[Reprinted from CLINICAL CHEMISTRY, 31, 2031 (1984).]

Global Approaches to Quantitative Analysis of Gene-Expression Patterns Observed by Use of Two-Dimensional Gel Electrophoresis

N. Leigh Anderson,[1] Jean-Paul Hofmann,[2,3] Anne Gemmell,[1] and John Taylor[1]

A major difficulty in the use of two-dimensional protein maps to identify and classify cell types is the problem of acquiring, selecting, and analyzing quantitative data on hundreds of protein spots. Here we use methods of multivariate statistics to analyze the differences among a panel of human cell lines, in some cases involving quantitative data on more than 250 proteins. Principal-component and cluster-analysis techniques show that the lines can be easily distinguished, even by using the subset of proteins present in all cells. A preliminary analysis of the protein changes brought about by phorbol ester-induced differentiation of the line U937 is included.

Additional Keyphrases: *computerized pattern interpretation · "marker sets" of proteins · multivariate statistics · principal-component analysis · cluster analysis · numerical taxonomy*

Two-dimensional (2-D) electrophoresis of proteins has been used primarily to detect small differences in protein composition between pairs of related samples. It is useful in this context because sequential isoelectric focusing (with the sample in a medium containing NP-40 detergent and urea) and pore gradient electrophoresis of samples treated with sodium dodecyl sulfate (SDS) can resolve thousands of proteins (*1*), and thus one can detect changes in the abundance, modification, or rate of synthesis of proteins never before observed. In most comparisons of two samples, visual inspection alone allows detection of the alteration or disappearance of a single protein in a field of 2000. This accounts for the widespread usefulness of the technique without computerized data reduction.

In at least three situations, however, visual analysis has proved inadequate: (*a*) inter-comparison of a large series of gels, to evaluate a panel of differences (where the observer's memory and notation methods are the major limitations); (*b*) interpretation of complex qualitative differences (such as the multiple charge changes seen in comparison of species); and (*c*) comparison of patterns in search of complex quantitative pattern differences (as in comparisons of gene-expression patterns in cell types from one species).

The first situation can be dealt with by using a computer system capable of quantifying and stretching gels into registration, combined with a simple database capability for remembering facts about spots (*2*).

The second case has been attacked by using visual interpretation (*3–5*), often with discouraging results (*3, 5*). We will describe more-promising approaches to this problem in a separate paper.

The third situation requires multivariate statistical analysis of computer-generated quantitative data from two-dimensional gels, an initial exploration of which forms the subject of this paper. In this mode 2-D analysis can make perhaps a unique contribution to biological understanding. Alternative methods can usually be devised for the discovery or measurement of one or a few proteins, but no other technique provides as good a means of characterizing comprehensively the major activities of the cell. A multivariate analysis of 2-D mapping data may thus provide the only workable approach for unravelling the complex program of gene expression.

The use of 2-D electrophoretic data for multivariate analysis has only recently been explored. Tarroux (*6*) and Westerbrink et al. (*7*) have successfully used statistical techniques to analyze manually encoded data from 2-D gels of different cell types from the same organism. By calculating a measure of distance (or, alternatively, the similarity) between gels, they were able to use cluster analysis and multi-dimensional scaling to examine the relative relationships between samples. The manual collection of data for such an investigation involves several limitations, however: (*a*) the difficulty of obtaining large data sets routinely (the human factor), (*b*) the lack of any mathematical assurance that the spots measured are the correct ones (the eye can deduce some incorrect associations), and (*c*) the lack of precise quantitation. Clearly, it would be better to be able to apply statistical methods to larger data sets generated and selected by the computer in an unbiased way.

At the outset, it is useful to outline what the goals of a fully computerized statistical analysis of 2-D pattern data would be. Certainly it is possible to differentiate between the 2-D patterns of many cell types on the basis of groups of "marker" proteins discovered during the course of exploratory comparisons. A trained observer can glean important information about a cell type by examining a small set of familiar protein spots, for example cytokeratins (often characteristic of epithelial cells; ref. *8*) or tropomyosin polypeptides (*9*). The difficulty with this approach is that expression of other proteins among the vast majority not currently regarded as specific markers may vary in ways not correlated with a recognized marker set. This would be the case if, for example, there were two subtypes of epithelial cells that shared the expression of certain cytokeratins. Classification of expression patterns based on limited sets of markers is thus likely to be incomplete, and if the chosen "markers" happen not to represent fundamental functional characteristics of the cell, such a classification may be misleading as well.

