[14] Sancar, A., Kacinski, B. M., Mott, D. L. and Rupp, W. D., *Proc. Natl. Acad. Sci. USA* 1981, *78*, 5450–5454.
[15] Reeve, J., *Methods Enzymol.*, 1979, *68*, 493–503.
[16] Clarke, L. and Carbon, J., *Cell* 1976, *9*, 91–99.
[17] Tabor, S. and Richardson, C. C., *Proc. Natl. Acad. Sci. USA* 1985, *82*, 1074–1078.
[18] Studier, F. W. and Moffatt, B. A., *J. Mol. Biol.* 1986, *189*, 113–130.
[19] Latter, G. I., Burbeck, S., Fleming, J. and Leavitt, J., *Clin. Chem.* 1984, *30*, 1925–1932.
[20] Blose, S. H., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, VCH Verlagsgesellschaft, Weinheim 1986, pp. 552–555.
[21] Aebersold, R. H., Leavitt, J., Saavedra, R. A. and Hood, L. E., *Proc. Natl. Acad. Sci. USA* 1987, *84*, 6970–6974.

Peter F. Lemkin[1]
Eric P. Lester[2]

[1]**Image Processing Section, Laboratory of Mathematical Biology, National Cancer Institute/FCRF, Frederick, MD**
[2]**Veterans Administration Medical Center, and Coleman Building, University of Tennesee Central Health Services, Memphis, TN**

Database and search techniques for two-dimensional gel protein data: A comparison of paradigms for exploratory data analysis and prospects for biological modeling

Two-dimensional (2-D) polyacrylamide gel electrophoresis can detect thousands of polypeptides, separating them by apparent molecular weight (M_r) and isoelectric point (p*I*). Thus it provides a more realistic and global view of cellular genetic expression than any other technique. This technique has been useful for finding sets of key proteins of biological significance. However, a typical experiment with more than a few gels often results in an unwieldy data management problem. In this paper, the GELLAB-II system is discussed with respect to how data reduction and exploratory data analysis can be aided by computer data management and statistical search techniques. By encoding the gel patterns in a "three-dimensional" (3-D) database, an exploratory data analysis can be carried out in an environment that might be called a "spread sheet for 2-D gel protein data". From such databases, complex parametric network models of protein expression during events such as differentiation might be constructed. For this, 2-D gel databases must be able to include data from other domains external to the gel itself. Because of the increasing complexity of such databases, new tools are required to help manage this complexity. Two such tools, object-oriented databases and expert-system rule-based analysis, are discussed in this context. Comparisons are made between GELLAB and other 2-D gel database analysis systems to illustrate some of the analysis paradigms common to these systems and where this technology may be heading.

1 Introduction

This papers discusses paradigms for exploratory data analysis on two-dimensional (2-D) gel data and prospects for more global biological models of cell activity using such data. Why do we need 2-D gels and computer generated 2-D gel protein databases? In short, because major biologic processes (growth, differentiation, malignancy, *etc.*) are indeterminate, *i.e.* too complex for complete determinate analysis of their component parts and explanations of all of their interrelationships and functions. Biological, and particularly eucaryotic systems, consist of a vast number of parts, of which genes and gene products are key. Most estimates suggest 30 000–50 000 structural genes coding for proteins in eucaryotes, of which 5000–10 000 are expressed at some level in a given cell [1]. While current gels usually fail to deal with such numbers, the 2-D gel technique is theoretically capable of such resolution and sensitivity [2], a point confirmed by recent practical advances [3–5]. Each of these protein parts in a cell has complex interactions with other parts and the whole system is too complex for total description.

Thus, understanding these biologic systems is analogous to understanding ocean currents, the weather, or macroeconomics. For all such systems, understanding begins with classifying as many component parts as possible, describing their interactions, and ultimately developing idealized models of the system which serve as explanations of our understanding. The economy inherent in biological processes suggests that while many proteins are essential for the life of the cell, no

Correspondence: Dr. P. F. Lemkin, IPS, LTB, NCI/FCRF, Bld 469 Room 150B, Frederick, MD 21701, USA

Abbreviations: ADB, annotated databases; **ALL**, acute lymphoid leukemia; **AML**, acute myelogenous leukemia; **CLL**, chronic lymphoid leukemia; **CM**, composite match files; **CP**, composite group of spots; **2-D**, two-dimensional; **DB**, database; **EP**, extrapolated spot; **eRspot**, extended Rspot; **GCF**, gel comparison file of paired spots; **GM**, group match files; **GSF**, gel segmentation file of quantitated spots; **HCL**, Hairy cell leukemia; **LMN**, local morphologic neighorhood; **MW**, molecular mass; **NLDE**, nonlinear differential equation; **OODB**, object-oriented databases; **PCA**, principal component analysis; **PCG**, paged composite gel database file; **p*I***, isoelectric point; **Rgel**, reference gel; **Rspot**, reference spot; **SRL**, search results list

0173-0835/89/0202-0122 $02.50/0

one protein or metabolic pathway is of overriding importance – which is not to say that some pathways are not critical. Rather, each has its own role in concert with all the rest, and all must ultimately be included in our understanding.

Thus the attraction of the 2-D gel approach is its capacity to lay out in a single image, theoretically at least, a complete display of the genetic expression occurring in a population of cells. In a single 'snapshot' one has measured the qualitative and quantitative setting of the genome – how the DNA is 'tuned'. By cleverly comparing the data inherent in 2-D gels from different but related samples, we can begin to discern the organization of genetic expression – which genes are expressed coordinately during a biologic process and the sequence of changes in their expression. Two problems arise however: (i) The large quantity and complexity of the data ("it's too complicated, I want a simple problem!"), and (ii) the lack of a relationship between the 2-D gel data and traditional biochemistry ("What are all those spots anyway?"). The solution to the first is, obviously, creation and analysis of computerized databases (DB). For this, each of the computer systems described in this issue (as well as others mentioned) has faced a common set of problems: (i) record study data and digitize corresponding 2-D gel images; (ii) locate and quantify the component spots in each image; (iii) map the data from one image to the locations of homologous data in one or more similar images in an experimental series; (iv) create a unified database suitable for analysis. Such a database has, of necessity, a structure analogous to a three-dimensional (3-D) stack of 2-D gels (*cf.* Fig. 1), thus implying certain practical constraints on data manipulation as well as the possibility of analysis of the stack from various points of view (by gels, by individual spots, by groups of spots, by external ordering of the stack, *etc.*); (v) develop an analytic strategy, and (vi) display results of analysis with derived (marked) images and graphs.

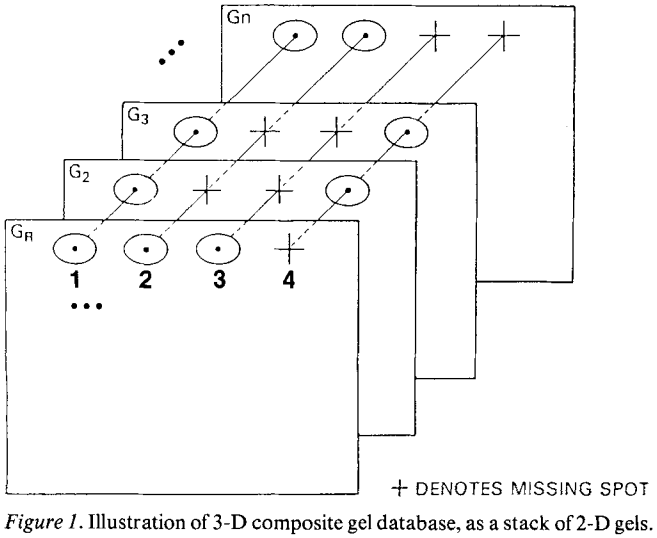

Figure 1. Illustration of 3-D composite gel database, as a stack of 2-D gels. The set of gels are tied together by the Rgel or Reference gel G_R. Corresponding spots are tied together and assigned to Rspot sets. Spots present in at least the Rgel are assigned to unextended Rspot sets (*e.g.* Rspots 1, 2 and 3), while those missing from the Rgel are assigned to extended eRspot sets (*e.g.* ERspot 4). Spots missing from any gel may be extrapolated as EP spots (*e. g.* G_R spot 4, *G*2 and *G*3 spots 2 and 3, G_n spots 3 and 4). Not all gels need be analyzed at any one time. The current subset of gels under analysis at any one time is called the working set of gels. This may be further subdivided into classes of gels which represent different experimental study conditions. An example of Rspot set data is given in Table 3.

In the course of this, problems such as recognition of data external to the gel image (*eg.* sample source, time sequence in an experiment, *etc.*), correction of digitization noise and other errors, and normalization of quantitative data must also be addressed. While important and potentially instructive differences between various computer systems exist, we believe that the fundamental similarity imposed by this hierarchy of data structures is more significant. Ideally, future systems will be modularized to permit the interactive choice of various solutions to each of these problems so as to optimize the result for a given problem. Analytic strategies likewise have a common set of problems. Fisher *et al.* [27] suggest a definition that is appropriate for 2-D gel databases: "Exploratory data analysis can be characterized as a search for regularity or structure among objects in an environment, and the subsequent interpretation of discovered regularity."

(i) Identify individual proteins whose presence or absence (qualitative change) serve as a marker for a class of samples (*eg.* a class of leukemias or cell lines transformed by a particular oncogenic virus). (ii) Identify consistent quantitative changes in protein expression (relative spot size) which can serve as markers for a class of samples. (iii) Comparison of groups of proteins identified in (i) and (ii) using set operations after various classes of samples have been compared and marker proteins identified. Thus proteins which are most consistently different between classes can be identified as markers for these differences. Such proteins, by virtue of their consistency, are more likely to represent key gene products which underlie the biologic differences between sample classes. (vi) Ultimately, a complete catalog of proteins observed in all the gels from each sample in various databases can be constructed. Creation of such large databases, particularly those containing summaries of the analytic comparisons and set operations noted above, provides a powerful inferential tool for comparison of mechanisms operative in diverse biologic systems. (v) Alternatively, gel patterns may be analyzed in a global fashion, in order to show the relatedness of samples to each other based upon a summary of their features. Various cluster analysis algorithms and other approaches will generate such 'summary statistics' and patterns of relatedness among samples. Such analyses also of course lead back to the individual marker proteins since a natural outcome is to ask 'which proteins contribute the most to the final arrangement of sample relatedness'.

The larger problem alluded to earlier, of relating 2-D gel data to other forms of biologic data, is more difficult. Ideally the protein databases can provide a scaffolding upon which more traditional biochemical information can be arranged, thus relating function to the underlying structure (proteins involved). The most obvious example of this sort is the capacity of 2-D gel database analysis to identify sets of proteins which are coordinately regulated during differentiation or other biological processes. In order to achieve this coordination, proteins in such sets must share common regulatory mechanisms. The recent explosion of knowledge about classes of DNA binding proteins provides examples of such common regulatory mechanisms operative upon groups of genes. The problem is that most 'spots' on 2-D gels are unidentified and most proteins which have been isolated and studied functionally have not been located on a 2-D gel map. Furthermore, there is as yet no universal 2-D gel map of proteins, transferable from lab to lab, and allowing meaningful communication based on the location of proteins in the map. Such

a solution to the current 'Tower of Babel' in the 2-D gel world is clearly feasible but technically non-trivial.

The best link, at present, between 2-D gels and the rest of biology appears to be the generation of partial amino acid sequences from proteins isolated from 2-D gels [6–8]. Thus key proteins identified by comparing sets of proteins defined by statistical analysis of 2-D gel databases can be partially sequenced and compared with the extant gene and protein sequence databases to obtain clues to their function. Such an approach also allows generation of nucleic acid probes which can be used to isolate, sequence and aid in defining the regulation of the gene coding for the original protein. By following this path for multiple members of a set of co-regulated proteins defined with a 2-D gel database for a variety of conditions, we may be able to elucidate and interrelate the various regulatory mechanisms operative. Thus our 2-D gel analyses may be used to lead our efforts at DNA sequencing in a rational way. The key point here is the need to relate expression (the evidence of regulation) and coordination of expression with the underlying genetic structure. It is not enough to simply sequence the human genome. We must understand its order, regulation, and coordination – a process in which 2-D gel database analysis will play an important role.

We propose in this paper first to briefly recount how, in GELLAB-I [17–19], we sought to deal with the set of problems common to all 2-D gel database systems using examples from our human leukemia and chicken embryo axonal regulation 2-D gel databases. Next we will indicate the alterations and elaborations embodied in our new system, GELLAB-II. These versions will be compared with the other major published 2-D gel database systems to illustrate these common paradigms. A not exhaustive list of these systems includes TYCHO/KEPLER [9], ELSIE-IV [10], MELANIE [11, 12], QUEST [13], PDQUEST [14], and HERMeS [15, 16] as well as *other* systems including some of these functions [20–25]. A number of these systems are reviewed by Dunn and Burghes [26]. These gel analysis systems are all really performing an 'exploratory data analysis'. Finally, we will sketch possible avenues of future work, including relating 2-D gel databases to gene and protein sequence databases and, ultimately, the goal of developing mathematical modeling capacity within such a unified biologic database system. Although this paper deals with a biological subject, it does so in the context of computer oriented database analysis. For that reason, some of the discussion is multidisciplinary and somewhat technical. We feel this is necessary to adequately describe the algorithms and database problems with which we deal.

