrated density D′, and the #gels/*Rspot* set = size of the WS. The pairing labels are: SP+PP+US (including eRspot DB)]. 2) Do a ratio type search of the CGL DB to compute the normalization factor *Ng*, by equation (1), for all gels *g* in WS to find a set of consistent *Rspots*. *D′gi* is the background-corrected total integrated Optical Density (OD) for *Rspot i* in gel *g*.

$$Ng = \sum_{Rspot_i \text{ meets limits}} D'gi \qquad (1)$$

and the normalized density is computed by equation (2).

$$D^{r}\text{gi} = (D'gi/Ng) * 100\,\% \qquad (2)$$

2.2.2.2 Modified least squares method

1) Set the Rspot set feature limits to select robust spots found in any gels in the working set of gels and the Rgel (all gels may be used – even if they are marginal gels or not of immediate interest). Typically, the limits are constrained [for area, OD range, CV area, minimum total integrated density D′. The pairing labels are: PP+SP]. 2) Do a least squares type search of the CGL DB to compute a regression line (Mg, Bg), mapping D′g to the density domain of the Rgel. This is done for all pairs of spots in Rspot sets meeting the feature limits criteria (*i. e.* prefilter concept to be discussed) and for which the Rgel as well as gel g are present. A piecewise linear function is then used to map D′g to D″g such that for $|D'g| > |Bg|$ the function of D″g = D′gMg + Bg. Otherwise it is modeled by D″g = D′g (Mg/2) for Bg > 0 and D″g = D′g (2 Mg) for Bg < 0. This model minimizes the er-

Table 1. Landmark root mean square deviation of a gel[a)]

```
Valid landmarks [T is OK, F is NG or SM (same) LMS validity check]
 GEL   | A B C D E F G H I J K L M N O P Q R S T U V W X Y
-------------------------------------------------------
0269.1| T T T T T F T T T T T T T T F F F F T T
0266.1| T T T T T T T T T T T T T T T F T F T T
0267.1| T T T T T T T T T T T T T T T T T T T T
0268.1| T T T T T T T T T T T T T T T T T T F
0270.1| T T T T T T T T T T T T T T T F T T T T
0270.2| T T T T T T T T T T T T T T T T T T T T
0271.2| T T T T T T T T T T T T T T T T T T T
0272.2| T T T T T T T T T T T T T T T T T T T T
0273.1| T T T T T T T T T T T T T T T T T T T T
0273.2| T T T T T F T T T T F T T T F F F F T T
0274.1| T T T T T T T T T T T T T T T T T T T T
0275.1| T T T T T T T T T T T T T T T T T T T T
0276.1| T T T T T T T T T T T T T T T T T T T T
0277.1| T T T T T T T T T T T T T T T T T T T T
0278.1| T T T T T T T T T F T T T T T T T F T T
0279.1| T T T T T T T F T F F T T T T F T F T F
0280.1| T T T T T F T F T F F T T T T F F F T F
0281.1| T T T T T T T T T T T T T T T T T T T T
0282.1| T T T T T T F T T T T T T T F F T F T T
0283.1| T T T T T T F T T T T T T T T F F F T T

Percentage of gels in which landmark is present (T=100%)
-------------------------------------------------------
  A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y
-------------------------------------------------------
  T  T  T  T  T 75  T 90  T 85 90  T  T  T 95 80 65 75 65  T 75 %

Global estimate of LMS centroid of gel and RMS deviation from Rgel
-------------------------------------------------------
[0269.1] Mean LMS centroid (276,167), LMS RtMnSqDev from Rgel= .0
[0266.1] Mean LMS centroid (258,186), LMS RtMnSqDev from Rgel= 4.8
[0267.1] Mean LMS centroid (280,195), LMS RtMnSqDev from Rgel= 7.3
[0268.1] Mean LMS centroid (263,185), LMS RtMnSqDev from Rgel= 8.2
[0270.1] Mean LMS centroid (269,185), LMS RtMnSqDev from Rgel= 5.9
[0270.2] Mean LMS centroid (257,179), LMS RtMnSqDev from Rgel= 4.8
[0271.2] Mean LMS centroid (262,186), LMS RtMnSqDev from Rgel= 112.7
[0272.2] Mean LMS centroid (286,185), LMS RtMnSqDev from Rgel= 7.0
[0273.1] Mean LMS centroid (263,168), LMS RtMnSqDev from Rgel= 7.0
[0273.2] Mean LMS centroid (301,177), LMS RtMnSqDev from Rgel= 4.5
[0274.1] Mean LMS centroid (264,186), LMS RtMnSqDev from Rgel= 6.7
[0275.1] Mean LMS centroid (267,161), LMS RtMnSqDev from Rgel= 5.5
[0276.1] Mean LMS centroid (260,186), LMS RtMnSqDev from Rgel= 6.8
[0277.1] Mean LMS centroid (260,180), LMS RtMnSqDev from Rgel= 6.2
[0278.1] Mean LMS centroid (274,249), LMS RtMnSqDev from Rgel= 12.6
[0279.1] Mean LMS centroid (266,264), LMS RtMnSqDev from Rgel= 15.6
[0280.1] Mean LMS centroid (266,255), LMS RtMnSqDev from Rgel= 16.3
[0281.1] Mean LMS centroid (257,253), LMS RtMnSqDev from Rgel= 9.6
[0282.1] Mean LMS centroid (262,243), LMS RtMnSqDev from Rgel= 110.1
[0283.1] Mean LMS centroid (268,264), LMS RtMnSqDev from Rgel= 14.5
```