A very similar problem was perceived in the 1950's with regard to the taxonomic classification of organisms on the basis of quantitative data. This led to the introduction of a group of statistical methods collectively termed "numerical taxonomy" (*10*), a basic tenet of which is that a large set of equally-weighted quantitative characters (or markers) is likely to give a more objective picture of an organism than will a few special characters chosen by a taxonomist. A proper statistical analysis of these (it is hoped) unbiased data is then likely to yield a more general classification. This notion, and its implied challenge to the classification

[1] Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory, Argonne, IL 60439.
[2] Institut National de la Recherche Agronomique, Orsay, France.
[3] Sungene Technologies, Palo Alto, CA.
Received July 17, 1984; accepted August 13, 1984.

methods devised by orthodox sytematists, has been debated ever since. The question of who chooses the characters used for classification, either qualitative (e.g., eye color or absence of feathers) or quantitative (e.g., the length of a particular bone or the rate of utilization of glucose), exposes a remnant of human intervention in the procedure that leaves the result more or less arguable.

There are, nevertheless, several reasons for believing that the methods of numerical taxonomy might provide an effective approach for analyzing quantitative 2-D gel data and, conversely, that such data may provide ideal material for numerical taxonomic analysis:

- We may assume with some confidence that each protein represented by an observed spot has some function in the cell, and hence that a measurement of its abundance is a measurement of something intrinsically relevant to the cell. There seems thus to be inherent in the data some protection against choosing irrelevant characters.
- A great deal of data can be rather easily obtained, making it possible to avoid reliance on a small number of characters. This practical aspect (the facility of data generation) allows us to avoid the pitfalls inherent in classifying complex objects on the basis of too few characters.
- The type of data involved is available almost uniformly for all living things. Because all life consists of cells, and the majority of working parts of all cells are proteins, it follows that comparisons of proteins between and within organisms can (in theory) be universally applied. The consistency of this approach avoids some of the problems inherent in present species taxonomy, which classifies birds on the basis of plumage, beak shape, etc., and mammals or sponges on the basis of entirely different criteria.
- On account of our almost complete ignorance as to the function or importance of individual cellular proteins, the assumption of equal statistical weight for all characters (generally required in numerical taxonomy) seems well justified. In fact some ambiguity remains, because it is not clear whether the characters should be assumed to be the proteins or the amino acids from which they are made. If the latter is the case, we should weight the information in each protein by the size of that protein (which is proportional to the mutational target size of the corresponding gene). We currently favor equal weights for proteins in comparisons within a species, and weighting proportional to protein size when comparing different species. Someday scientists may dispute whether the importance of (e.g.) F1-ATPase is three or five times that of tubulin in fibroblasts, but that will first require generation of a classification of spots based on specific properties related to their differential expression. From a logical point of view, we conclude that there is a useful match of aims and capabilities between 2-D gel data and numerical taxonomic approaches.

These various arguments apply not only to species taxonomy, but to the classification of cells according to stages of differentiation and to the exploration of differences between normal and pathological cells.

Two principal types of questions can be asked in the sort of analysis we are developing. First, what do differences in abundance (or rate of synthesis) between cell types tell us about individual proteins? What proportion of proteins appear to be qualitative markers, what proportion are expressed similarly in a variety of cell types (i.e., might constitute a "base set"), and what proportion vary perceptibly but not radically? Can we determine whether each "regulable" protein varies independently, or whether there are co-regulated sets which always vary together? Second, what can be learned about the relationships between cell types? Do the various bone-marrow-derived cell types resemble each other more than they resemble a fibroblast, for example? In essence, can we construct from the gene expression patterns alone an "ontogenetic tree" that correctly describes the lineage and/or functional relationships among cell types? If such a tree (and an associated classification method) can be generated, then "new" cell types such as cancer cells might be classified more meaningfully, and differentiation could be better understood. From a practical viewpoint, we also could ask whether the various cultured cell types widely used as model systems really represent in vivo cell types.

In this paper we discuss methods for generating statistical data sets in which the abundances of numerous proteins are measured across many samples (gels). Using actual 2-D gel data, we have applied the techniques of principal component and cluster analysis to the problem of determining relationships between a series of human cell types, and have made some progress in demonstrating the potential of this approach.