2 Materials and methods

2.1 Description of the GELLAB paradigm

All 2-D gel analysis systems commonly perform a data reduction from image space to protein-concentration frequency-distribution space, and finally to a subset of named spots. This data reduction may be thought of as the following sequence of operations: Experimental information and gel images → spot-lists → paired-spot-lists → composite-gel-database → analysis-derived-images-and-spot-lists. The analysis-derived spots can then be used to select proteins to extract from the gel and sequence, or which can be used as marker proteins. In GELLAB, this is performed by a series of programs [17–19], with the following subgoals: experimental study and scanned gel image information acquisition – getacc; spot extraction and quantitation from original scanned gel image – sg2gii; pairing spots between gels – cmpgl2; merge and analysis of paired spot-lists as composite gel database – cgelp2; derived images and plots – markgel, mosaic, dendrogram.

2.2 Definitions of GELLAB terminology

In order to more easily describe GELLAB and compare it with other 2-D gel analysis systems, we define some of the terminology used in GELLAB in the Appendix. Although the terminology is different in other systems, many of these concepts are the same. Since there are so many terms, they are organized by topic: gels, spots, images, database files and data filters. The Appendix may be skipped but referred back to as needed for the rest of the paper. The term database will refer to a multiple gel composite database.

2.3 GELLAB-I description

As this paper deals primarily with 2-D gel database analysis, we will not detail the method of spot quantitation and pairing between gels. This is discussed for our system in [18, 28]. Other systems use different algorithms with the same expectation for data reduction and similar results for creating files of quantified and matched spots between pairs of gels [9, 10, 13–15, 21, 22, 25, 29–33]. Using statistical searches (such as listed in the Appendix) for differences between gel classes, one can find highly correlated marker proteins for use in clinical diagnosis. As an example, Fig. 2 shows an Rmap of gel 324.1 with some of the adult human acute myelogenous leukemia marker proteins (Rspots 106, 235, 273, and 466) found by analysis of the HM5 database [34]. These markers were found using several different statistical tests (parametric and non-parametric) at various minimum probability p-values and verified with mosaic images. The data was further analyzed by performing a complete-linkage cluster analysis using the dendrogram program, illustrated in Fig. 3, of some of these leukemia gels as a function of Rspots 235 and 273. Spots found in a search are saved in a Search Results List or SRL.

One would like to find patterns in groups of proteins indicative of a given state of differentiation as well as putative co-regulated proteins, *i.e.* sets of proteins whose relative synthetic rates show changes under specific physiologic circumstances. These can be visualized using spot expression profiles of proteins under different experimental conditions. Fig. 4a shows some expression profiles of key axonal-growth related proteins under different experimental conditions, and Fig. 4b is the cluster analysis dendrogram of these spots as a function of expression profiles. Another form of analysis uses constraints on relations between several spots in the same gel and between other gels. Constraint analysis has been used in the detection of putative coordinately regulated proteins [35], detection of point mutations [36], and polymorphisms [37]; the quantitative effects of phosphorylated charge shifts are discussed in [38].

2.4 Composite gel 3-D databases

Choices in structuring data representing a problem domain affect how that data can be effectively analyzed. The implemen-

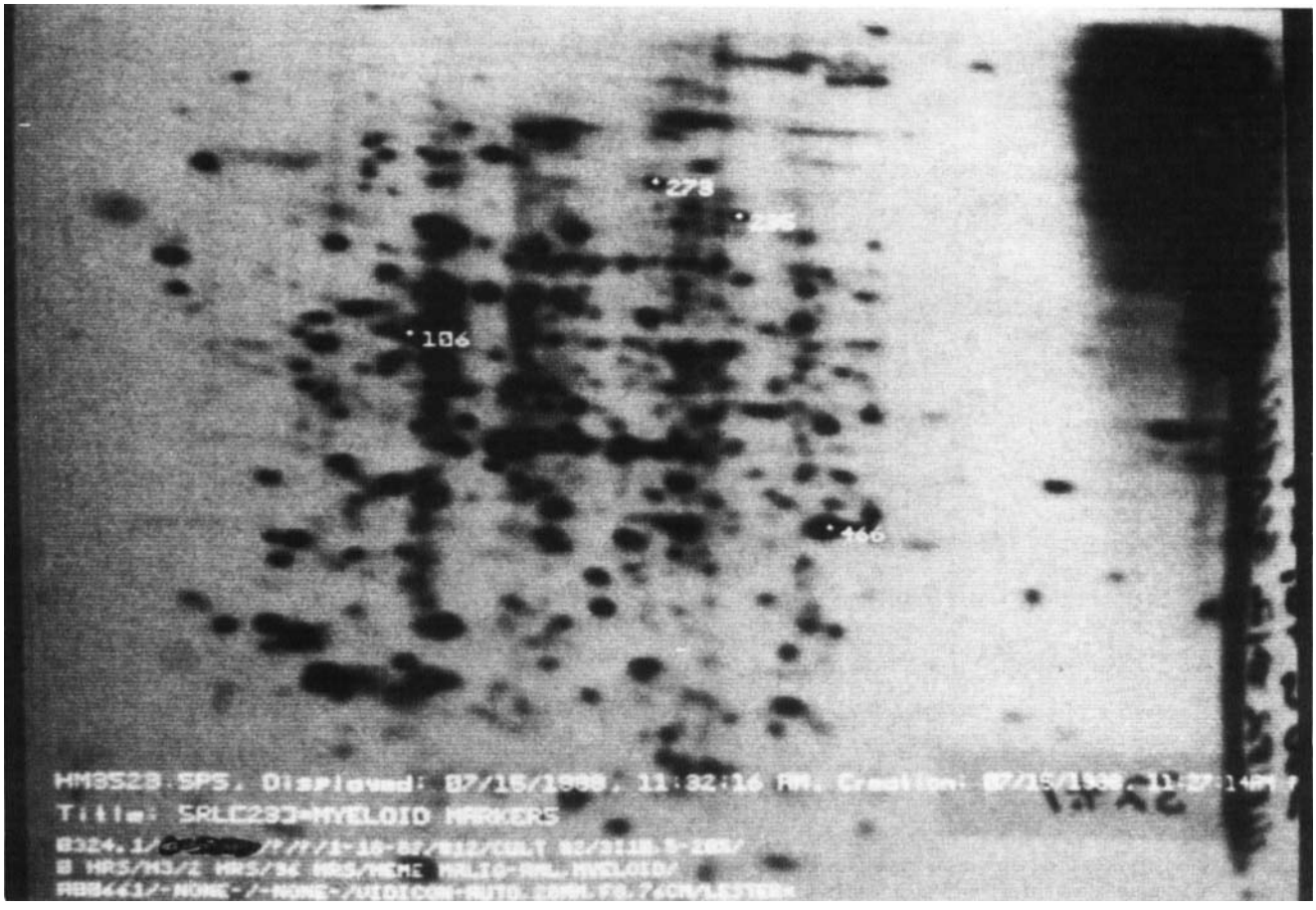

Figure 2. Rmap image of HM5 gel 324.1 with some of the myeloid marker spots 106, 235, 273 and 466 selected as a subset of putative marker spots. The putative marker spots were selected after running both parametric (*t*-test), non-parametric (Wilcoxon-Mann-Whitney) tests at probability *p*-value thresholds of .95 and .99 [60]. Spots surviving these tests were then visually inspected using Rmap and mosaic images to remove false positives and negatives.

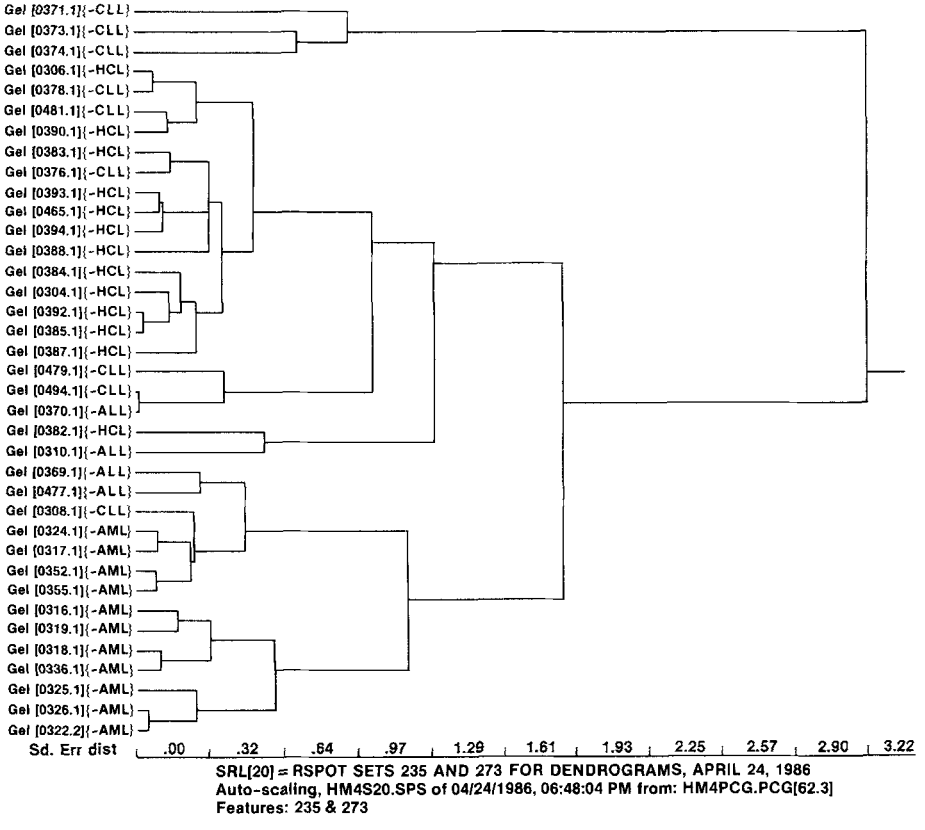

Figure 3. Dendrogram of some gels from the HM5 leukemia database clustered as a function of myeloid marker proteins 235 and 273. Note the AML gels clustered together as would be expected using marker proteins for that class. However, the preliminary data for ALL shows that it is split between two major clusters suggesting possible bimodality within the ALL data.

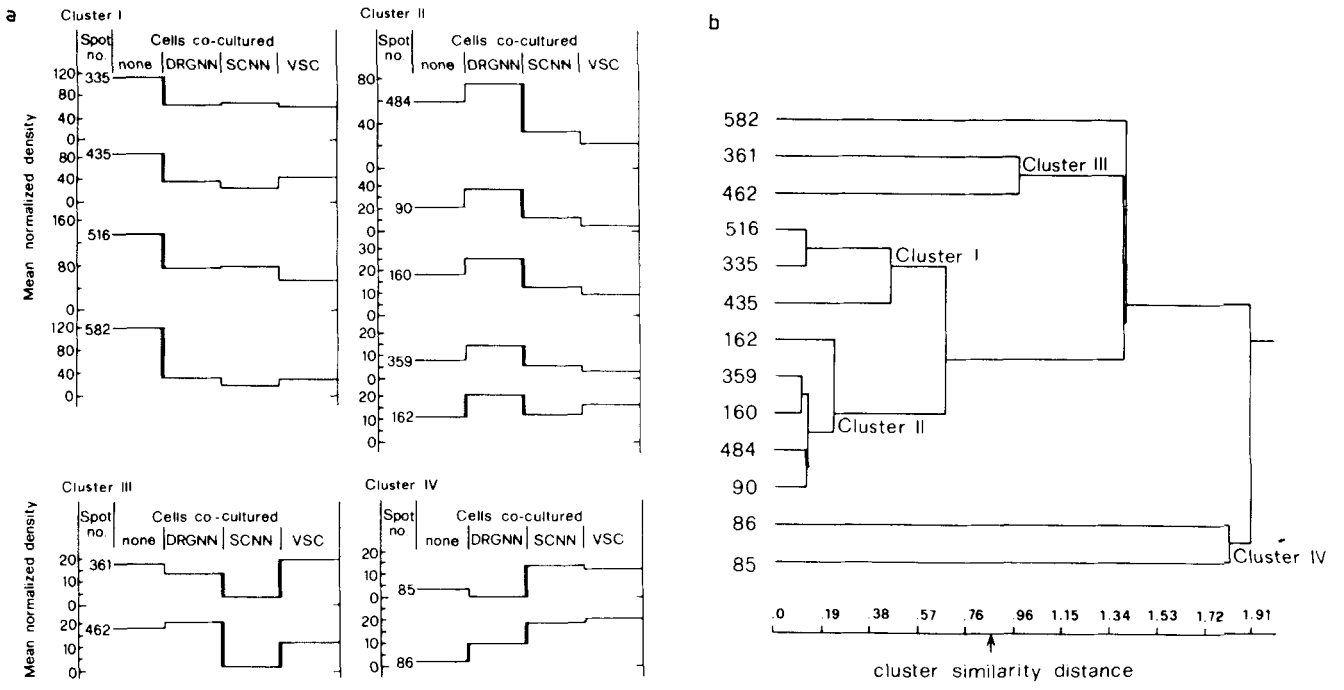

Figure 4. (a) Expression profiles of the 13 proteins previously determined to be modulated by environmental influences. Histogram-type plots of mean normalized density versus coculture condition. The thick vertical bars indicate where statistically significant differences in the comparison of the adjacent experimental classes occur. The expression profiles were manually clustered according to their most prominent feature. Heterogeneity of these groups is obvious at more detailed inspection and has been dealt with by computerized cluster analysis (see Fig. b). Cluster I proteins: proteins with high relative abundance in the absence of any cocultured cells (none). Cluster II proteins: proteins with high relative abundance under local coculture with peripheral nonneuronal cells (DRGNN). Cluster III proteins: proteins with depressed relative abundance under coculture with central nonneuronal cells (SCNN). Cluster IV proteins: proteins with high relative abundance under local coculture with central vervous system cells (SCNN and VSC). (b) Dendogram representing the clustering pattern of the 13 environmentally modulated proteins. Computerized cluster analysis was performed with the 13 modulated spots as the objects and the ratios of the mean values of all permutations of pairwise comparisons of all experimental classes as the features. The numbers to the left of the tree indicate the 2-D gel spot numbers. The abscissa represents cluster similarity distance as measured in standard error distance. As a cutoff criterion of "significant clusters" the mean cluster similarity distance plus one standard deviate was used (cutoff indicated by vertical arrow) This data is taken from [59] and discussed there in the context of coordinate regulation of expression of axonal proteins by the microenvironment.