a) Two estimates of landmark deviation in a gel computed by GELLAB are the valid landmark table and the landmark root mean square deviation estimate. The valid landmark table is used to indicate that a landmark was within the specified distance (typically of the centroid within 3 to 5 pixels) of a spot. If it is not valid, then the user interactively-specified coordinates are used rather than a spot's segmenter computed centroid. At the bottom of the table the number of landmarks as a function of the number of gels is given, indicating landmarks which were difficult to landmark. The global estimate of gel deviation is given in the second table as the LMS root mean square deviation. Note that the validity table and root mean square deviation metric can be used to indicate that something is wrong with the landmarking of a particular gel or that it is geometrically quite different from the Rgel. In this table, the case where gels 271.2 and 282.1 have high values was traced to faulty landmarking by the operator. Gels 278.1, 279.1, 280.1 and 283.1 had somewhat larger global distortion than the other gels when compared to the Rgel.

ror for the set of spots in the range of D′ used to estimate the line. It is better at forcing the intercept to go through zero (thus better estimating low values of D′ which have a large margin of error if a piecewise estimate of the curve is not used).

2.2.3 Types of changes expected and corresponding searches

There are basically three types of real changes found between two or more 2-D gels: 1) qualitative (*i. e.* spots missing in one gel which are present in the other), 2) quantitative (*i. e.* varying amounts of protein in one gel relative to the other when the amount of material in one gel with respect to the other is taken into account for normalization), and 3) shifts in MW or pIe. The third type of change can be handled by GELLAB for small deviations whereas, at this time, large changes cannot. Of course one must realize that, since 2-D gels are a finite precision detection system, qualitative changes may in reality be quantitative changes. So care must be taken in interpretation – possibly by verification with a darker exposure (in the case of an autoradiograph). Shift changes need to be separated from changes due to the position variance of simp-

ly running gels. Results obtained by the computer must be further checked. We will be discussing this in the context of: false positive (F+, spots mispaired), true positive (T+, spots correctly paired), and false negative (F-, spot not segmented) events. Proteins suspected of extensive pIe-MW migration need to be detected using other methods – gel flickering [2, 26] or possibly protein extraction techniques [27].

Qualitative changes can be succinctly defined as follows for the 2-class problem. Let *C1* and *C2* be the two classes. Let *r* be an Rspot set. Then, a search results list (*SRL*) of missing gel class spots (using ~ to indicate a logical NOT condition and & to indicate a logical AND condition) is defined by equation (3).

$$SRL = \{r \mid (\sim(r \text{ in } C1) \,\&\, (r \text{ in } C2)) \text{ or } ((r \text{ in } C1) \,\&\, \sim(r \text{ in } C2))\} \quad (3)$$

The test can be made more robust by requiring that, in order for (*r* in *Ci*) to be true, the subset of spots in class *Ci* meet the statistical limits criteria for the subset of gels belonging to class *i* for Rspot set *r*. The use of this prefilter will be discussed later.

Quantitative changes are generally found using a standard parametric (t- or F-) test or non-parametric (%-change- or Wilcoxson-Mann-Whitney rank order-) test of total integrated density (per spot) distributions for each Rspot set of gels [9–12]. Sometimes the researcher has indications of the change to expect, which is based on external biological evidence or experimental conditions. Often, however, this is not the case and so both types of searches should be performed.

2.2.4 Multiple gel analysis problems

The term multidimensional is used in this paper to refer to a set of data which can be partitioned into two or more classes of gels. For example, experimental condition-1 × ... × experimental condition-n. A set of gels from a particular experiment suggests by its biological protocol a partitioning into an n-dimensional abstract problem space. An experiment may be described as an n-class problem. Some typical examples might be: 1) 2-class problem: (control *vs.* experiment); 2) n-class problem: (control *vs.* exper[1] *vs.*... *vs.* exper[n-1]); 3a) n-class × m-condition problem: (control *vs.* exper[1] *vs.*... *vs.* exper[n-1]) × m-time samples; or 3b) n-class × m-condition problem: (control *vs.* exper[1] *vs.*... *vs.* exper[n-1]) × m-dose samples., 4) n-class × m1-condit. × m2-condit. problem: (control *vs.* exper[1] *vs.* ... *vs.* exper[n–1]) × m1-dose samples × m2-time samples. For all parametric and some non-parametric (rank order) tests, one needs at least two gel samples/class. This implies replicate gels of the sample specimen or duplicate cultures.

We present an example summarizing a partial analysis of a 4-class problem (four different types of leukemia) by breaking it down into several 2-class sub-problems. Thus 2-D gels of one class of leukemic cells are sequentially compared with a series of different classes of leukemic cells. Manipulation of subsets of spots is used to perform this problem decomposition. Leukemic cells from AML (acute myeloid leukemia), ALL (acute lymphoblastic leukemia), CLL (chronic lymphocytic leukemia) and HCL (hairy cell leukemia) patients were prepared and run as ^{3}H fluorograph 2-D gels as described by Lester *et al.* in [23], where a full discussion and analysis of this data is detailed. There are three possible combinations of 2-class difference searches where AML is one of the classes. We denote a search *i* by '*Si*' in equation (4).

$$\begin{aligned} &S1 = AML \text{ versus } ALL \\ &S2 = AML \text{ versus } CLL \\ &S3 = AML \text{ versus } HCL \end{aligned} \quad (4)$$