Materials and Methods

Preparation of Samples

Cells were labeled for 18 h in methionine-free RPMI 1640 medium (GIBCO Laboratories, Grand Island, NY 14672) containing, per liter, 100 mL of fetal bovine serum, and 60 mCi of [^{35}S]methionine. Samples were prepared by solubilizing cells directly in a solution containing, per liter, 9 mol of urea, 20 mL of NP-40 detergent (Particle Data Inc., Elmhurst, IL 60126), 10 mL of mercaptoethanol, and 20 mL of ampholytes (Ampholines; LKB Instruments, Gaithersburg, MD 20877) pH 9–11. Adherent cells were solubilized from the bottom of tissue-culture wells where they had grown; cells in suspension were pelleted by a 2-s centrifugation (Beckman Microfuge; Beckman Instruments, Fullerton, CA 92634) in capillary-bottom Microfuge tubes (Walter Sarstedt, Princeton, NJ 08540) and solubilized immediately by aspirating into and expelling from a Hamilton (Reno, NV 89510) syringe. Most of the lines we used were obtained from the American Type Culture Collection. Monocytes were obtained by density gradient and adherence fractionation of fresh blood (*11*).

Two-Dimensional Electrophoresis

For this we used either the 18 × 18 cm or 20 × 25.5 cm ISO-DALT system (*12, 13*). In the former system, sodium dodecyl sulfate (SDS) acrylamide gradient (90 to 180 g of total acrylamide per liter) slab gels were used in the second dimension; in the latter (for this study) we used 100 g/L acrylamide slabs produced by a microprocessor-controlled device. All samples were processed in groups of 20 or 40. Fixed, stained, and destained gels were autoradiographed on either XAR or GTA films. Unless otherwise noted, we used LKB ampholytes (100 mL of pH 2.5–4 and 900 mL of pH 3.4–10 per liter).

Image Processing, Spot Quantitation, and Gel Matching

Two-dimensional autoradiograms were scanned with an Optronics P-1000 microdensitometer at 100-μm resolution, and analyzed with the Argonne TYCHO II system (see ref. 2). Briefly, this system (based on a VAX 11/780 computer, Digital Equip. Co., Maynard, MA 01754) applies film corrections, removes background, detects spots, and fits them with two-dimensional gaussian forms. The result is a list of spots, each characterized by position, amplitude, and shape (x- and y-halfwidths). For matching patterns to one another we used the TYCHO I display system, and constructed a master

pattern containing all the major spots on any gel. Finally, we stretched the whole set of patterns corresponding to a single electrophoresis run into superposition with the master pattern.

Selection of Data

We sought to apply statistical procedures only to data that could be objectively defined as "good." This is necessary for two reasons: (*a*) some regions of 2-D gels are crowded with spots at these resolutions, and some large spots can be fitted by different numbers or shapes of spots on different gels. The possibility thus exists for mismatching some spots. (*b*) The stretching process cannot always remove all differential distortions between a pair of gels. In a small distorted region, a few spots may not be matched at all, or may be matched incorrectly. Failure to match some spots makes it difficult to know whether a particular protein is "missing" (i.e., undetectable) on a gel or instead is present but unmatched. To minimize these potential difficulties we applied a selection procedure based on a computation of the likelihood that a given protein is correctly matched. Once the gels are stretched into registration, the "overlap" of each spot on the "master" gel with every other spot on the "object" gel is computed (Figure 1). In this case we used as the overlap measure the value of a gaussian function of user-assigned width (typically 0.7) and height 1.0 evaluated at d, which is the physical distance between the spot centers.

Previous work (*17*) has shown that, when gels giving the resolution of these are used, approximately 95% of spots will be placed within 0.7 mm of the expected positions when the pattern is stretched onto a master pattern by the methods described (*18*). For each master spot m, data from a corresponding object spot o are included if the overlap of m with o exceeds a certain threshold (usually 0.5), and if the m–o overlap is more than 90% of the total overlap of m with all object spots or of o with all master spots. These conditions assure a reliable assignment. If this condition is not fulfilled, a "missing data" flag is inserted for the value of the protein m in the current object. The one exception occurs when the total overlap of m with all object spots falls below some threshold (usually 0.001) indicating complete absence of any spot in the area. In this case a "minimum detectable spot" value is used, usually set equal to the smallest spot quantified in the current series of gels.