tation of the GELLAB composite gel 3-D database shown in Fig. 1 may be visualized as a linear random-access file consisting of a state-of-the-database header, followed by Rspot sets Rspot[1], . . ., Rspot[n], eRspot[1], . . ., eRspot[m]. We should point out that the composite database does not contain the original image, raw spot quantification or raw pair-list file data at this final stage. The header 'state' contains current information on which gels are in the database, their current gel accession information, lists of SRL and gel subsets, the working set of gels, prefilter limits, *etc.* From the point of view of the computer program data structures involved, each spot instance from a given gel is stored as a member of a sorted linked-list for each Rspot set. Accessing an Rspot set means random accessing that linked-list using an in-core cache. If that Rspot set (as well as adjacent ones) is already in the cache, then processing will proceed rapidly. The file is always created (or expanded if need be) with additional capacity in each Rspot's linked list so that adding spots (*i.e.*) one spot or more per gel) is not difficult. As noted in the Appendix, Rspots are defined as spots found in at least the Rgel. They may or may not be found in other gels. Similarly, eRspots are spots not found in the Rgel. They are found in other gels and then extrapolated to the Rgel. Therefore in any case, this form of database can represent all of the spots found on any of the gels. Any Rspot may be extrapolated (with zero protein concentration assumed) to any gel in which that Rspot is not present. We may optionally include these EP spots in subsequent statistical calculations.

The Fspot or foreign spot is simply the mapping of the numbering of spots from a similar sample 2-D gel database made in another laboratory with those Rspots in the data base under discussion. Obviously it is not one-to-one given the state of the art of inter-laboratory gel reproducibility. Therefore not all spots can be mapped. GELLAB-II permits the use of mapping files to discourse about the Rspot data base using Fspot numbers. Computer search results need to be validated. This is done by inspecting computer synthesized images. The Rmap image (*cf.* Appendix) can be used to locate potentially interesting spots and remove from further consideration false positive matches to 'noise spots' on the gels. The mosaic image (*cf.* Appendix) is inspected to detect extrapolated spots (EP) which are false negatives (*i.e.* actually exist but are below the limits of detection) or mispaired spots. In GELLAB, any labor intensive visual analysis is done only after statistical analysis of unedited data is performed. This, in order to verify the accuracy of results, greatly reduces the quantity of data which must be reviewed.

2.5 Second order database tools

In addition to the standard types of univariate parametric or nonparametric statistical tests, there are a number of special tools which are useful for second-order analysis (the prefilter and statistical test being the first-order analysis):
Ratio-histogram – histogram of means of each Rspot in two experimental classes for a subset of spots.
Order-table – table of ratios of a subset of spots for all permutations of gel experimental classes for each spot.
Expression-profile-table – table of protein expression of spots relative to the first experimental class.
Correlation tables – correlation coefficient table of a set of gels or of a set of spots with each other.
Dendrogram tree – cluster analysis plot of a set of objects as a function of a set of feature properties of each object; *e.g.* cluster a set of {*gels*} as function of set of {*Rspot*} objects, or cluster {*Rspots*} objects as a function of spot expression profiles in various gels (see Fig. 3 for example.).
Histogram plots – global histograms of spot and Rspot set features.
Density *vs.* density plots – The density of a spot is plotted against the density of the corresponding spot in a second gel.
Density *vs.* experimental-class-associated-values – 2-D scatter plots for viewing global correlations of protein concentration as a function of experimental-class associated feature data. Coordinately regulated spots may be indicated by clustering on the same expression curves.
Ratio-table plots – of a set of individual spots (corresponding to individual gels) as a function of Rspot set number. These are adjacent plots of individual spot (per gel) expression for several Rspot sets.

There are a number of constraint-based searches which do not fall into the normal category of univariate statistics. These include:
Missing-class search – a qualitative test to find spots which do not appear in one of two classes, where the definition of 'does not appear' can be defined to take noise into account.
Least-squares-search – a quantitative search for spots having similar expression response profiles. The profiles are computed as a function of experimental class associated value (*e.g.* culture time, dose, *etc.*). This search could be used for finding spots with the same (assumed linear) dose-response or time-response expression profile.
Coordinate-pair search – a quantitative search for a list of spots that meet coordinated-pairing constraints (discussed under constraint analysis).
%-search – a quantitative search for spots having a ratio > some threshold when mean values for two classes are compared.
Expression-profile-search – a non-parametric quantitative search to find spots with an expression profile (as a function of experimental class) similar to a specified user model expression profile with least square error < some threshold.

2.6 Set operations

By partitioning gels and spots in the composite gel database into their respective subsets, one can refine such subdivisions using the set operations union, intersection, and difference. Of course, the resulting derived sets can be saved and used in other set operations and as part of the prefilter. This is especially useful for comparing results obtained by exploring the database under different views. Another class of useful operators are relational operators which let us search on relations between spots, or relations between previously performed searches based on different partitions of the database. For example, we can find which searches contain a particular Rspot or which searches meet a relational expression criteria. (For example, to find search result list subsets dealing with "AML and CLL but with a probability p-value threshold > 0.8", one enters "AML & CLL & $\sim p > 0.80$"). The ability to add and remove spot numbers from spot sets allows us to implement parts of an annotation database to keep track of spots found subsequently to belong to a particular annotation set.

2.7 2-D Gel spreadsheet

Although we have not stated so explicitly, we are effectively analyzing the gel database as if it were a 2-D gel spreadsheet by looking at the database from different views. The gel database manipulation program cgelp2 contains a "command-language" for manipulation of 2-D gel data views. The language may be thought of as allowing the user to apply "Operators" to "Objects" to produce "Views" as illustrated in Table 1. One example of changing the view is by changing the density normalization method [18]. We could first normalize the database using the ratio-sum method where each spot's density is expressed as a ratio to the sum of the densities of some set of normalization spots for that gel. We could alternatively use a least-squares normalization to take the overall protein concentration of the gel into account. The first is a local and the second a more global view.

Another use of different views is in trying to find gel subclassifications based on evidence discovered in previous views (*cf.* Table 2). Table 3 is a tabular view of rank ordered Rspot set data from the HM5 leukemia database. This shows Rspot 466 HCL data as being putatively bimodal. This is illustrated in Fig. 5 as an idealized frequency histogram of spot concentrations (as a function of experimental class). Plotting this data makes the bimodality of the HCL data obvious. Fig. 3 shows a dendrogram in which gels from the leukemia database are clustered as a function of myeloid markers showing putative bimodal ALL data. These types of results suggests that no assumption of sample normality be made and other multi-unimodal tests be applied.

Table 1. 'Operators' applied to 'objects' produce 'views'[a)]

Objects	Gels, spots, sets of spots, sets of sets, parameters
by mapping them with various	
Operators	Prefilter operators Set operators: union, intersection, difference, Statistical search operators: parametric, nonparametric Constraint-based search operators: pair-spot-ratios, fuzzy test for missing-spots
can generate different	
Views	Tables of: Rspot, correlation, histogram, *etc.* Lists of: spots, gels Gray-scale and color image displays and plots of: mosaics, Rmaps and feature-feature plots, correlations, histograms, dendrograms.

a) Based on analysis of the current views one adjusts the objects and operators iteratively, deriving new objects when attempting to converge on the essential class differences characterized by sets of spots.

Table 2. Example of changing the 'view' of a lymphocyte leukemia database[a)]

View	# Classes	All-classes	
Main cell lineage	2	myeloid	lymphoid
Primary leukemia diagnoses	4	AML	ALL CLL HCL
Putative secondary subclasses	> 6	AML1 + 2	ALL1 + 2 CLL1 + 2 HCL1 + 2

a) The two subclasses ALL1 and ALL2 are suggested by the dendrogram cluster analysis of the four classes of gels as a function of myeloid marker proteins (*cf.* Fig. 3); The two subclasses HCL1 and HCL2 are suggested by the bimodal distribution of Rspot 466 in the HM5 database [34] (*cf.* Table 3 and Fig. 5).

Table 3. Protein quantitative data for HM5 Rspot 466 indicating possible bimodal distribution for HCL class gels[a)]

Gel ACC #	Class	Normalized spot density	Gel ACC #	Class	Normalized spot density
0369.3	2	294.78	0352.1	1	32.48
0388.1	4	294.11	0355.1	1	17.91
0389.1	4	251.47	0382.2	4	14.94
0384.3	4	239.38	0370.1	2	14.66
0333.1	1	183.66	0382.3	4	14.03
0369.1	2	180.69	0382.1	4	13.81
0338.1	1	148.88	0354.1	1	10.75
0339.1	1	142.52	0356.1	1	9.11
0336.1	1	119.05	0376.1	3	7.19
0369.2	2	117.71	0360.1	1	6.86
0324.2	1	107.93	0374.2	3	6.44
0324.1	1	107.70	0374.3	3	6.30
0324.3	1	105.41	0353.1	1	6.20
0337.1	1	100.27	0374.1	3	6.13
0326.3	1	74.33	0390.1	4	5.06
0326.1	1	72.71	0378.3	3	3.50
0346.1	1	71.13	0376.2	3	3.39
0383.1	4	70.77	0378.1	3	3.36
0326.2	1	69.88	0378.2	3	3.00
0325.1	1	34.12	0377.1	3	1.48
0325.3	1	34.02	0385.1	4	1.38
0325.2	1	33.06			

a) Rspot set data for Rspot[466] for 43 gels taken from the HM5 leukemia lymphocyte database. Note suggestive bimodal HCL, class 4, distribution. In the table, ACC # is the gel accession number uniquely defining each gel sample, class is the experimental class (1 = AML, 2 = ALL, 3 = CLL, 4 = HCL, the density is the value of normalized integrated spot density (protein concentration). The data is taken from the HM5 adult human leukemia database [60]. Rspot[466] is also one of the marker proteins differentiating AML and CLL.

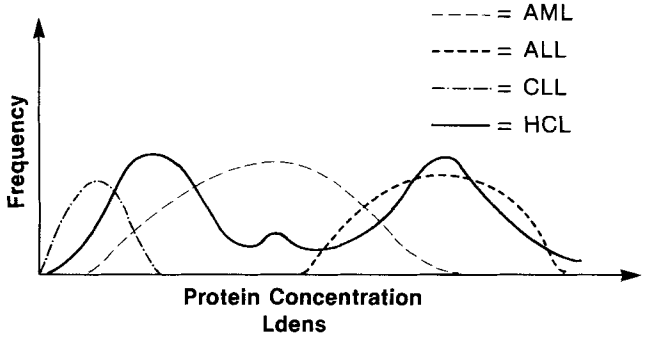

Figure 5. Illustration of frequency distributions of protein concentration of four classes of gels for HM5 leukemia database Rspot set 466. The histogram data from Table 3 are drawn as smooth frequency distributions to indicate possible trends in this type of data – not the actual data itself. It indicates a putative bimodal frequency distribution plot for HCL class gels for this polypeptide which is more apparent when graphed. Such bimodal distributions suggests possible subclasses. One consequence of such results is that univariate model tests should not be applied blindly since they assume unimodal distributions. Protein concentration is in least-square normalized integrated density units.

In addition to using different operators and normalizations, we can change the views by adjusting prefilter parameters to change the view of a slice into the '3-D' gel database. Some of the parameters often changed include: subset of gels in the working set of gels, experimental classes partition, normalization method (ratio to particular spots, least-squares normalization to range of Rgel, % total density of a gel, *etc.*), set operations on different search results, and externally imposed constraints. We can use relational and set operations on sets of spots or sets of gels to describe higher level views of the database, *eg.* annotation of properties of gels or spots. Another use of the prefilter is to reduce thrashing of the database between main computer memory and the disk database file. For example, by restricting the region of the gel under search, only that part of the database required will be brought into memory during the search – thus speeding up the search. Generally, one would iterate between setting of prefilter and search

Table 4. Example of failure code histogram used to help adjust prefilter limits[a)]

Prefilter 'failCode' Histogram		
Frequency	Test #	NAME OF TEST
341	0	PASSES ALL TESTS!
000	1	Rspot set is NULL
000	2	No gels in Working Set meet label, DP, DL or OD-range limits
382	3	# gels outside of #-gels-required per class limits
000	4	# gels meeting relative (dx, dy)-from-landmark limits = 0
000	5	Rspot set Rgel position outside of pI-MW region
356	6	No gels in working set meet spot-pairing-label limits
064	7	No gels in working set meet DP limits
000	8	No gels in working set meet DL limits
000	9	No gels in working set meet OD-range limits
003	10	Failed mean AREA statistics
203	11	Failed mean DENSITY statistics
000	12	Failed coefficient of variation of AREA statistics
000	13	Failed coefficient of variation of DENSITY statistics

a) Failure code statistics are an optional result of performing a search through the composite spot data base. During the search, if the prefilter fails for any reason for any spot, that reason is counted in a histogram. These data base search statistics were from a search of a subset of the leukemia database [60]. For this search, the histogram indicates that the prefilter limits were set so that about 25 % of the Rspot sets passed the prefilter. By adjusting those prefilter values with the highest rejection rates (for example, Code 03 #gels-required-per-class), one could increase the number of Rspots passing the prefilter. If we failed on the lower bound of a prefilter limit, then the next time the limit would be decreased whereas failure on the upper bound would cause the upper bound to be increased. In the integrated expert-system being added to GELLAB-II, this will be suggested (and optionally performed) automatically.

probability *p*-value threshold parameters for optimizing results of statistical tests. Failure-mode statistics can be accumulated and used to aid the setting of prefilter parameters. Table 4 shows a sample failure-code histogram which indicated that in the previous search of the database, most prefiltered spots were rejected because the number of gels required per experimental-class was out of range. This suggests that the lower limit for this parameter might be lowered further in order to permit more spots to be analyzed.