A possible biological question might be which polypeptides are possibly correlated with differences between AML and all of the lymphoid leukemias (*i. e.* ALL, CLL and HCL). Differences correlating with the AML class are found in the S1, S2 and S3 searches. Therefore a potential list of spots which meet the initial constraints of the question might be computed as S4 using the following set operations in equation (5), where *INTERSECTION* is the standard set theoretic operator [31]. If the missing class-test is used, then spots in S4 will be those either missing from AML while present in ALL, CLL and HCL, or present in AML while missing from the other three classes.

$$S4 = INTERSECTION\ (S1, S2, S3) \quad (5)$$

The result of any search is a search results list of which *S1*, *S2*, and *S3* are examples. GELLAB has SRL subset operations [10–12] for saving and manipulating (using set theoretic operators) up to 88 SRL subsets. The user would typically: (1) do the three searches required to generate the {Si}, saving the SRL in subsets called *S1*, *S2*, *S3*; (2) use the SRL subset UNION, INTERSECTION and DIFFERENCE operations to compute the final additional SRL subsets. In addition, any or all of the SRL subsets may be written into files and selectively restored at a later date. The set of Rspots in *S4* is a potential list of significant spots which 'however' must be verified with manual observation of Rmap and mosaic images.

2.2.5 Selection of search constraints-prefilter and succeeding test

The search procedure may be thought of as a two-stage process as illustrated in Fig. 4. The first stage is called the PREFILTER, detailed in Fig. 5, and attempts to select those Rspot sets which would give a robust response to the second part of the procedure. The second part is the parametric or non-parametric test. The use of the prefilter stage is critical in reducing the number of false positive estimates of significant Rspot set differences found when applying any test. Of course, any decrease in the false positive rate will result in an increase in the false negative rate. (A false positive is an event which is called significant by a test which, on further analysis, is determined not to be a real event. A false negative is an event which is real but was missed by the test. The rate is computed as the ratio of the number of specific events to the total number of events.)

2.2.6 Adiustment of search parameters in the prefilter

By adjusting search parameters, the false negative rate of the CGELP search could be set close to zero to pick up all real results. However, the side effect of doing this is to greatly increase the false positive rate up to near 100 %. Thus every

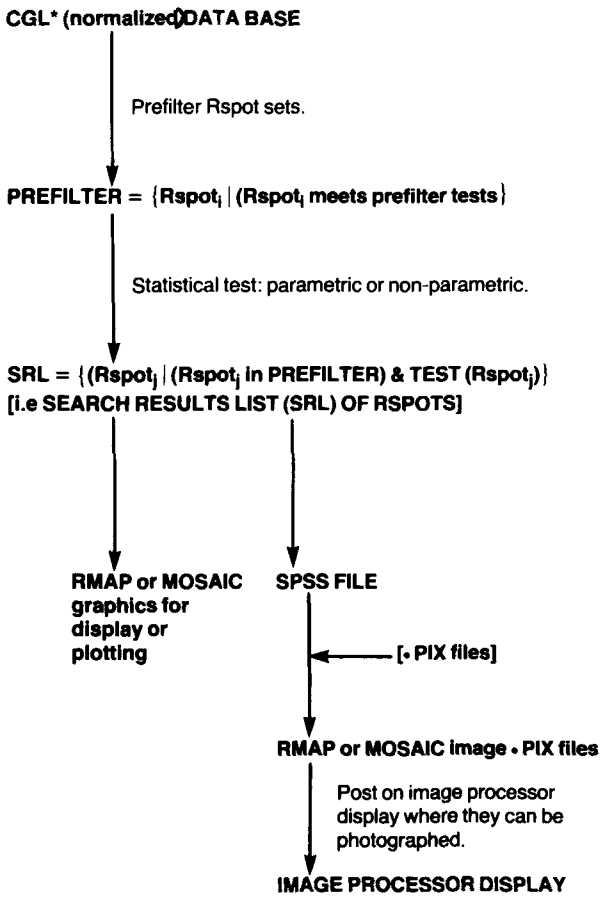


Figure 4. The CGL DB is searched as a two-stage process. Each normalized Rspot set is first tested with a prefilter to determine whether it is a candidate for further festing. If it is, then the statistical test selected is applied. Spots passing this second test are saved in the search results list, SRL, which is then used to make Rmap and/or mosaic images or plots, as well as for other SRL operations.

spot would be called significant! As in all practical statistically-based systems, a careful balance of the false positive and false negative rates must be found for each application. It is this issue of how one tunes the parameters for the search process which we will be addressing.

In the initial gel image data reduction to spot lists, the true positive (spot properly segmented) and false negative (spot not segmented) rates can be computed as follows: generate an image which is the original image with the segmented image [7] subtracted. An example of this type of image can be seen in [12]. Spots which were segmented appear as if they were cut out of the gel and are white. Spots which were not segmented are still in the image. A photograph is then taken and used to manually record spots that were not segmented by placing a pinhole through remaining spots. Counting pinholes yields the number of false negative events.

In the spot list pairing stage of the analysis, true positive (good spot pairing) and false positive (incorrect pairing) rates can be computed for a given set of spots paired by GELLAB. We have done this evaluation manually for a pair of similar gels of P388D1 macrophage-like cells [25] under different stimuli (72 h control and 72 h amosite asbestos). The relative error rates were then calculated. The flicker system [26] was used to manually check each spot pairing on derived images marked with the same sure pair, SP, or possible pair, PP, pairing label [8].

If replicate gels are used in the CGL DB with the constraint that prefiltered spots be in the replicate gels, then the false positive rate in the CGL DB can be greatly reduced since it is very unlikely for a false positive spot to occur in all of the replicate gels. The end product of a CGL DB search is a search results list and is a list of the Rspot set numbers which passed both the prefilter and the particular statistical rest. Fig. 6 shows the various options available in GELLAB for manipulating the SRL in continuing an analysis.