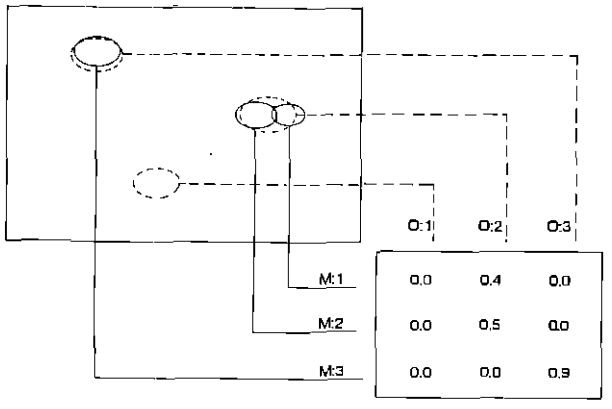

Fig. 1. Schematic description of criteria used in data selection

The *rectangle at upper left* shows a small region of a 2-D pattern containing three spots in the master spot pattern (*continuous ellipses, M:1,2,3*) and three spots of the superimposed object spot pattern (*dashed ellipses, O:1,2,3*). The *rectangle at lower right* shows results obtained from a computation of gaussian overlap between each master and each object spot. According to the selection criteria used in this work, spots *M:3* and *O:3* would be considered as valid data (>.5 overlap, overlap for this pair >90% of total overlap of either with all other spots), the match of *O:2* with *M:1* or *M:2* would be rejected as ambiguous, and *O:1* would be found to be unambiguously absent from the master pattern (overlap of *O:1* with all master spots is <.001)

The assembly of a data set then consists in the selection of spots that are either unequivocally assigned or unequivocally absent on all gels in the set, i.e., have no missing data flags. This procedure results in an objective selection of "well-assigned" spots. In particular cases, a further selection involving restriction to a section of the gel or a predefined spot group can be used. In data sets where the same or similar samples are analyzed more than twice, it is possible to "fill in" some missing data flags by using the average of the values for that spot on the other gels of the same sample. We have used this method to fill in a maximum of one missing data flag per sample type per spot.

Multivariate Analysis

For these initial studies we used the ARTHUR 81 package of multivariate statistical-analysis software (Infometrix, Inc., Seattle, WA 98125). The initial data matrix of spot abundances was autoscaled to give a column (gel) mean of zero and a variance of 1.0. Principal component (PC) analysis was performed with use of the KAPRIN routine, and the data were transformed by using KATRAN into a space consisting of the first 10 principal components. A distance matrix was computed using euclidean distances in the transformed space after weighting of each coordinate by the percentage of total data variance represented by the corresponding PC axis (eigenvalue). This distance matrix was analyzed by a complete linkage cluster algorithm (routine HIER) to yield a dendrogram of similarity. The above analysis was performed both for the data viewed as spot measurements characterizing gels (gel space) or gel measurements characterizing spots (spot space).

Results

A series of experiments is necessary to determine the usefulness of multivariate statistical approaches in cell-type comparisons. We have examined the relationships of differences between cell types to (*a*) variations among different gels of the same sample and (*b*) variations among samples of the same cell type prepared on different occasions. Such experiments are required to demonstrate the importance of any major differences observed between cell types. Finally, we examined differences between groups of related cell types.

Comparison of Five Cell Types: Assessment of Gel-to-Gel Reproducibility

Replicate 2-D analyses were performed on [^{35}S]methionine-labeled samples prepared from each of five human cell lines (one sample per line). All the gels were run as one ISODALT batch, for maximum self-consistency. Gel resolution (defined in ref. *17*) was between 13 000 and 16 500 for these gels. The five cell lines were selected as in vitro representatives of genuinely distinct cell types; "GM607" is lymphoblastoid line, "1494" a normal fibroblast, "HTB-63" a melanoma line, "BT-20" a breast-tumor line, and "HTB-3" a bladder-tumor line. We analyzed a total of 16 gels (four triplicate runs and one quadruplicate run). The resulting patterns ("spotfiles") were treated and data selected as described in *Materials and Methods*, yielding a final data set of 285 spots from each of 16 gels, or 4560 measurements.