2.8 Composite group of spots

It is sometimes useful to postulate that a group of spots on the same gel are the result of fragmentation of a single spot caused by some systematic process whether in the biology, the production of the gel or the image processing. We can define a composite group of spots (CP) and the method used to track fragmented or posttranslationally modified spots [39]. Posttranslational modification chains in 2-D gels are discussed in [1, 38]. The group of spots constituting a CP may be different on different gels and are defined, in GELLAB-I, during the spot pairing phase of the analysis. They could also be defined after the paged composite gel database was constructed. Any system treating a group of spots as a single spot needs to specify which spots in which gel are in any particular group. In general, for corresponding groups in different gels one may not always have all members of a group present in all gels. GELLAB-I allows manual entry of CP spots as a set of GSF spot numbers specific to a CP spot for each gel. A spot is denoted by cc_j, *i. e.* its connected-component number *j* (see GSF definition in Appendix) defined by the spot segmentation (quantification) algorithm. Denoting CP spots in GELLAB-I is done by manually editing a file of entries of the form:

$$CP_{\text{spot-name, ACC \#}} = \{ cc_i, cc_j, \ldots, cc_n \} \quad (1)$$

e.g.

$$CP_{\text{III,0056.1}} = \{ 295, 296, 297, 298 \} \quad (2)$$

Then spot features are automatically computed for the CP spot. For example for a CP spot *z* in gel *g* as illustrated for some of the following features: integrated density as $D_{zg} = \Sigma_{i \epsilon CPzg} D_{ig}$, area as $A_{zg} = \Sigma_{i \epsilon CPzg} A_{ig}$, and the estimated CP centroid is:

$$X_{zg} = \frac{\Sigma_{i \epsilon \text{ CP zg}} X_{ig} \cdot D_{ig}}{D_{zg}} \quad (3)$$

$$Y_{zg} = \frac{\Sigma_{i \epsilon \text{ CP zg}} Y_{ig} \cdot D_{ig}}{D_{zg}} \quad (4)$$

Having manually defined the connected-component numbers in group spots *z* and CP_{zg} for a set of gels, program cmpgl2 will declare these as additional spot CC numbers. These in turn are paired (by definition) and are given a CP spot pairing label. Note spots constituting a group could be counted twice in the database if we are not careful. Therefore, if the CP label is used in the database prefilter, spots which are part of a group are not tested individually, but only as part of a CP spot. They may alternatively be viewed as individual spots by setting the prefilter to ignore CP Rspots.

2.9 Constraint analysis

Views can also be specified by constraint analysis models. We illustrate one example of constraint analysis below, while another might be the detection of posttranslational modification chains of spots and another the detection of polymorphism. The search for putative precursor-product pairs is an example of a constraint analysis search based on a particular spot pair model [35]. This is sketched out in Algorithm 1.

Algorithm 1: Finding putative precursor-product pairs. The object is to find pairs of spots constrained by limits on some of their properties. Pair-ratio constraints can be used to find putative precursor-product protein pairs in which, for example, a small peptide is cleaved from a precursor yielding the product.

$$\text{precursor}_1 = (signal_2 + product_3). \quad (5)$$

This model assumes a small peptide with small or no resulting p*I* change. Then the following constraints can be used to find putative precursor-product spot pairs:

$$|pI_1 - pI_3| < T_{\text{pI}}, \quad (6)$$

and

$$|MW_1 - MW_3| < T_{\text{MW}}, \quad (7)$$

where T_{pI} and T_{MW} are limits meeting the expectations of the model and the protein concentrations (*O* and *o*) pair-ratio for spots 1 and 3 for two gels a and b

$$(O_1 + o_3)_a/(o_1 + O_3)_b \sim 1.0 \quad (8)$$

Constraint-analysis-based spot search models are being increasingly investigated [15, 35, 40]. This type of constraint-based search algorithm could also be applied to finding putative point mutations (discussed in [36]), or polymorphisms in genetic studies [37, 41] both of which exhibit similar small shifts in *pI* and/or *MW* [38]. Constraint analysis has also been proposed using predicate calculus and relational database models [15, 16].

3 Results

3.1 GELLAB-II: Enhancements of GELLAB-I

GELLAB-I was written in the SAIL language [42] which only runs on DEC 10/20 systems. Over the past few years, we have developed a SAIL to C compiler called PSAIL [43, 44] which has been used to translate the approximately 70 000 lines of GELLAB SAIL code to C. All of the code is currently translated to C and is being converted to run in the UNIX* operating system using the portable X-window System graphics environment. The portable UNIX C version of GELLAB-II is available for SUNs and microVAX-II class computers. Later non-UNIX versions such as for PC/386, VMS and other systems with C compilers will be considered. GELLAB-II uses the X-window System [45] for providing a

* UNIX is a trademark of ATT; microVAX and VMS are trademarks of DEC; POSTSCRIPT is a trademark of Adobe Systems Inc.

portable color display and an optional interactive pop-up menu system to aid the inexperienced user. X-windows now run on a large number of UNIX systems and will be running on several non-UNIX systems (VMS and MS-DOS) as well. The Xpix application program [46] has been written as an image interface between X windows and GELLAB-II and provides us with portable image graphics as well as the ability to flicker-compare gel images. GELLAB-II also has online HELP manual documentation based on keyword search. Another graphics extension to GELLAB-II allows the generation of portable Tektronix 4010 as well as POSTSCRIPT language graphics output. POSTSCRIPT can be used with many laser printers to give rapid high quality plotter output for derived gel data. The Tektronix output enables the GELLAB user to generate graphics output when running on "dumb" terminals where a workstation environment is not available. GELLAB-II has increased many of the GELLAB-I limits: larger images (up to 4096 × 4096) rather than the 512 × 512 pixel images of GELLAB-I; up to 8000 gels and 16 000 Rspot sets per paged composite gel (PCG) database. Image compression is also used and is essential to preserve disk space – especially when using these larger images. Derived images are still 512 × 512, which is adequate to visualize proteins of interest while minimizing storage requirements.

We plan to use a high resolution charge coupled device (CCD) scanner: Data Copy model 612 F with 1728 × 2400 × 8-bit 0-2.3 optical density units, normally scanning a 1024 × 1024 × 8 bit image. This is connected *via* a standard small computer system interface (SCSI) to the SUN and is relatively machine independent. (This UNIX driver, written by Chris Johnson of Falcon Systems, will be made available by NCI) Scanner dependency is restricted to one small program module. However, using the ppxcvt program, GELLAB can use image data from other sources such as TV cameras, an optronix scanner, and other CCD/photodiode array scanners. Command history (also called journaling) and macro substitution, like that in the UNIX C shell history mechanism, can be used to save typing and to keep a record of all database changes in the cgelp2 database program.

An expression-profile (see Fig. 3 and [59]) is a pattern consisting of an ordered list of mean protein concentrations of a particular protein as a function of experimental-study class. For n classes, let protein k for class c have density m_{ck}, then the expression profile is an n-tuple: $(1.0, m_{2k}/m_{1k}, \ldots, m_{nk}/m_{1k})$. GELLAB-II has two new statistical operations for finding and presenting expression profiles. The ordered expression profile table operation creates two tables for selected Rspot sets: (i) Rspot number *vs.* expression profile, and (ii) Rspot number *vs.* an ordered list of other Rspot numbers in that set ranked by minimum least square error between the profile of the spot being considered and that of each spot in its list. Only those with minimum least square error < a specified threshold are reported. Those reported, indicate clusters of Rspot sets with similar expression profiles. The other operation is an expression-profile search through the database looking for Rspot sets whose expression profile is similar (in the same sense just discussed) to a user-specified expression profile ratio and which have a least square error < some threshold. The cgelp2 database analysis program also runs in either an interactive menu or command-driven user interface. The latter can optionally be fed from a batch script with results going into a log file. GELLAB also provides a higher level user interface, makjob, which can generate GELLAB command script files for

Table 5. Possible objects for object oriented 2-D gel databases[a)]

(A) Name
- (A.1) Name of protein for Rspot set or Composite Rspot Set
- (A.2) Alternative name(s) of protein – if discrepancy

(B) Amino acid sequence (through pointers to GenBank, EMBL, *etc.*)
- (B.1) Amino acid sequence (or name of sequence) for polypeptide spot
- (B.2) Alternative sequence(s) or sequence fragment(s)
- (B.3) Genetic variant/polymorphism(s) sequence(s)
- (B.4) Chromosome location(s)
- (B.5) Amino acid composition(s)
- (B.6) Gene structure(Intron/Exon, *etc.*)

(C) Disease
- (C.1) Disease(s) where marker-spot present or increases, and % concentration w/variance
- (C.2) Disease(s) where marker-spot absent or decreases, and % concentration w/variance

(D) Structure
- (D.1) Structure/location(s) where protein occurs in the cell
- (D.2) Crystalographic structure(s) references
- (D.3) NMR structure(s) references

(E) Function
- (E.1) Protein's function(s) in cell
- (E.2) Subcellular localization(s)
- (E.3) Enzymatic activity of this polypeptide
- (E.4) Multimeric-protein(s) it is part of if any and
  - (E.4.1) Number of subunits present

(F) References
- (F.1) Reference(s) of papers detailing that spot
- (F.2) Estimated value(s) and variance of *pI*, *MW* – published
- (F.3) Source(s) – who has or sells the antibody

(G) Protein maps
- (G.1) Mapped Rspot # in GELLAB-II standard PCG DB
- (G.2) Mapped Rspot #(s) in other system's spot DB(s) – 2-tuples (DB-name, Fspot#)
- (G.3) Local morphologic neighborhood naming Rspots adjacent to this spot
- (G.4) Statistics on reliability of local morphologic neighborhood of spot
- (G.5) List of composite group spot member spots (if this is a composite group of spots)

(H) Regulation of this protein
- (H.1) List of other Rspot #s which appear to be co-regulated with this spot
- (H.2) Inducers/repressors for this polypeptide
- (H.3) Post-translational modifications of this polypeptide
- (H.4) Half-life of this polypeptide
- (H.5) N-tuple association of one Rspot with other Rspot(s) or 'object' features

(I) Taxonomy
- (I.1) Species
- (I.2) Strain
- (I.3) Organ
  - (I.3.1) Cell type

(J) Experimental external data
- (J.1) Researcher(s)
- (J.2) Sample preparation and time stamp
- (J.3) Gel preparation and time stamp
- (J.4) Experiment description

(K) *ETC*

a) These are some features of polypeptides to be stored in 'slots' in a possible object-oriented database. The object may contain data or pointers to entries in other objects or to external file databases (probably the most common implementation because of the amount of data and duplication involved). For all slots – only those which are known need be entered into the database. All entries are source (*i.e.* where it came from) and time stamped. This compilation of entries was derived from a number of suggestions in many of the 2-D gel papers referenced in this paper.

many of the routine GELLAB programs. These scripts are automatically generated, based on user-prompted problem description using 4th generation program generation techniques – thus greatly magnifying analysis power without the user having to worry about the details.

3.2 GELLAB-II future enhancements

GELLAB-II will be using object-oriented database techniques for merging 2-D gel and other types of information. Some of the information we might want to have available in a database are listed in Table 5. Relational databases can be a subset of object-oriented databases; therefore relational techniques are not excluded. We will be able to apply (through additional UNIX processes) new data analysis programs to prefiltered database data without leaving the GELLAB environment. These new programs can then be thought of as command extensions and treated the same way as other built-in commands. In GELLAB-I, this was done by exporting cgelp data through SPSS files which were then analyzed by running separate programs after exiting the cgelp database program. Similarly, GELLAB-II will merge the ability to view Rmaps, mosaics and dendrograms, *etc.*, without leaving the GELLAB cgelp2 database program. Any number of programs or copies of cgelp2 can read the same gel database simultaneously to analyze the same data in different ways. However, only the owner of the file can change the database.

We will examine additional automation tools for the CG problem. One can use tight constraints to look for a chain of spots as a model of a posttranslational modification chain. Such constraints might include: (i) *pI* slope, (ii) inter-Rspot-spacing (as suggested by [47]), (iii) variance about a line fit through them, (iv) variation of size of spots in the group (related to their order in the chain) and (iv) periodicity of the spots in the chain. Currently, however, no automatic algorithm can be completely trusted, especially when additional problems such as different numbers of spots in the chains in different gels and interfering neighboring spots which may shift MW are taken into account.