2.2.7 Evaluation of search results using Rmaps and mosaics

A typical 2-D gel CGL DB analysis protocol is shown in Fig. 7. For the case where a good guess has been made on the

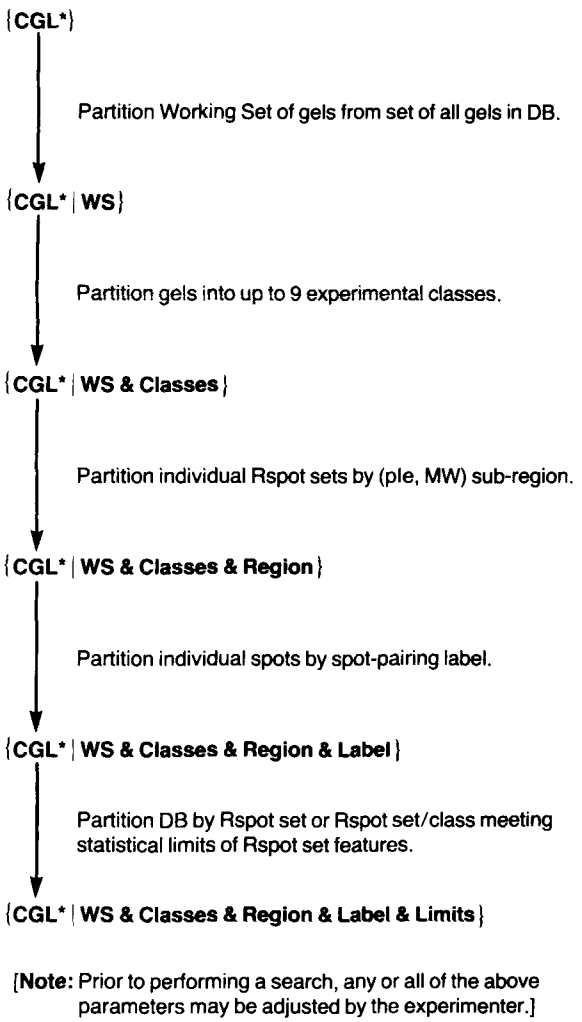
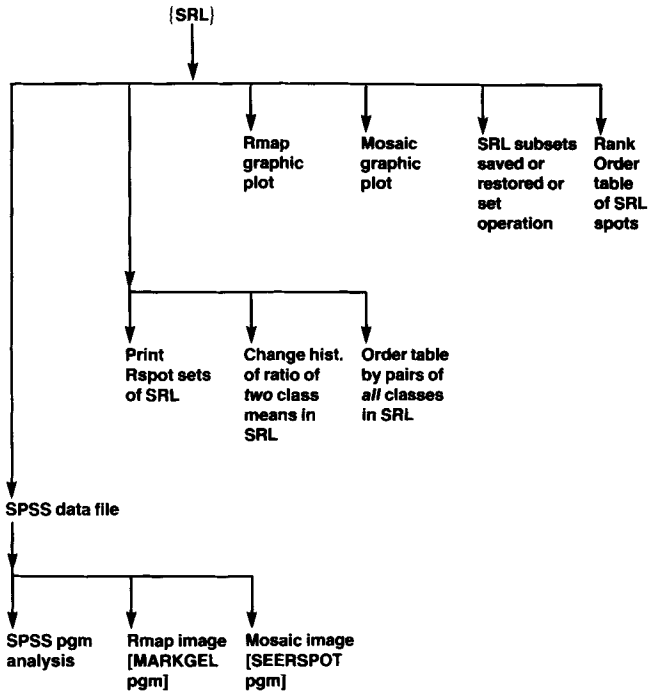


Figure 5. The prefilter test consists of five Rspot set or subset specific tests which determine whether or not an Rspot set is to be visible for further processing. Any of the parameters of the five tests may be changed by the user. The statistical limits are applied on such Rspot set features as: relative distance from the landmark, mean DL, mean DP, mean area, mean density, CV Rset area, CV Rset density, OD difference within spot, significance level, # gels in Rspot set. For example, restricting the CV area, OD difference, and # gels/*Rspot* set limits would be useful in finding robust stable spots possibly suitable for normalization.



[Note: Each of these options *may* be performed.]

Figure 6. Additional CGELP processing options are available to check and manipulate the search results lists. Used judiciously, these can aid the refinement or termination of the search process.

initial prefilter parameters and only two classes of gels exist, a single search through the CGL DB might be adequate. Each SRL search to be viewed in an Rmap or masaic image is saved in an SPSS data file (suitable for further analysis using the SPSS statistical package [29]). When the user is finished making a number of SPSS files, he would leave the CGELP program and proceed to make Rmap and mosaic images using the two GELLAB programs MARKGEL and SEERSPOT discussed in [12]. Particular Rspots can then be manually evaluated as true or false positive events. If more information is desired, CGELP is reentered and the process repeated until those outstanding questions (which can be answered by such a 2-D gel DB system) are answered.

It is more difficult to realize those questions which cannot be answered by probing a given CGL DB. If spot changes are subtle, the normalization marginal or noisy, inadequate numbers of replicate gels used, the statistical results not particularly robust, etc., other experiments should be performed to establish the tentative results. This may include adding more replicate gels or changing the experimental conditions under which the gels are run.