In this experiment, a gel (or, more precisely, a sample) can be looked upon as a point in a 285-dimensional space, marking by its position along each of 285 perpendicular axes the abundance of one of the 285 protein spots. We wish to know the distribution in this space of the 16 points that correspond to the gel patterns. In particular, we want to know whether the points corresponding to replicates of one sample are tightly clustered and whether the sets of points

corresponding to different cell types are well separated. Because it is not possible to "look" into a 285-dimensional space, we applied principal-component analysis as a means of reducing the dimensionality of the data while retaining as much of the information as possible. With this procedure one selects a new set of orthogonal axes in the original space along which most of the variation occurs. For the present data, the first three new axes (the largest three principal components) represent respectively 24.8%, 23.0%, and 21.6% of the variation in the original data (a total of 69.4%).

Figure 2 shows plots of the positions of the 16 gels on the three planes corresponding to the first and second, first and third, and second and third principal components. These views represent three different aspects of a cube comprising the first three dimensions of the principal component space. The groups of points representing replicates of a sample are in general tightly clustered, while the five cell types are all well separated in at least one plane. In order to display more clearly the distance relationships between the gels in the PC space, we calculated a distance matrix of all possible gel-to-gel distances. Each element consists of a euclidean distance in which separation along each PC axis was weighted by the eigenvalue associated with that axis, considered over the first six PC axes (91.7% of total variance). A complete-link cluster analysis of this distance matrix yielded the dendrogram shown (Figure 2). With the exception of one rather poor gel, all replicates of a single sample are more than 90% similar, whereas the similarity of the cell types ranges from 0% to 50% on this arbitrary scale.

The above results were obtained by using a large population of spots that included many qualitative markers—i.e., spots that were detected in some but not other lines. When the calculation was restricted to only the 58 spots that were detected (or filled in, as described earlier) on all 16 gels, the results shown in Figure 3 were obtained. Although the replicates are perhaps not quite as tightly clustered as when

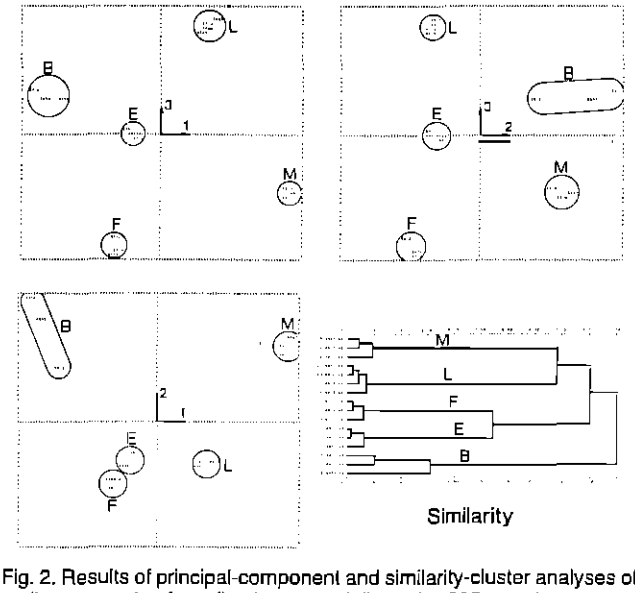

Fig. 2. Results of principal-component and similarity-cluster analyses of replicate samples from five human cell lines, for 285 protein spots

The three *square panels* show the positions of gels on three planes formed by all pairs of the first three principal components. Components 1, 2, and 3 account successively for 24.8, 23.0, and 21.6% of the total data variance (for a total of 69.4%). Euclidean distances between samples calculated in the PC space (first six coordinates after weighting each coordinate by the variance represented), and analyzed by using a complete-link clustering procedure, yield the dendrogram shown. Five groups of samples emerge, corresponding to the five cell types (*M* is the HTB-63 melanoma, *L* is the GM-607 lymphoblastoid, *F* is the 1494 fibroblast, *E* is the HTB-3 bladder epithelium, and *B* is the BT-20 breast tumor). *Circles* surrounding sample groups are intended to indicate only the rough group size, not variance