3.2.1 Expert systems

We have been experimenting using expert-systems ([48] is a good introductory text). One application is as an aid in running software such as the cgelp2 exploratory data analysis part of

Table 6. Example of part of expert-system rule set for helping select statistical test[a)]

```
TITLE: STATISTICAL-TEST-TO-USE-RULES,

@Rules: STATISTICAL-TEST-TO-USE-RULES
R1:  if CHANGES-EXPECTED
     then assert: DO-TESTS
R2:  if @not CHANGES-EXPECTED
     then assert: IS-DON'T-DO-TESTS
R3:  if DO-TESTS and QUALITATIVE
     then assert: IS-MISSING-CLASS-TEST
R4:  if DO-TESTS and QUANTITATIVE and EXPECT-GAUSSIAN-DISTRIBUTIONS
     then assert: PARAMETRIC-TESTS
R5:  if DO-TESTS and QUANTITATIVE and @not EXPECT-GAUSSIAN-DISTRIBUTIONS
     then assert: NON-PARAMETRIC-TESTS
R6:  if PARAMETRIC-TESTS and @val NBR-CLASSES @== 2 and @val GELS/CLASS @> 1
     then assert: T-TESTS
R7:  if PARAMETRIC-TESTS and @ val NBR-CLASSES @ > = 2 and @ val GELS/CLASS @ > 1
     then assert: IS-F-TEST
R8:  if T-TESTS and EQUAL-VARIANCE
     then assert: IS-STANDARD-T-TEST, IS-T-TEST
R9:  if T-TESTS and @ not EQUAL-VARIANCE
     then assert: IS-BEHRENS-FISHER-T-TEST, IS-T-TEST
R10: if NON-PARAMETRIC-TESTS and 2-CLASSES and @val GELS/CLASS @== 1
     then assert: IS-%CHANGE-TEST
R11: if QUANTITATIVE and NON-PARAMETRIC-TESTS and @val CLASSES @== 2 and @val GELS/CLASS @> = 3 and
       @val P-VALUE @> = 0.80
     then assert: IS-WILCOXON-MANN-WHITNEY-TEST
R12: if QUANTITATIVE and NON-PARAMETRIC-TESTS and @ val CLASSES @ > = 2 @ val GELS/CLASS @ > = 5
     then assert: IS-KRUSKA-WALLIS-TEST

@Hypotheses:
IS-F-TEST
IS-STANDARD-TEST
IS-BEHRENS-FISHER-T-TEST
IS-WILCOXON-MANN-WHITNEY-TEST
IS-KRUSKA-WALLIS-TEST
IS-%CHANGE-TEST
IS-MISSING-CLASS-TEST
IS-DON'T-DO-TESTS
@end
```

a) The notation expresses facts or hypotheses as UPPER-CASE-WORDS, lower case words and '@' prefixed words are rule connectives and data evaluation operators, respectively. For example '@val CLASSES' will return the current number of experimental classes currently visible in the PCG database and which we are using to compare gels. Then, the predicate '@val GELS/CLASS @> 1' is true if it evaluates to true – *i.e.* there is more than one gel in a class. The objective of using the expert-system with the above rule set is to select one of the hypotheses which makes sense given the current state of the GELLAB database. Information which is not found in the database will be prompted for from the user.

GELLAB. Table 6 shows part of an expert-system rule-set which can be used to help select a statistical test to apply to the database. We are investigating rule sets in parameter selection by analyzing "failcode" histogram data (*cf.* Table 4) as well as normalization method selection, detection of correlated spots, constraint analysis, gel classification, assisting in batch script generation, and the use of auxiliary databases in the context of 2-D gels. Tukey [49] discusses some of the uses of expert-systems for exploratory data analysis. In particular, expert-systems can be especially useful for helping occasional or novice users. Because of the complexity of any 2-D gel analysis system, a system which worries about forgotten or unknown details (and prompts the user where appropriate) will be much more usable by users who cannot and do not want to remember all of the idiosyncrasies of performing an analysis.

We will be adding the ability to GELLAB-II to use spot data from other sources. GELLAB-I was only able to read GCF spot pairing files. GELLAB-II will be able to access other GELLAB gel databases through mapping files so that different Rspot numberings for different data bases can be translated. In contrast, GELLAB-I only allowed analyses of a single gel database – although that database could be extended by adding additional gels. GELLAB-II will also be able to use data from non-GELLAB spot pairing files, (such as matchset or match-pair type data files [10, 13] to construct a gel database file. Thus data analysis will be independent of the particular spot pairing and spot quantitation method used. It will also be able to merge and use non-gel spot-associated information with the gel database. Algorithms 2 and 3 show how this mapping can be done by mapping matchset or match-pair data files into the single GELLAB database file. The first converts match-pair data into a form required by cgelp2 to build the database. The second shows how one might extrapolate landmarks for gels where they are missing or new landmarks are being propagated so that spots can be extrapolated.

3.2.2 Algorithm 2: Conversion of match-pair data into Rspot and eRspot sets

Because a set of match-pair files and a set of gel comparison files for the same set of gels contain equivalent paired-spot transitivity information, conversion is fairly straightforward. (i) Pick any gel as the Rgel and create the initial Rspot sets from it. (ii) Scan the match-pairs repeatedly, looking for spots which are in the Rspot set, and add that (gel, spot) pair to that Rspot set. If a match-pair entry does not get entered into the database during one pass, it will enter at some later pass if it is linked through some other gel to the Rgel. (iii) Finally, when no more changes can be made, for each match-pair file in turn, put the match-pair spots not in any Rspot sets into eRspot sets and iterate until there are no more spots added. Using Local Morphologic Neighborhoods (described in Algorithm 3), estimate missing spots for gels not present and extrapolate them into the Rspot and eRspot sets.

3.2.3 Algorithm 3: Local morphologic neighborhood replacement for landmark spots

The local morphologic neighborhood (LMN) algorithm described here can estimate "effective landmarks" for gels where the landmark spot is missing. It can also be used for generating pseudo-landmarks (as spots are propagated through the gel database using Algorithm 2 above). We will be replacing landmark spots with LMN's in future cgelp2 versions so that paired-spot data from other gel analysis systems can be used with GELLAB-II. Cgelp2 currently has the data structures and algorithms in place to handle this type of data. Each Rspot has its own LMN which is a list of the numbers of all neighboring Rspot sets found within any gel for which the Rspot is defined such that they are constrained to be within a local neighborhood region. Briefly, for any particular gel, only some of these LMN spots will be present. The algorithm lets us extrapolate a spot in that gel even though the entire LMN set of adjacent spots is not present. Each Rspot's LMN is stored in the associated data dictionary for that Rspot set. As will be discussed, each spot (*i. e.* gel) in the Rspot set will have a subset set of the LMN Rspots visible to it. As will be discussed, each spot (*i. e.* gel) in the Rspot set will have a subset set of the LMN Rspots visible to it. For example the 'LMN' for Rspot [273] might be:

$$R_{\text{neighbors}}(273) = LMN(Rspot_{273}) = \{25, 76, 77, 235, 236\} \quad (9)$$

Each spot in the Rspot set has one or more of these R_{neighbor} Rspot numbers in its own LMN.

$$Rneighbors(j) = \bigcap_{\text{Rspot j } \epsilon \text{ Gel k}} LMN(spot[Rspot_j, gel_k]) \quad (10)$$

The subset of Rspots in $R_{\text{neighbors}}$ can be used to compute an effective landmark and displacement vector (Dx, Dy) (from the landmark to the estimated extrapolated EP spot) for any gel for each Rspot set. As mentioned, each spot (or corresponding gel position) in the Rspot set contains a subset of LMN spots. The default LMN of the current gel pairing schema is simply the interactively defined landmark spot. If a landmark is missing for a particular gel, we can estimate its position and (Dx, Dy) for that gel in order to calculate the estimated Rspot position that it would have for that gel.

Given a set of LMNq spots ($Rspot_a$, $Rspot_b$, ..., $Rspot_k$) associated with Rspot q then the LMNq centroid (in Rgel space) is estimated by,

$$(\bar{X}, \bar{Y})_{\text{LMNq}} = \frac{1}{N_q} \cdot \sum_{\text{Rspot j } \epsilon \text{ LMNq}} (X_j, Y_j) \quad (11)$$

where N_q is $|LMN_q|$.

The displacement vector for the LMNq centroid to one of the LMNq spots j is

$$V_{\text{LMNqj}} = (D_x, D_y)_{qj} = (X, Y)_{\text{LMNq}} - (\bar{X}, \bar{Y})_j \quad (12)$$

Then for any gel g missing from the Rspot set q, but for which at least one of the LMNq spots exists in subset LMNgq, the estimated centroid $(X, Y)_{gq}$ may be extrapolated by

$$(X, Y)_{gq} = \frac{1}{N_{gq}} \cdot \sum_{i \epsilon \text{ LMNgq}} ((X, Y)_{\text{Rspot i}} + V_{\text{LMNqi}}) \quad (13)$$

where N_{gq} is the number of i such that ($i \varepsilon LMNgq$). The assumption used in this extrapolation is that the magnifications for the gels are similar and differences can be neglected when working in small regions such as the LMN.

4 Discussion

There are a number of different areas dealing with the 2-D gel data reduction problem and we will discuss several of them to show what the current practice is and where it appears to be heading. These are: data reduction paradigms and organization of 2-D gel databases to meet the needs of these paradigms; parametric data analysis as a possible global gel analysis model; comparison of cluster analysis techniques as one of the components of the data reduction model; some possible options for extending future 2-D gel databases to reflect the needs of new types of analyses; and finally, perspectives for future 2-D gel database structures using new types of computer database tools.

4.1 Data reduction paradigms

We want to reiterate the point that there are common problems – some of which are addressed differently by different groups. Many of these systems are constructed in a more or less modular fashion so that one could theoretically substitute different algorithms or program sub-systems for spot quantitation, spot comparison, statistical analysis or clustering algorithms, and recording results. Ideally, any system should be modular enough to allow the researcher to easily investigate other algorithms. In practice, researchers in the electrophoresis field have not caught up with software engineering. The paradigms used in structured data analysis have several characteristics: (i) access to all stages of data reduction simplifies the analysis by facilitating back checking and adding additional computations when the research dictates; (ii) structuring data along the lines of the problem domain simplifying the analysis; (iii) ability to extend database concepts and corresponding structures; (vi) the ability to generalize the data. Tools are generally available to aid in both modeling and higher level database maintenance.

GELLAB-I's analysis sequence is: prefilter, test, cluster analysis. This approach approximates the results of the principal component analysis methods, but does not rank-order spots in order of correlation. However, one might rank-order a set of previously determined statistically significant spots by probability p-value to get similar results. GELLAB is designed to use overtly designed prefilters which can be easily changed, tested and used in a non-subjective manner. If there are large volumes of data, one can not hope to review and edit all this data, and thus automatic detection of marker spots becomes increasingly useful. In experiments, one prefers using replicate gel samples which can then be used for statistically valid database analysis. This is as opposed to using a single image which has been subjectively improved using interactive graphics. Thus the use of the GELLAB analysis sequence allows us to handle statistically the experimental noise always found in real gel data.

4.2 Organization of 2-D gel database

Table 7 compares 2-D gel database organizations for some of these systems. Different types of data structures are required for sequential statistical searches rather than for creating the initial spot pairing structure (which is a type 2 organized by gel). For that reason GELLAB uses type I data (organized by spot) resulting in more efficient sequential searches. That is, $\{gels\}$ are sorted by spot so that spot expression profile data is available efficiently. However, this organization is at the cost of making updating the database somewhat more difficult. In practice, the additional overhead of adding a gel is less than one minute/gel, is only done once, and is not really significant in the context of the total analysis. In either type of structure, performing an analysis of some small part of the database involves accessing subsets of $\{spots\}$ and $\{gels\}$. What is being analyzed determines how efficient the database access is. If an operation is to be repeated, it may be cost-effective to generate a temporary derivative database structure in a more convenient form – a paradigm both GELLAB and other systems use.

4.3 Parametric data analysis

Differentiation during cell development or reaction to the environment often causes drastic switches in cell physiology between different states [51]. In discussing a common mechanism of transcription of yeast and mammals, Guarente [53] suggests the importance of an interacting network: "... an interacting network of regulatory proteins appears to lay down the program of spatial development in drosophilia cell types." We suggest that genetic expression in cells may be modeled as a network capable of discrete states determining the expression of each gene. Thus parametric models of pro-

Table 7. Comparison of 2-D gel database organizations[a)]

Database type	Advantages	Disadvantages
(1) Gels sorted by Rspot in sequential random access file	One file for faster access to whole DB, faster DB search looking at data for gels of same Rspot.	If preallocate space for more gels then waste space otherwise more time computing to add gels. To add gel (a) add spot list, (b) add spot pairing file with/Rgel, (c) merge spot pairing file w/PCG DB.
(2) Rspots sorted by gel in separate gel spot list file	Easier and faster to add new gel: (a) add spot list, (b) add spot pairing file which is the DB.	Too many directories-files slows down analysis if large # gels. Slow search through DB when look at all spots for gels of same Rspot. Worst case is N × M match-files for N gels.

a) For a case (1) sweep through the database, the data is readily available for a sequential search since all spots in the DB have data for all gels for each spot immediately available. In case (2), for M gels and N spots, this requires doing NxM file lookups and scans for the particular (spot, gel) pair of interest. If the file system used random access of single gel file data and M gel files were opened simultaneously, then one could keep a cache for each file and the overhead would not be that bad. GELLAB is a type (1) DB while ELSIE/MELANIE, QUEST, HERMeS are type (2). in type (2) systems. DB data may be extracted from match-pair files into a different view of the database (file or table which can then be treated as type (1) DB. Although GELLAB final form DB data is type (1), the Gel-Comparison-Files are equivalent to type (2) data – the linkage however is made through Rspots in the type (1) DB.

tein j's expression e_j can be considered as a function of (i) environmental conditions, (ii) expression of a subset of other proteins e_j^* – a subset k of all proteins in the organism E, and (iii) time t. Eq. (14) illustrates this general statement. We can refine this model (Eq. 15), which is suggested in the following discussion.

$$e_j(class, envirn, e_j^*, t) \approx e_j(class, envirn, \{e_i|_i \, in \, SRL_k\}, t) \quad (14)$$

"Chaos theory" is an area of mathematics dealing with modeling complex systems which exhibit radically different states of behavior. For example, behavior seen during cell differentiation, heat shock, *etc.* Such behavior is not always predictable using a linear extrapolation of previous events and is often sensitive to the initial state of the system. Many such chaotic systems can be modeled using parametric nonlinear differential equations (NLDE) as one way of describing a system with multiple global states which can pass from one state to another [54]. Busse has modeled nonlinear parametric chemical network oscillators using a system of NLDE [55]. Much of the work on understanding stable and unstable states (called points of attraction and saddle points, respectively) had been done by Poincaré [56], as noted in [57]. However, the general solution to a system of coupled (thus synchronized) nonlinear differential equations is not usually obtainable except for some particular forms [57]. The main problem is that actual systems often do not have these solvable forms. Another problem is that much information is missing, leading to an underdetermined system.