The making of mosaic images is a computationally expensive procedure. In the case where the SRL contains just a few spots, mosaic images of all of them can be generated and evaluated manually. In the case where there are many spots (greater than, say, 10 or 20), a brief manual pre-analysis of the Rmap(s) of these spots can be used to eliminate many of the false positive spots in advance of making mosaics, thus reducing the number of mosaic images to be actually generated. Most of the false positive spot differences can be quickly resolved by flicker, comparing Rmaps (derived images may be flickered as well as original images) from each of the two classes used in the search. Fig. 8 illustrates this iterative parameter adjustment process.

3 Results

Recently the missing class search (*i. e.* qualitative differences) has been applied successfully to several different problems in which missing spot differences occurred ([23] and in preparation: Howard, R. J., Aley, S. B. and Lemkin, P. F.; Wirth, P. J., Lemkin, P. F., Alexander, L. A., Thorgeirsson, S. S. and Lemkin, P.; McGuire, B., Colbert, D. A., Lemkin, P. F., Wirth, P. J., Heilman, C. A. and Thorgeirsson, S. S.). Fig. 9 illustrates the Rmap 9.a and mosaic 9.b of a spot found with the missing class search in the comparison of clones of malarial intracellular parasite *Plasmodium knowlesi* (Howard, R. J., Aley, S. B. and Lemkin, P. F., in preparation).

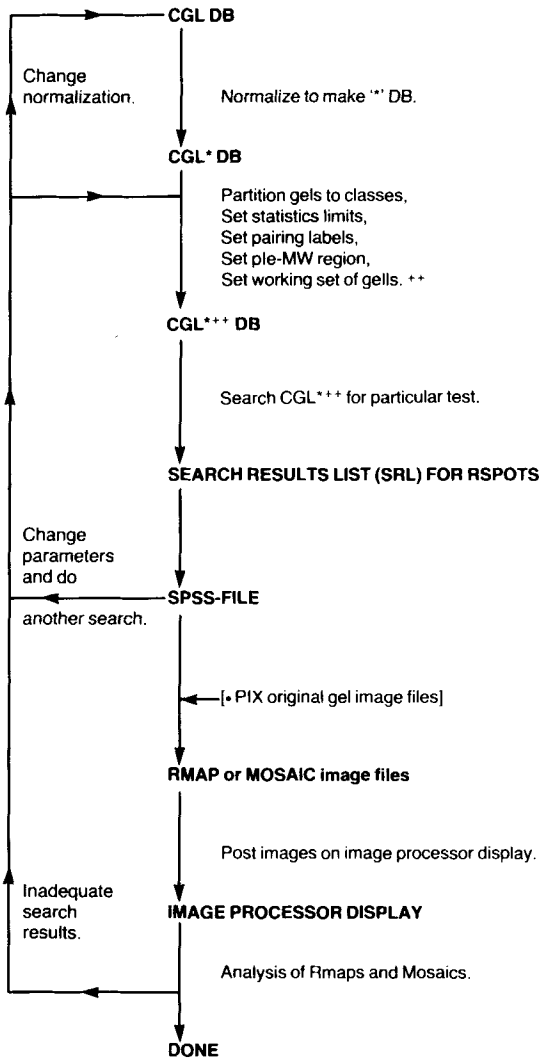


Figure 7. A typical 2-D gel CGL DB analysis protocol. Modification of prefilter parameters or tests may be indicated for iterative search based on an analysis of the SRL found in the current search. Set operators on SRL subsets found in a series of tests may also be necessary to detect multidimentional class differences.

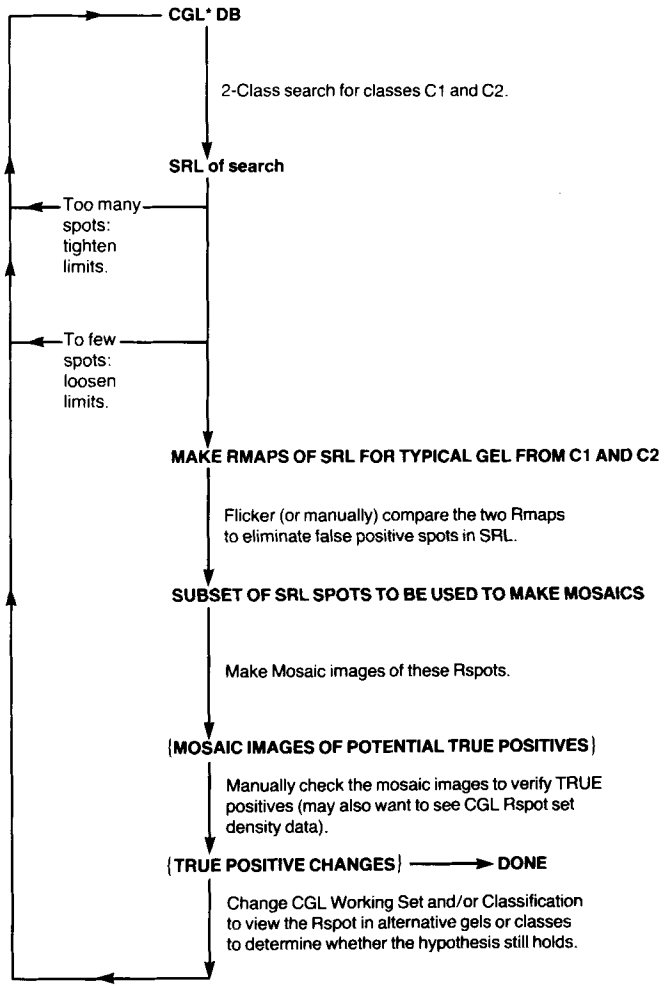


Figure 8. Iterative adjustment of prefilter search parameters can be used to modify the TRUE/FALSE positive rate of spot differences detected. Decreasing the false negative rate (to find more real spot differences) will result in also increasing the false positive rate (finding more incorrect spot differences) and vice versa.