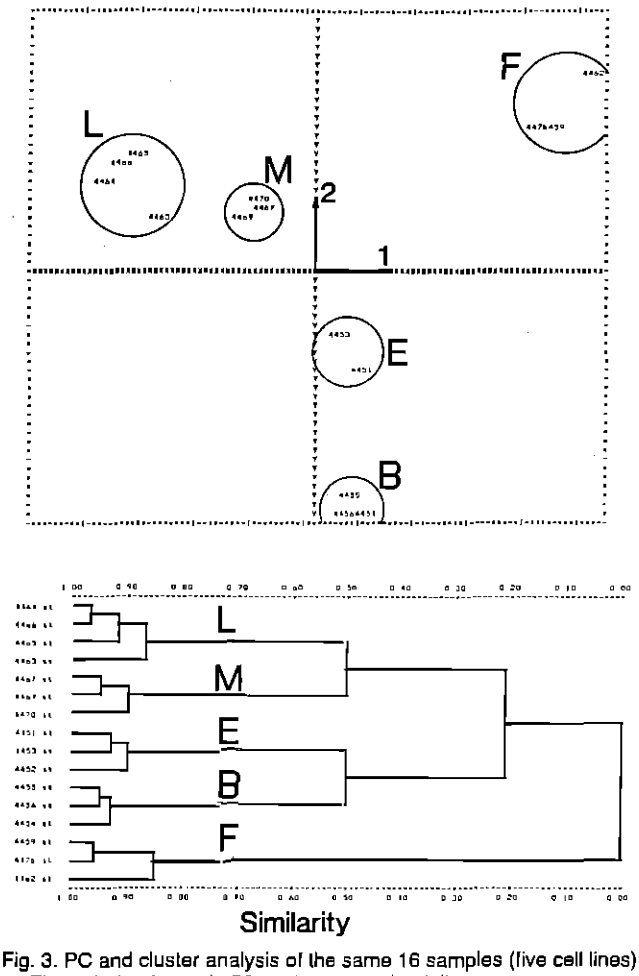

Fig. 3. PC and cluster analysis of the same 16 samples (five cell lines) as Fig. 2, but using only 58 spots present in all lines

The calculation is displayed as in Figure 2, except that only the first two principal components are shown. These two account for 29.1% and 24.9% of total data variance, respectively, and in the distance calculation used to produce the dendrogram we used the first four principal components (87.2% of total variance). The cell types are nearly as well resolved as in Figure 2

all the data were used, the cell types are nevertheless very clearly distinguished in the similarity dendrogram. The dendrogram resembles but is not identical to that obtained with the full data, indicating that relationships obtained with different protein sets and at low levels of similarity must be compared with caution.

In order to explore the relationships between proteins, we computed PC analyses and distance matrices as above, but with the additional step of autoscaling the values for each of the 58 spots. Figure 4 shows an image of the master spot pattern with an indication of one set of coregulated (or, more properly, "similarly regulated") spots derived from the full data dendrogram shown at the right.

Comparisons of the Same Cell Types Labeled on Different Days

Three of the cell lines used above were labeled on a series of separate occasions after subcultures were established. The gels analyzed (resolution 12 000 to 13 000) show BT-20 at one, two, three, and seven days, HTB-63 at two, three, and seven days, and the fibroblast 1386 at one, two, and seven days. The PC analysis and dendrogram shown in Figure 5 demonstrates that the cell types are well separated as compared with these replicate cultures, which remain well clustered. Differences in protein pattern ascribable to growth rate or state of medium depletion (variables that

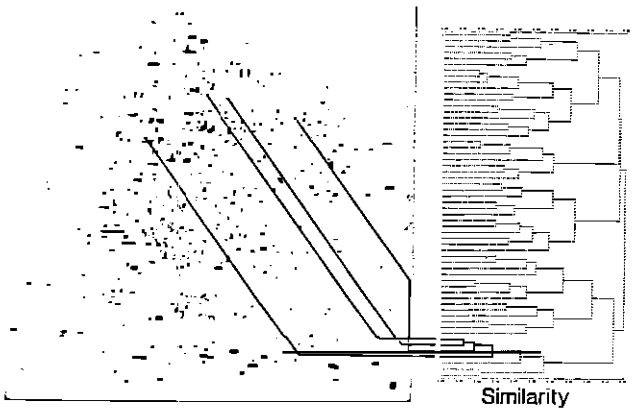

Fig. 4. An image of the synthetic master pattern (i.e., one containing all proteins observed in the five cell types), showing a set of four similarly regulated proteins taken from the full expression-similarity dendrogram (at the right)