Eq. (14) can be expressed in more explicit form. We can assume a first order time dependency for protein expression although it is not clear that this would hold adequately for all regulatory mechanisms. The form of a set of e_j might be modeled with the following time-dependent first order NLDE where E is a column vector of e_j, $C_0, K_0(t)$ are constant vectors, C_n and $K_n(t)$ are constant matrix terms for equations on the order of n, $n > 0$. The C terms could model the coupling between e_j (the linkage of expression between various proteins), while $K(t)$ terms model the environment (for a relatively constant environment). E^T is the matrix transpose of E.

$$\frac{\partial E}{\partial t} = C_0 + C_1 \cdot E + E \cdot C_2 \cdot E^T + \ldots + K_0(t) + K_1(t) \cdot E + E \cdot K_2(t) \cdot E^T + \ldots \quad (15)$$

Of course most of $E(t)$ will be 0 (*i. e.* that part of the genome not being expressed) and of those proteins remaining, one might hypothesize on the basis of observing 2-D gels for some model systems that many e_j are uncoupled from one another so that the C_n and $K_n(t)$ could be sparse. By carefully controlling the cell environment, one might force K_n to be sparse for some $n > 0$ and nearly constant. However, the equations are still underdetermined and may not fit the form of those equations which can be solved exactly – although numerical solutions may be possible if various assumptions are made. However, the solutions of these types of coupled equations may yield the type of multistate biological behavior observed and expected.

We propose using subsets of spots as one basis for generating initial estimates. The form of equations fitting the 2-D gel data (modeling it well) might suggest or rule out possible underlying biological mechanisms as explanations for such a model. In any case, such biological explanations must conform to quantitative data from the 2-D gel analysis. Thus the resulting forms put real constraints upon hypotheses explaining the molecular biology of a network model of genetic expression. One may postulate laws which may constrain this higher order analysis: (i) There are a limited number of permissible (nonlethal) states of differentiation. (ii) There are also a limited number of thermodynamically optimal states. (iii) Conservation of cell resources is necessary for normal activities and this may restrict the number of such states. There is a theoretical upper bound on the number of states S of a eucaryote based on the number of proteins expressed, N, – and the number of proteins which could interact with each other, K. In the human genome, if the number of total proteins in the genome is on the order of 50 000 while N is on the order of 10 000, then S is on the order of

$$C_K^N = \frac{N!}{K!(N-K)!} \quad (16)$$

For some model systems, if K is very small, then the magnitude of S may be manageable by computer. Biological model systems which could most productively use this higher order anolysis are those such as *E. coli.* Another question which might be addressed is whether the expression of genes is in a continuous or discontinuous set of states for a set of asynchronous samples. That is, for all gels in a database containing various experimental classes, does the ratio of Spot A to Spot B change continuously? Is the distribution of the ratio Poisson (1 state) and with what variance, bimodal (2 states), trimodal (3 states), *etc.*? Is this due to simple genetic regulation of possibly to some "chaotic" parametric control of regulation which could produce multiple stable states?

4.4 Cluster analysis as a data reduction tool

Table 8 lists the cluster analysis methods used by some of the gel database analysis systems. Cluster analysis can be used to find sets of gels which are similar in some sense and detection of outlier gels in an experimental class. It can also be used to find sets of marker proteins which can be used to discriminate between two or more classes of gels. We have used it for the separation of marker spots into expression profile groups which are putatively functionally related [59]. Some of the clustering algorithms are able to rank-order these marker proteins by their ability to discriminate gel classes [61]. Cluster analysis is discussed in terms of the numerical taxonomy [62] and principal components analysis as a form of multivariate statistics [63], two subdisciplines with different ways of viewing the data. Conceptual clustering [27, 64] attempts to cluster objects representing a set of features and to discover rules defining relationships between objects in these classes. Instead of the numerical similarity measures representing clusters normally associated with numerical taxonomy methods, the conceptual clustering methodology represents sets of objects by sets of concepts derived from the cluster analysis. Conceptual clustering has been used for gel classification [61]. Ho *et al.* [65], using conceptual clustering, describe a general algorithm for generating rules for expert systems from observations. Other recent advances in the optimization of tree construction from the distance matrices make the particular clustering method used less important in minimizing "linkage" effects [66]. It also reduces the uncertainty of cluster membership of objects midway between two clusters. The two main objectives using cluster analysis seem to be (i) discriminating gel classes based on marker proteins for

Table 8. Comparison of cluster analysis methods: N spots, K spot subset, M gels

Variables	Objects	Purpose	Methods	References
K of N selected marker spots	M gels	Subclassify gels as function of marker spots	Prefilter + *t*-Test + complete-linkage cluster-analysis	*cf.* Fig. 3 GELLAB
M gels for K of N spot expression profiles	K selected spots	Group putatively functionally related proteins	Prefilter + *t*-Test + complete-linkage cluster-analysis	[59] GELLAB
M gels for K of N spot expression profiles	K selected spots	Group putatively functionally related proteins	Principal component analysis	[90] TYCHO/KEPLER
M gels for K of M spot expression profiles	K selected spots	Group putatively functionally related proteins	Heuristic clustering analysis	[61] MELANIE/ELSIE-IV
M gels, N spots expression profiles	N spots as eigenvectors	Group putatively functionally related proteins	Principal component analysis Predicate calculus	[91] [15] HERMeS

clinical applications, and (ii) finding groups of coregulated proteins.

4.5 Options for future 2-D gel databases

Future 2-D gel database systems will need to integrate other information including gene and protein sequence, structure and function databases with the 2-D gel databases [6–8]. Table 5 lists some of this associated information. Several investigators developing 2-D gel analysis systems have suggested using auxiliary biological databases (both nucleic acid and protein sequence as well as others) to help perform more informed experiments at the molecular biology level [1, 16, 40, 67]. A number of detailed specific gel databases have been assembled over different cell lines and species. These include: human lymphocytes [34, 68, 69, 70]; human fibroblasts [71]; HeLa [72]; correlation between mouse and human patterns [73]; human plasma proteins [74]; *E. coli* [75]; and the rat REF52 cell line (normal and transformed clones) [76]. Anderson's group has published extensively on protein maps for a variety of domains [77]. New databases are beginning to appear in a number of different domains.

Anderson *et al.* [1, 67] suggested the Human Protein Index and its extension to other species and dimensions of data as a way of both approaching the complexity of biological systems and of communicating such information between laboratories. They suggest [40] the use of annotated databases (ADB) associated with each gel quantitative databases (QDB) (the type of databases we have been discussing so far) to point into a wide variety of auxiliary database files using relational database formalisms. Using simple hierarchical and relational architectures, relations are hierarchically defined on: properties (of anything in the data base), mappings (of spots between gel databases), people (who work on a database), bibliography, definitions (of terms used in the ADB), synonym (for users to map their jargon to that of the DB), contexts (experimental systems), functions (of the data), and rules (to relate other objects in the ADB). Given these basic relations in the ADB and programs to operate on it and the QDB, one could then use such a database to aid new research. Anderson [40] suggests that given a sufficient ADB plus QDB it would be possible to exchange such a database between laboratories since the ADB includes definitions which could be used to interpret the QDB for other laboratories.

Garrels [13] suggests using a hierarchy of databases consisting of databases for: an individual experiment, other experiments in the system, and/or a species-specific database. He also suggests some of the implications of the network structure of a database based on matchsets [76]. "... First, there is no limit to the number of matchsets that can be entered. Second, there is no need to directly match highly divergent protein patterns; these can be linked through intermediate gels (for example a gel representing a mixture of two samples). Third, any amount of connectivity is allowed. Matchsets from a group of related experiments may be highly connected to each other, yet these may not be connected at all with matchsets for another cell type or another species. However, when additional information (such as protein identity of amino acid composition) allows a match between spots of representative linker gels from the distanct matchsets, a connection is immediately made between two potentially large bodies of data." As we discussed in Section 3, a GELLAB gel database is equivalent to a matchset or match-pair database; therefore, these comments also apply for data in other database forms.

Miller and Olson's ELSIE-4 system [10, 30, 79] is a UNIX-based system using the concepts of match-groups and match-pairs. The philosophy is to construct a set of match-pair files between gels (a match-pair file contains lists of matches between gels – similar to Garrels' matchsets). These in turn are processed with UNIX shell level sequences of program "tools" to produce match-groups of corresponding spots across the gels of interest. Some effort is spent in detecting and then interactively correcting these match-pair files using the powerful tools they have developed. This can result in the construction of a highly reliable database. Editing includes: (i) splitting a spot, (ii) merging two (or more) spots into one, (iii) repairing mis-paired spots. Some of this can be automated but manual review of computer decisions is required. Miller [30] notes that matchpair errors can be multiplicative if a cross-matching is performed on a sequence of gels. If, on the other hand, all gels are matched to a single gel (as is done with the Rgel in GELLAB), these errors do not propagate. However, for the latter method we trade off this advantage for another problem. A gel matched with the Rgel may in fact have a much better match with some other gel with a similar geometry but will not be able to take advantage of this fact.

GELLAB's approach to spot mispairing (although it has minimal spot editing built into the GELLAB-I database program) is to handle mispairings statistically. Whenever possible, replicate gels should be used so that false positive or false negative spots for different gels can be quickly checked using mosaic and Rmap images [19]. Spots found to be mispaired can be corrected in the GELLAB I database using the spot editor on the gel database rather than lower level paired gel files. GELLAB-II will expand this spot editing facility using interactive graphics. Any (*pI, MW*) position in a gel *g* can be mapped to the Rgel as $(pI, MW)_{Rgel}$. Then, the mappings of $(pI, MW)_{Rgel}$ for all Rspot numbers are kept in the gel database. A list of Rspot sets containing all spots in any gel around any given (*pIe, MW*) region can be retrieved and edited. The position could be specified interactively using gel images. The MELANIE system of Hochstrasser *et al.* [11, 12], is based on Miller's ELSIE system and uses heuristic clustering extensions to principal component analysis (PCA) to find characteristic sets of spots which can be used for classifying gels. They suggest capturing expert-system rules inferred from results of PCA using conceptual clustering [61, 80].

The HERMeS system of Vincens, Tarroux *et al.* [15, 16, 81, 82] uses predicate calculus and relational database technology to express and implement "prefilter" constraints. The relational database facility gives the ability to delve into other auxiliary tabular databases. Expressing search conditions using predicate calculus is more powerful than the GELLAB-I approach because of its generality, but seems more awkward and unintuitive – especially for inexperienced or occasional users because of the necessity of specifying the constraints in one of the predicate calculus equivalent languages. They propose creating expert-system rules to aid the data reduction and have used such rule-based systems in parts of their implementation [15, 32]. A learning program would automatically detect regularity in the study data and deduce these rules [16]. Later, Leavitt *et al.* [83] have an interesting approach which uses amino acid compositions measured for a set of gels of the same material with different amino acids radiolabeled to establish the linkage between 2-D and sequence databases. Alternatively, one could cut out spots from gels, microsequence them [6], and then find homologies in sequence databases such as GenBank (Intelligenetics, Los Alamos) EMBL (European Molecular Biology Laboratory, Heidelberg) or NBRF (National Biomedical Research Foundation) to make the linkage as "connector databases" [84].

Hanash, Neel, Kuick, Skolnick *et al.* [20, 50, 70, 85] have concentrated on the spot matching aspects of the database problem. Using their approach, one can match more than two gels at a time. Data is currently obtained by using the bioimage system [20] which summarizes data as a list of spots with values for matched gels. These are subsequently merged into group-match (GM) files. These in turn are used to construct composite-match (CM) files. The CM file is similar to the GELLAB Cgel' estimate of a 'canonical gel' [17]. However, the CM gel is used extensively to improve spot matching by matching two derived CM files to produce a new GM file and this process may be iterated. Our approach with GELLAB-II, which has been partially outlined in this paper, is to layer an object-oriented database with the cgelp2 database program as one of its components. This auxiliary database and its 'data manager' will have access to various types of data objects and 'methods' for handling them. We will discuss object-oriented databases later in this paper.