Fig. 10 shows the potential set of spot differences found between AML and all of ALL, CLL and HCL in the human leukemia cell DB [23] using the Rspot subset processing discussed before. Subsets S1 through S3 were constructed by applying the missing class test on Rspots prefiltered so as to consider only spots with a total least squares normalized density greater than 10, *i. e.* dark spots. The Rmap of Rspot subset S3 is illustrated in Fig. 10a and a mosaic of one of the major spot changes from subset S4 is shown in Fig. 10b.

An order by class table for some of the S3 potential spot differences is shown in Table 2. In such a multiple-class prob-

Table 2. Order by class table for some S3 *Rspot* sets[b)]

R-spot:	m1/2	m1/3	m1/4	m2/3	m2/4	m3/4
167	–	<	–	–	–	–
536	<	–	–	–	–	–
754	–	–	–	–	>	–
803	–	–	–	<	<	<
896	–	–	–	–	–	–

R-spot:	m1/2	m1/3	m1/4	m2/3	m2/4	m3/4
167	–	0.4	–	–	–	–
536	0.1	–	–	–	–	–
754	–	–	–	–	1.2	–
803	–	–	–	0.4	0.3	0.9
896	–	–	–	–	–	–

b) The *Order by Class* table is a way of presenting the relationship of spots in all classes to one another (*i.e.* normalized mean class spot density ratio). This table is computed only for those spots in the SRL. Entries for which the subclass did not meet the prefilter test are not counted and appear in the table as a '–'. A few of the S3 SRL subset (missing class search of AML Vs. HCL) spots were selected to show a representative illustration of the order by class table. Class 1 is AML, class 2 is ALL, class 3 is CLL and class 4 is HCL. Since any SRL resulting from a search needs to be visually checked, any inferences to be drawn from such tables need wait for visual confirmation of the spots using Rmaps and mosaics.

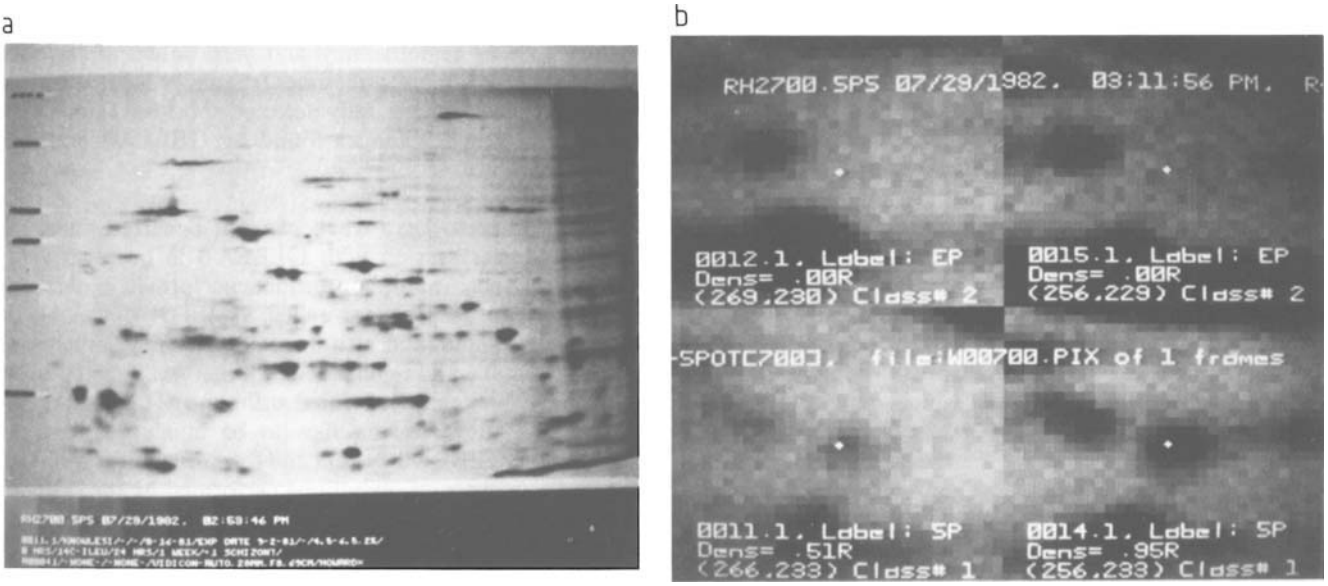


Figure 9. An Rspot found with the missing class search in the comparison of clones of malarial parasite *Plasmodium knowlesi* grown in Rhesus monkey erythrocytes (Howard, R. J., Aley, S. B. and Lemkin, P. F., in preparation). (a) Rmap showing global position of Rspot (700) within the gel. (b) Mosaic of Rspot (700). Class 1 is the +1 schizont clone and class 2 is the +2 schizont clone.