4.6 Perspectives for future 2-D gel database structures

Object-oriented databases (OODB) are a relatively new way of organizing data. Such a methodology employs more advanced ways of organizing and processing data, by encapsulating data and the methods which know how to process this data as objects. An object definition is a set of named data instances and possible procedures, called methods, for performing operations on that data. Just because a method exists does not mean that it has to be used – it is a capability which is available to operate on that data. No other methods can manipulate that data. In addition, an object can inherit other object definitions as well as modify these inherited definitions. Examples of some object categories are: spot (for an individual gel), Rspot (across a set of gels), sets of spots, gel, sets of gels, gel-databases, sets or subsets of gel databases, *etc.* Table 5 suggests the derived object spot and the subclasses of objects from which it was derived. Thatle [88] describes object-oriented databases for other domains but the basic schema might be used in the 2-D gel domain. Programming languages, such as C++ [89], which embody many of these concepts, are becoming increasingly popular. Other similar approaches are being applied in the HERMeS and MELANIE systems. Object-oriented databases are a way of dealing with the complexity of tree or graph structured lists and of hiding complexity. Associated with different objects, in addition to being used to store data, are procedures or methods which are available to operate on that data. Objects can inherit attributes of other object classes (such as might appear in other types of data bases) – both as single and multiple inheritance (the latter being a many-to-one mapping). OODB overcome the shortcomings of network, hierarchical or relational database in modeling power because as Thatle notes "OODB do not have the restrictions on the necessary diversity and richness of object types and structures" [88]. The methods one defines with the object know how to operate on the objects. The different pieces of data in an object are sometimes called slots; therefore adding another type of data consists of just adding another slot (data structure and methods to process it) to its object definition. Anderson's annotated database concept [40] is major progress in that direction. Some of the types of slots which might be used in such an OODB for 2-D gels are listed in Table 5. Because OODB can handle inheritance, duplication of information found in the existing nucleic acid or protein sequence data bases can be avoided. This has the added advantage that the chance of error is reduced since the information is entered once – by the maintainers of those other databases.

Other issues need to be addressed with these database hierarchies. Ways of discussing data in a global context as well as a local context are required. For example, one might use a primary spot accession index (which is citable) when talking about the master index (such as the protein sequence database) and secondary spot numbers when working a local database. The database needs a data dictionary such as mentioned by Anderson [40] in his ADB, but again this should be divided into two types: the standardized and the local. In some cases, relations can be created to map one definition into the other, while in other cases additional terms can be introduced through the local data dictionary which have not yet been

standardized. In the past few years GenBank and EMBL began an effort to standardize their annotation and feature formats. As 2-D gel data bases will be accessing this data as auxiliary databases, these sequence databases should be tracked to avoid duplication and to maximize their utility. Associated with this development will be increased use of artificial intelligence and expert-system techniques for both spot analysis and experimental design management [12, 15, 16, 19, 32, 65, 80]. Ease of use is important. As Miller [79] points out ". . . programs which are 'user friendly' are made so by putting the user into a procedural straight jacket that allows no flexibility in the way gels are processed. On the other hand, systems that provide a wealth of flexibility, with many ways of processing the data may be a joy to the expert, but frustratingly incomprehensible to the new or computer-naive user". In ELSIE-IV, he gains the flexibility for experimenting with a new analysis algorithm by prototyping it using the UNIX AWK language to process data obtained by his "listgroups" program. Successful prototype algorithms are then recoded more efficiently and added to a library of such functions. Such packages of sequences of such functions could be repackaged to be more transparent and thus more user friendly.

It has been our experience that user friendliness need not be constraining. (i) User-friendly systems can be made powerful enough to do almost everything required for a particular application domain. (ii) Escape mechanisms can be built into such systems which will also satisfy the expert user. Use of high level tools such as command macros or scripts (generated by 4th generation type tools), or expert-system interfaces can facilitate access to new developments in the current context of the system for inexperienced users. GELLAB allows experimentation with new algorithms on gel database prefiltered data by generating derived data sets which are SPSS (statistics package for the social sciences) compatible but which can be (and are) used for a variety of analyses by GELLAB. This is similar to Miller's use of "listgroups". Our philosophy has also been to test new algorithms outside of the gel database program using these SPSS files and then incorporate the algorithm into the database program if found to be useful. In addition, GELLAB-II will be able to transparently run other programs on such derived data as if these programs were incorporated into the main gel database program. Cgelp2 will also be able to act as a "database server" for other programs on a computer network.

In summary, there are common themes seen in all the 2-D gel analysis systems reviewed here. They all handle and attempt to automate to various degrees the processes of data acquisition, spot-quantification, spot-pairing, composite database generation and its subsequent manipulation. Composite database analyses perform statistical and logical searches through this data and record the results which are then available for further analysis. Flexible database offer the promise of being able to apply additional constraints from information supplied by external databases containing other types of biological information. A 2-D gel analysis system embodying this paradigm should be easy to use both in the sense of minimum training required and offering on-line expert-system assistance to help plan exploratory data analysis strategies. The system should be usable for other types of analysis besides 2-D gels, ideally including mathematical modelling of other biological activities as well as one-dimensional gel data. A system should be extensible so that new gel analysis algorithms may be easily added, allowing the researcher to try out new analytic procedures without any drastic modifications to the system. Databases should be self-organizing using OODB concepts and be in formats that are portable to different gel analysis systems across laboratories. No restrictions should be placed on how much of the OODB needs to be previously defined. Rather, it should allow the individual user to fill out objects as data is required or becomes available – possibly with data from other laboratories as the reproducibility problem is better understood.

We would like to thank the referees for their many useful suggestions in improving the style of this paper.

Received August 7, 1988

5 References

[1] Anderson, L. and Anderson, N., *Clin. Chem.* 1984, *30*, 1898–1905.
[2] Taylor, J., Anderson, N. L. and Anderson, N. G., *Electrophoresis* 1983, *4*, 338–346.
[3] Young, D. A., *Clin. Chem.* 1984, *30*, 2104–2108.
[4] Klose, J. and Zeindl, E., *Clin. Chem.* 1984, *30*, 2014–2021.
[5] Hochstrasser, D. F., Harrington, M. G., Hochstrasser, A. C., Miller, M. J. and Merril, C. R., *Anal. Biochem.* 1988, *173*, 424–435.
[6] Hood, L. and Smith, L., *Issues in Science and Technology*, Spring 1987, 36–47.
[7] Delisi, C., *Science*, 1988, *240*, 47–52.
[8] Pabo, C. O., *Nature*, 1987, *327*, 467.
[9] Taylor, J., Anderson, N. L., Scandora, A. E., Willard, K. E. and Anderson, N., *Clin. Chem.* 1982, *28*, 861–866.
[10] Olson, A. D. and Miller, M. J., *Anal. Biochem.* 1988, *169*, 49–70.
[11] Funk, M., Thesis #2236, Department D'Informatique, University of Geneva, 1987.
[12] Appel, R. D., Thesis #2241, Department D'Informatique, University of Geneva, 1987.
[13] Garrels, J. I., Farrar, J. T. and Burwell IV, C. B., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*. Academic Press, New York 1984, pp. 37–91.
[14] Blose, S. H., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, VCH Publishers, Deerfield Beach, FL 1986, pp. 552–555.
[15] Vincens, P. and Tarroux, P., *Electrophoresis* 1987, *8*, 173–186.
[16] Tarroux, P., Vincens, P. and Rabbiloud, T., *Electrophoresis* 1987, *8*, 187–199.
[17] Lemkin, P. F., Lipkin, L. E. and Lester, E. P., *Clin. Chem.* 1982, *28*, 840–849.
[18] Lemkin, P. F. and Lipkin, L. E., in: Geisow, M. and Barrett, A. (Eds.), *Computing in Biological Science*, Elsevier/North Holland, Amsterdam 1983, pp. 181–226.
[19] Lemkin, P. F. and Lipkin, L. E., *Electrophoresis*, 1983, *4*, 71–81.
[20] Kuick, R., Sing, C. and Hanash, C., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press of Nancy, Nancy 1986, pp. 91–96.
[21] Skolnick, M. M., *Comp. Vis. Graph. Img. Proces.* 1986, *35*, 306–332.
[22] Skolnick, M. M. and Neel, J. V., *Adv. Hum. Genetics*. 1986, *15*, 55–160.
[23] Ridder, G., Von Bargen, E., Burgard, D., Pickrun, H. and Williams, E., *Clin. Chem.* 1984, *30*, 1919–1924.
[24] Smith, K. A. and Dunn, M. J., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, VCH Publishers, Deerfield Beach, FL 1986, pp. 560–562.
[25] Mann, R. C., Mansfield, B. K. and Selkirk, J. K., in: Gelsema, E. S. and Kanal, L. N. (Eds.), *Pattern Recognition in Practice II*, Elsevier Science Publishers, North-Holland, Amsterdam 1986, pp. 301–312.
[26] Dunn, M. J. and Burghes, M. H. M., *Electrophoresis* 1983 *4*, 173–189.
[27] Fisher, D. and Langley, P., in: Glae, W. A. (Ed.) *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, MA 1986, pp. 77–116.
[28] Lipkin, L. E. and Lemkin, P. F., *Clin. Chem.* 1980, *26*, 1403–1413.
[29] Miller, M. J., Olson, A. D. and Thorgeirsson, S. S., *Electrophoresis* 1984, *5*, 297–303.

[30] Miller, M. J., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 322–335.
[31] Vincens, P., *Electrophoresis* 1986, *7*, 357–367.
[32] Vincens, P. and Tarroux, P., *Electrophoresis* 1987, *8*, 100–107.
[33] Mann, R. C., Mansfield, B. K. and Selkirk, J. K., *Cancer Res.* 1988, *48*, 1110–1118.
[34] Lester, E. P., Lemkin, P. F. and Lipkin, L. E., *NY. Acad. Sci.* 1984, *428*, 158–172.
[35] Lemkin, P., Sonderegger, P. and Lipkin, L., *Clin. Chem.* 1984, *30*, 1965–1971.
[36] Giometti, C. and Zhang, J.-S., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, VCH Publ., Deerfield Beach, FL 1986, 670–673.
[37] Goldman, D., Giri, P. R. and O'Brian, S. J., *Proc. Natl. Acad. Sci.* USA, 1987, *84*, 3307–3311.
[38] Steinberg, R. A., O'Farrel, P. H., Freidrich, U. and Coffino, P., *Cell* 1977, *10*, 381–391.
[39] Stoeckli, E. T., Lemkin, P. F., Kuhn, T. B., Ruegg, M. A., Heller, M. and Sonderegger, P., submitted, 1988.
[40] Anderson, N. L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313–321.
[41] Goldman, D. and Merril, C. R., *J. Psychiat. Res.* 1987, *21*, 597–608.
[42] Reiser, J. F., SAIL, 1976, # AD-A045-102 from NTIS, Springfield, VA.
[43] Lemkin, P., *Comp. Languages* 1985, *2*, 39–45.
[44] Lemkin, P., *SIGPLAN. Notices* 1988, *23*, 149–171.
[45] Scheifler, R. W. and Gettys, J., *ACM. Trans. Graph.* 1986, *5* (2).
[46] Lemkin, P., *Manual* 1988 (March), pp. 1–16.
[47] Johansson, K.-E., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press, Nancy 1986, pp. 7–40.
[48] Hayes-Roth, F., Waterman, D. A. and Lenat, D. B., (Eds.) *Building Expert Systems*, Addison-Wesley, Reading, MA 1983.
[49] Tukey, J., in: Gale, W. A., (Ed.) *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, MA 1986, pp. 401–409.
[50] Kuick, R. D., Hanash, S. M., Chu, E. H. Y. and Strahler, J. R., *Electrophoresis* 1987, *8*, 199–204.
[51] Lewin, B., *Genes III*, John Wiley, New York 1987.
[52] Gold, D. W. and Gasson, J. C., *Sci. Amer.* 1988, *259*, 62–70.
[53] Guarente, L., *Cell* 1988, *52*, 303–305.
[54] Gleick, J., *Chaos Making a New Science*, Viking Press, New York 1987.
[55] Busse, H. G., *Zeitschr. Naturforsch.* 1979, *34a*, 1518–1527.
[56] Pincaré, *Acta. Math.* 1885, *7*.
[57] Minorsky, N., *Nonlinear Oscillations*, Van Nostrand, New York 1962.
[58] Lester, E. P., Lemkin, P. F., Lowery, J. F. and Lipkin, L. E., *Electrophoresis* 1982, *3*, 364–375.
[59] Sonderegger, P., Lemkin, P. F., Lipkin, L. E. and Nelson, P. G., *Dev. Biol.* 1986, *118*, 222–232.
[60] Lester, E. P., Lemkin, P. F., Lipkin, L. E., in Rowley, J. D. and Ultmann, J. E., *Chromosomes and Cancer: From Molecules to Man*, Academic Press, New York 1983, pp. 226–245.
[61] Appel, R., Hochstrasser, D., Roch, C., Funk, M. and Muller, A. F., *Electrophoresis* 1988, *9*, 136–142.
[62] Sneath, P. H. A. and Sokal, R. R., *Numerical Taxonomy*, Freeman Co, San Francisco 1973.
[63] Srivastava, M. S. and Carter, E. M., *An Introduction to Applied Multivariate Statistics*, North-Holland, New York 1983.
[64] Michalski, R. S. and Stepp, R. E., *Artific. Intell.* 1986, *28*, 203–226.
[65] Ho, T. B., Diday, E. and Gettler-Summa, M., *Pat. Rec. Let.* 1988, *7*, 265–271.
[66] Henaut, A. and Delorme, M. O., *Pat. Rec. Let.* 1988, *7*, 207–213.
[67] Anderson, N. G. and Anderson, L., *Clin. Chem.* 1982, *28*, 739–748.
[68] Willard-Gallo, K. E., *NY Acad. Sci.* 1984, *428*, 201–222.
[69] Willard-Gallo, K. E., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press, Nancy 1986, pp. 205–214.
[70] Hanash, S. M., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press, Nancy 1986, pp. 197–203.
[71] Celis, J. E., Bravo, R., Larson, P. M., Fey, S. J., Bellatin, J. and Celis, A., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 307–362.
[72] Bravo, R. and Celis, J. E., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 445–476.
[73] Fey, S. J., Bravo, R., Larsen, P. M. and Celis, J. E., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 169–189.
[74] Anderson, N. L. Tracy, R. P. and Anderson, N. G., in: Putnam, F. (Ed.), *The Plasma Proteins*, Vol. 4, Academic Press, New York 1984, pp. 221–270.
[75] Neidhardt, F. C., Vaughn, V. and Phillips, T. A., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 417–444.
[76] Garrels, J. I. and Franza, Jr., R. B., *J. Biol. Chem.* 1989, in press.
[77] *Publications of Protein Mapping* Large Scale Biology Corporation, LSB, Rockville, MD July 1987.
[78] Garrels, J. C. and Franza, B. R., Jr., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press, Nancy 1986, pp. 85–90.
[79] Miller, M. J., *Exper. Bio. Med.* 1986, *19*, 235–260.
[80] Funk, M., Appel, R. D., Roch, C., Hochstrasser, D., Pellegrini, C. and Muller, A. F., in: Fox, J., Fieschi, M. and Engelbrech, R. (Eds.), *Lecture Notes in Medical Information: Proc. AIME 87*, Marseille, 1987, pp. 97–103.
[81] Vincens, P. and Rabilloud, T., in: Galteau, M. M. and Siest, G. (Eds.), *Recent Progress in Two-Dimensional Electrophoresis*, University Press, Nancy 1986, pp. 121–130.
[82] Vincens, P., Paris, N., Pujol, J.-L., Gaboriaud, C., Rabilloud, T., Pennetier, J.-L., Matherat, P. and Tarroux, P., *Electrophoresis* 1986, *7*, 357–367.
[83] Latter, G. I., Burbeck, S., Fleming, J. and Leavitt, J., *Electrophoresis* 1984, *30*, 1925–1933.
[84] Anderson, N. L. and Garrels, J., *Electrophoresis* 1986, *7*, 295–296.
[85] Neel, J. V., Rosenblum, B. B., Sing, C. F., Skolnick, M. M., Hanash, S. M. and Sternberg, S., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 259–306.
[86] Goldman, D. and Merril, C. R., in: Celis, J. E. and Bravo, R. (Eds.), *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 241–258.
[87] Special issue on "Model Systems", *Science* 1988, *240*, 1377–1580.
[88] Thatle, S. M., in: *Proceedings of Software Development '88*, Miller Freeman Publ., San Francisco 1988, pp. 381–388.
[89] Strustrup, B., *The C++ Programming Language*, Addison-Wesley, Reading, MA 1986.
[90] Anderson, N. L., Hoffman, J. P., Gemmell, A. and Taylor, J., *Clin. Chem.* 1984, *30*, 2031–2036.
[91] Rabilloud, T., Vincens, P. and Tarroux, P., *FEBS* 1985, *189*, 171–178.