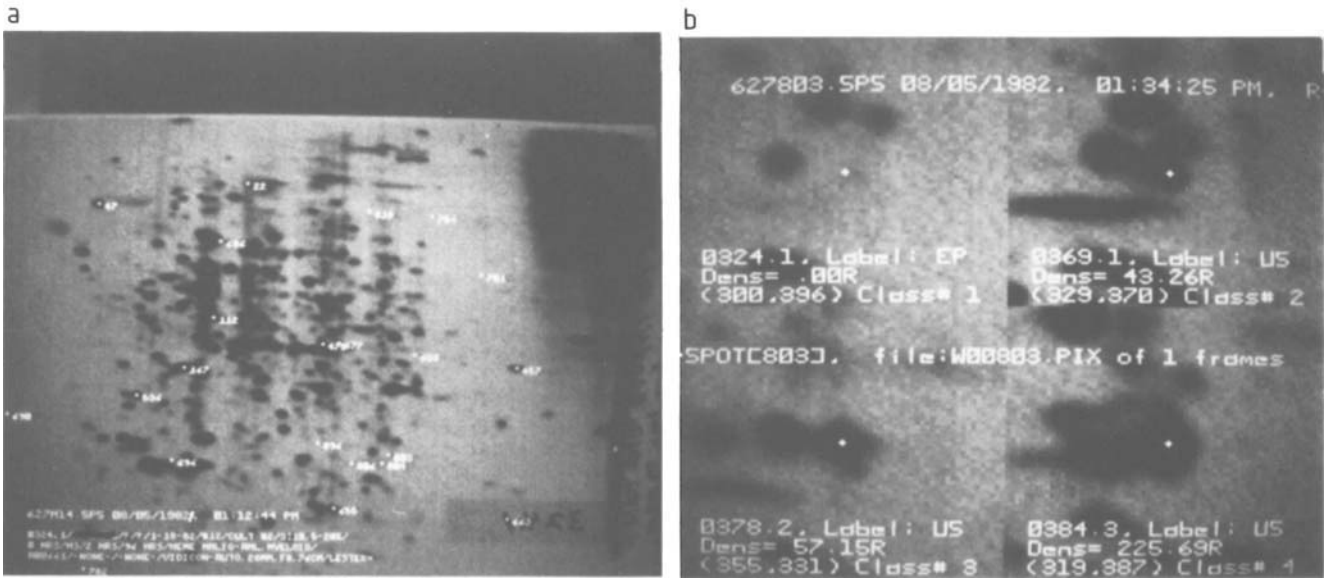


Figure 10. Some potential Rspot differences between human leukemic cells. A 'difference' for this example was the missing-class test for the two classes with the prefilter restriction that the mean least squared normalized Rspot density/class be greater than 10 which effectively removed very light spots from consideration (at the cost of increasing the F- rate). Rspots found using the spot subset processing discussed resulted in Rspot subset S4. (a) The potential list of Rspot differences in human leukemic AML cells and HCL cells (SRL S3). The original Rmap of Rspot subset S3 contains spots which must be manually checked to eliminate false positives and verify true positives. Note that some spots are obviously false positives, such as Rspot[663]. Others such as Rspot [803] appear in vacant regions. In the latter case, the Rmap was generated with extrapolated pairs so that we could see where the missing spot would be if it were present. (b) A mosaic of Rspot[803] from the S4 subset (INTERSECTION of S1, S2 and S3) where the spot is missing from AML but present in ALL, CLL and HCL. In the mosaic image, class 1 is AML, class 2 is ALL, class 3 is CLL and class 4 is HCL.

lem, after spots are found in a search between two classes, their mean density ratios may be observed between all combinations of classes. This procedure allows confirmation of suspected relationships between class differences.

In an experiment to manually check the true and false positive rates between the two similar P388D1 gels, we found that the relative total false positive rate (relative to all detected SP and PP labelings) was about 7 %. The false positive spots in the difficult edge-of-gel regions was about 5.6 %. Since these are spots which are, in general, difficult to run reproducibly, it is not surprising that most of the false positive spot pairing is in these regions. Ignoring these, the actual false positive rate is less than 2 %. These were primarily caused by major local distortions in the gel or by a local clumping of many spots close together.

The false negative rate, computed by (F–)/((T+)+(F+)+ +(F–)), was estimated to be 12 % and 25 % for the two gels. However, by far the great majority (at least 50 %) of these "missed" spots were barely at the level of detectability and were often only detected as an increase in the haze in a region. Taking this into account, the false negative error rate would be between 5 % to 10 %. Furthermore, every one of the well-defined false negative spots was extremely light compared to most of the other spots which were correctly segmented.

4 Discussion

We have presented techniques for increasing the effectiveness of the 2-D gel analytic process for finding polypeptide spot differences. The use of such tools is properly a reflection of the hypotheses brought to bear on the entire biological experimental design and execution, and 2-D gel preparation. By carefully taking these factors into account, the analytical process can be enhanced in effectiveness and shortened in time by taking advantage of the powerful logical and statistical techniques described in this paper, which can lead to zeroing in on results if properly applied.

Obviously checking the T+, F+, F– rates for one pair of gels will not give us the actual rates for any arbitrary set of gels. However, it does give an indication of the types of errors the GELLAB system may make, thus aiding in interpreting results. These include: edge region or streak mispairing; mispairing due to an inadequate number of landmarks in a highly variable region; hazy and very light spots are not always detected or are sometimes fragmented. Most of these errors occur with marginally detected spots or regions. As a result, robust spot changes found by GELLAB are given more credence.

It is necessary to keep track of what occurred during an interactive user session with GELLAB in order to verify parameters specified and procedures performed in obtaining search results. A program called SIMSES – available on TOPS-10 systems (and also available on many other computer systems) – allows all video terminal "traffic" to be saved in a file which may be printed at the end of the session. We have found these session files to be invaluable for both teaching the GELLAB system and for keeping track of complicated sequences of operations performed by the operator during an analysis without slowing down the user by requiring them to use a hard copy (*i. e.* printing) terminal.