6 Appendix: GELLAB terminology

The following five categories of terms used in GELLAB are briefly defined.

6.1 Gels

Gels are the basic experimental object of discourse. We may discuss them with respect to a particular gel sample or abstractly with respect to a database (DB) idealization of a set of gels or gel domain.

Rgel – is a 'representative' gel in the database under consideration where most spots appear to be found. It is the particular gel to which all other accessioned gels geometry is mapped in referring to corresponding spots.

Cgel – is the 'ideal' or 'canonical' gel which does not exist in an actual gel database. If it did, it would contain all spots seen in all gels.

Cgel' – is an estimate of the Cgel and may be computed for a subset of gels using a subset of the composite gel (PCG) DB (defined below). Each class of gels in the database could be represented by its computed Cgel'. This allows for extensive data compression in the database but retains individual spot variances for subsequent calculations.
Working Set of gels – is a subset of all gels in the composite gel (PCG) DB which are currently visible. The user can redefine which gels are in the working set at any time.
Experimental class – of gels is a subset of gels in the working set of gels which are all of the same experimental class, *e.g.* lymphoid *vs.* myeloid leukemia gels. The same gels may simultaneously belong to other classes, *e.g.* acute *vs.* chronic leukemia.
Gel subset – is any defined subset of gels in the database. A gel subset is defined by one of several criteria: gels in the same experimental class, gels in the working set, explicit definition, set operations on other sets of gels, *etc.* Although there may be gels in a subset which are not in the working set of gels, only those which are in *both* sets are visible during an analysis.

6.2 Spots and groups (sets) of spots

Spots are those polypeptides detected in the gels that we are trying to analyze. As with gels, there are several different levels of discourse.
Spot – generally refers to a particular polypeptide spot in a particular gel.
Rspot number – is a number which uniquely defines corresponding spots in different gels of a particular composite gel (PCG) DB. This Rspot number is an arbitrary sequential number which is assigned in the database. Corresponding spots in any gel can be retrieved by knowing the Rspot number and the gel accession number.
Rspot set of spots – a set of corresponding spots in different gels which correspond to the same Rspot number. There is one spot from each gel which corresponds to the same Rspot number.
Landmark spot – is a specific easily recognized spot interactively defined to be the same for all gels. A set of *Landmark spots* are all those spots (typically 10 to 20) determined for each and every gel in an experiment.
Extended Rspot (eRspot) – is an Rspot for any spot found in gels other than the Rgel and not found in the Rgel. Therefore the set of all Rspots (extended and unextended) includes all spots found on any gel.
Extrapolated paired spot (EP) – is a synthetic spot extrapolated to any gel (including the Rgel) when the corresponding spot is missing from that gel. It is assigned zero integrated density.
Composite pair spot (CP) – is a synthetic spot formed by considering several adjacent spots as part of the same spot group and treating them as one spot for statistical purposes.
Search results list (SRL) – is a list of Rspot numbers (*e.g.* those found to be significant after an analysis of the database). The list can be edited or created explicitly or may be the result of an automated statistical analysis of the database.
SRL subset – (SRL[n]) is an SRL saved as a separate subset *n* or Rspot numbers with its own user-assigned annotation title. The spot subset may be derived from a variety of database search and/or post-search operations. The SRL subset may be referred to by either its assigned set number (*e.g.* SRL[5]) or by its title. The SRL subset may also be retrieved by specifying relational expressions of key words found in the title – or by querying which SRL subsets contain which spots.

6.3 Images

Raw gel image – is the original computer readable scanned image of a gel which can be calibrated in optical density units or counts per minute.
Mapped gel image – is a transformed gel image such that the geometry of the initial gel image is mapped to the geometry of another so that they could then be superimposed to visually compare corresponding spots in the two gels.
Rmap image – is a synthetic image composed of a particular gel image with those Rspot numbers of interest overlayed (*cf.* Fig. 2). It is useful for getting a global view of spots of interest and at optionally different magnifications. It can be displayed on a video display or a line-drawing plot.
Mosaic image – is a synthetic image composed of similar regions or 'panels' from different gel images which surround a particular spot. These panels are sorted by mininum protein concentration in raster order (top to bottom, left to right). It is useful for validating marker spots identified by statistical analysis. It can be displayed on a video display or a line-drawing plot.

6.4 Database files

Database files are the raw material used by the data analysis programs at various stages of processing and data reduction.
Gel accession file – contains the basic auxiliary sample identification information for all gel experiments. Each gel is indexed by its unique accession number of the form XXXX.E.
Gel image files – are raw scanned gel images created at the time the gels are digitized into the system.
Gel spot file (GSF) – is a list of spots, with an identifying connected-component number for each spot, and their features, *e.g.* centroid (*pI*, *MW*), concentration (integrated optical density or counts/minute), area, *etc.* segmented or extracted from a single gel image. The data from GSF files is used to construct the gel comparison files.
Gel comparison file (GCF) – is a list of spots paired between the Rgel and another gel using GSF data and a set of landmark spots common to both gels. The GCF corresponds to matchset or match-pair files generated by algorithms used by other systems. Some of these other systems use just a few landmarks or guess the initial landmarks and 'grow' them through the gel – thus requiring no initial manual landmark definitions. The data from GCF files is used to construct the composite PCG DB file.
Paged composite gel database file (PCG DB) – is a random access composite 3-D gel database built from all of the GCF data. Subsequently, extrapolated spots missing from any gel but present in some gel may be added for the gels where they are missing. A PCG DB may be thought of as a complete gel database resulting from all the gels in a given experiment. The term paged is a computer science term meaning that only part of the information in a file resides in the high speed computer memory at any one time – that part which is currently being analyzed. Paging algorithms are commonly used to automatically and efficiently bring data required in (out) of local memory from (to) a random access PCG DB disk file.
Spot search results list file – is a list of SRLs, each of which consists of a title and list of Rspot numbers identified in the PCG DB analysis.
Composite gel file (CGL) – all of the data in the PCG DB as currently viewed by the cgelp2 DB program.
SPSS data file – is a list of Rspot set data for selected spots in the PCG database. Alternatively, it can consist of all protein

concentration data for all working set gels by all spots visible to the prefilter with missing spots defaulting to 0. This numeric data file is suitable for input to commercially available statistical analysis programs such as SPSS or SAS as well as other data analysis programs.
Inquire file – lists the results of performing a search or SRL post processing operations.
Table file – lists the results of performing statistical operations resulting in various types of global summary statistics tables (*e.g.* correlation, expression-profile tables, *etc.*).

6.5 Data filters

Data filters are programs or parts of programs which 'filter out' uninteresting data prior to performing a requested operation. Data needs to be filtered for a variety of reasons and at different levels of the data analysis.
Prefilter – is a set of constraints on spot properties which must be present for a spot to be considered for further processing. These limits test for features of individual spots and the Rspot set as a whole to determine its applicability for further tests.
Statistical test – is a subsequent Rspot set filter applied to all successfully prefiltered Rspot sets to produce an SRL. These univariate tests include parametric tests: such as a *t*-test (both standard and Behrens-Fisher based on F-statistic for equal/unequal variance, *F*-test, *etc.*; and non-parametric tests such as the Wilcoxon-Mann-Whitney, Kruska-Wallis, *etc.* [18]). The data in each Rspot set is tested for significant differences using one of the these tests comparing two or more classes of gels (*e.g.* AML *vs.* CLL).
SRL subset operations – (Union, Intersection, Difference) is considered a postfilter operation performed using the results of previous statistical tests or SRL subset operations resulting in additional SRL subsets (defined above). It is used to compare the members of various SRLs.
Clustering operations – on subsets of gels and/or SRL subsets can result in the definition of additional relationships between gels or Rspot sets. For example, spots can be clustered into coordinately regulated spots as a function of protein expression profiles; gels can be clustered *into groups which can be* predicted by marker proteins.

J. Klose

Institut für Humangenetik, Institut für Toxikologie und Embryonal-Pharmakologie, Freie Universität Berlin

Systematic analysis of the total proteins of a mammalian organism: Principles, problems and implications for sequencing the human genome

High-resolution two-dimensional electrophoresis (2-DE) has reached a technological level that allows us to resolve most of the numerous unknown protein species of a mammalian organism if appropriate strategies are used. We will discuss the problems of classification and characterization of proteins and propose a systematic approach to the analysis of the total protein complex. Both a comprehensive as well as a pragmatic approach towards systematic analysis have been considered. A "complex protein database" is suggested and considered with regard to various uses. A systematic analysis of the mouse proteins has been started and some of the preliminary results are summarized here. In particular, genetic properties of the proteins were investigated and are presented in order to demonstrate the significance of a systematic analysis of proteins for research and practical application (*e. g.* mutagenicity testing). A concept is presented for sequencing the coding DNA of mouse and man, starting with a systematic analysis of mouse proteins and then using two recently developed methods – microsequencing of proteins from spots of 2-DE protein patterns, and utilization of the relatively short *N*-terminal sequences obtained – to produce the corresponding cDNA's of these proteins.

1 Introduction

1.1 Early attempts in constructing protein databases

Since the development of high resolution two-dimensional electrophoresis (2-DE) of proteins in 1975 [1, 2], new techniques of fundamental significance for this method have been introduced, *e. g.* fluorography, silver staining, evaluation of the protein patterns by densitometry and computer analysis, and isoelectric focusing with Immobilines (for reviews see: [3–6]). Taking into account all the other improvements made in sample preparation, gel technique, buffer systems, running conditions, in the conventional staining procedures and the development of electrophoresis equipment, 2-DE has reached a technological level that makes it possible, using suitable strategies, to resolve almost all the different protein species present in a single cell type. In the early 1980's, therefore, a vivid discussion developed concerning a "human protein in-

Correspondence: Prof. Dr. Dr. J. Klose, Institut für Toxikologie und Embryonal-Pharmakologie, Freie Universität Berlin, Garystrasse 5, D-1000 Berlin 33, Federal Republic of Germany

Abbreviations: **2-DE**, two-dimensional electrophoresis; **IEF**, isoelectric focusing; **SDS**, sodium dodecyl sulfate

0173-0835/89/0202-0140 $02.50/0