Particularly useful is the grouping of short often-used sequences of operations in so-called 'command' files which can be instantiated as a high level 'command' with particular

arguments at execution time. These command files may be run under the TOPS-10 concurrent batch program processor or the interactive batch processor called MIC. For example, two frequently used MIC high level commands are MARK and MOSAIC. The former computes an Rmap image given the gel accession number and the SPSS SRL spot file name. The latter constructs a mosaic image, or images, given the specified Rspot number and the SPSS SRL list file which contains it. The command file MARK.MIC (MOSAIC.MIC) passes the user specified arguments plus default program switches to the MARKGEL (SEERSPOT) program(s) mentioned previously. Recent enhancements to CGELP involve generation of both an SPSS file and MIC command file to generate mosaic images contained in a specified SRL subset. Using these high level command forms permits one to concentrate on the problem and not get bogged down in details.

A practical problem, difficult to decide, is the choice between assembling a very large data base, or constructing several smaller ones. A 100-gel CGL DB not only takes a very large amount of disk space, but more important, with all gels in the working set, it takes a long time to search. One alternative is to reduce the working set of gels in such a large DB to a still statistically-significant smaller number of gels for probing purposes. After finding interesting spots using the smaller working set of gels, the working set is expanding back to the full set of gels for generating Rmaps and mosaics for manual visual backchecking.

We wish to thank Russell Howard and Steve Aley of NIAID at NIH and Eric Lester of the University of Chicago Medical School for permission to use partial results from their data bases as illustrations in this paper.

Received August 23, 1982

5 References

[1] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007-4021.
[2] Anderson, N. G. and Anderson, N. L., *Behring Inst. Symposium* 1977, Mitt. 1979, *63*, 169-185.
[3] Young, D. S. and Anderson, N. G., *Clin. Chem.* 1982, *28*, 737-739.
[4] Anderson, N. G. and Anderson, N. L., *Clin. Chem.* 1982, *28*, 739-748.
[5] Lester, E. P., Lemkin, P. F. and Lipkin, L. E., *Anal. Chem.*, 1981, *53*, 390A-404A.
[6] Lipkin, L. E. and Lemkin, P. F., *Clin. Chem.* 1980, *26*, 1403-1412.
[7] Lemkin, P. and Lipkin, L., *Comp. Biomed. Res.* 1981, *14*, 272-297.
[8] Lemkin, P. and Lipkin, L., *Comp. Biomed. Res.* 1981, *14*, 355-380.
[9] Lemkin, P. and Lipkin, L., *Comp. Biomed. Res.* 1981, *14*, 407-446.
[10] Lemkin, P. F. and Lipkin, L. E., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, Walter de Gruyter, New York 1981, pp. 401-409.
[11] Lemkin, P. F. and Lipkin, L. E., in: Geisow, M and Barrett, A. (Eds.), *Computing in Biological Science*, Elsevier/North Holland, Amsterdam 1983, pp. 181-231.
[12] Lemkin, P. F., Lipkin, L. E. and Lester, E. P., *Clin. Chem.* 1982, *28*, 840-849.
[13] Garrels, J. I., *J. Biol. Chem.* 1979, *254*, 7961-7977.
[14] Skolnick, M. M., Sternberg, S. R. and Neel, J. V., *Clin. Chem.* 1982, *28*, 969-978.
[15] Skolnick, M. M., *Clin. Chem.* 1982, *28*, 979-986.
[16] Bossinger, J., Miller, M. J., Vo, K. P., Geiduschek, E. P. and Xuong, N. H., *J. Biol. Chem.* 1979, *254*, 7986-7998.
[17] Vo, K. P., Miller, M. J., Geiduschek, E. P., Nielsen, C., Olson, A. and Xuong, N. H., *Anal. Biochem.* 1981, *112*, 258-271.
[18] Miller, M. J., Vo, P. K., Nielsen, C., Geiduschek, E. P. and Xuong, N., *Clin. Chem.* 1982, *28*, 867-875.
[19] Taylor, J., Anderson, N. L., Coulter, B. P., Scandora, A. E. and Anderson, N. G. in: Radola, B. J. (Ed.), *Electrophoresis '79*, Walter de Gruyter, New York 1980, pp. 329-339.
[20] Taylor, J. and Anderson, N. L., Anderson N. G., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, Walter de Gruyter, Berlin 1981, pp. 383-400.
[21] Taylor, J., Anderson, N. L., Scandora, A. E., Willard, K. E. and Anderson, N. G., *Clin. Chem.* 1982, *28*, 861-866.
[22] Lester, E. P., Lemkin, P. F. and Lipkin, L. E., *Clin. Chem.* 1982, *28* 828-839.
[23] Lester, E. P. and Lemkin, P. F., *Electrophoresis* 1982, *3*, 364-375.
[24] Baker, R. A., *Datamation*, May 1982, 139-142.
[25] Lipkin, L. E., *Environ. Health Perspect.* 1980, *34*, 91-102.
[26] Lemkin, P. F., Merril C., Lipkin, L. E., *Comp. Biomed. Res.* 1979, *12*, 517-544.
[27] Anderson, N. L., Nance, S. L., Pearson, T. W. and Anderson, N. G., *Electrophoresis* 1982, *3*, 135-142.
[28] Korfhage, R. F., *Logic and Algorithms*, J. Wiley & Sons, New York 1966.
[29] Nie, H. H., Hull, C. H., Jenkins, J. G., *et al., SPSS - statistical package for the Social Sciences,* McGraw Hill, New York 1975.