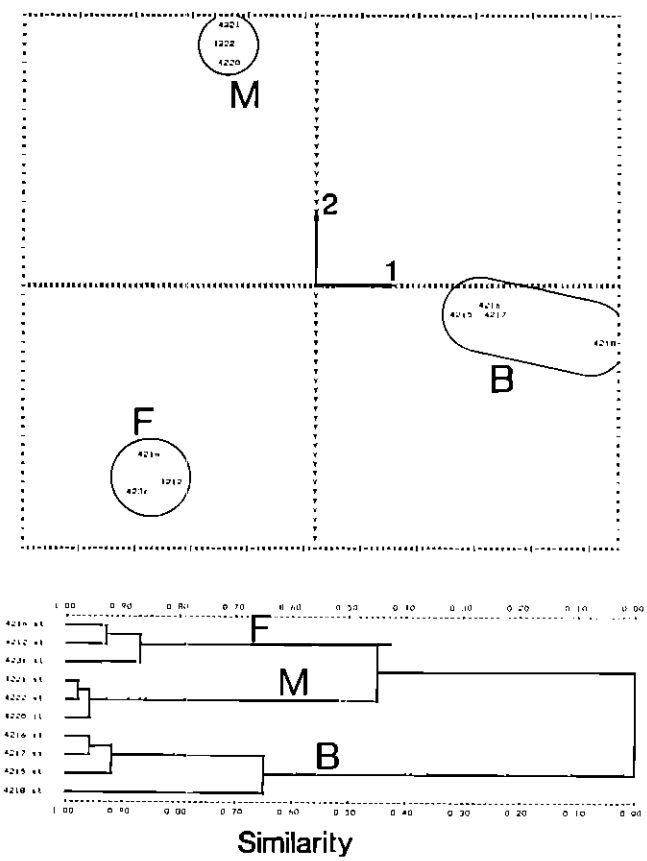

Fig. 5. PC and cluster analysis of 10 samples of three cell lines

In this case the samples from each cell type were prepared on different days, and thus include variation arising from cell-culture conditions. Principal components 1 and 2 account for 37.2% and 24.8% of total data variance, respectively. The dendrogram is based on the first four components accounting for a total of 79.1% of variance. Despite cell-culture variations, the three cell types are still well resolved

change substantially during a week of culture) thus appear to be small compared with the differences between types.

Analysis of a Differentiational Model System

Based on tests of specific marker proteins, it has been demonstrated that certain cultured cell lines can be made to alter their patterns of gene expression in vitro by treatment with various chemical agents. Figure 6 shows a PC analysis of the effects of an active phorbol ester (phorbol myristate acetate) and dimethyl sulfoxide on the human histiocytic

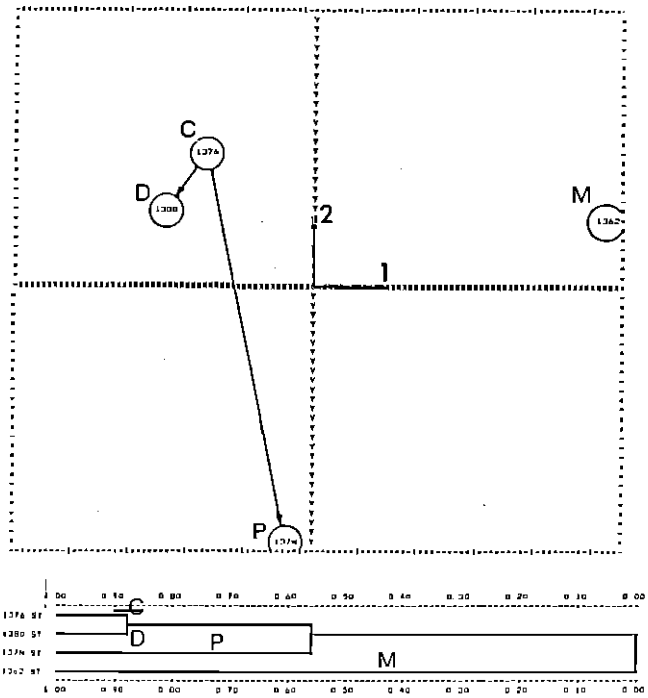

Fig. 6. Example of a prototype differentiation experiment

C is a control sample of U937 histiocytic lymphoma cells, *D* is a sample of U937 treated with dimethyl sulfoxide, 15 mL/L (little effect), *P* is U937 treated with 5×10^{-8} mol/L active phorbol ester (inducing differentiation), and *M* is a sample of normal peripheral blood monocytes. The induced differentiation is considerable, but does not appear to result in a shift towards the expression pattern of a monocyte. Components *1* and *2* account for 59.5% and 32.4% of total variance (92% total)

lymphoma line U937 (*14*), and the relationship of control and treated cells to the normal peripheral blood monocyte. Only proteins detected on all gels were used in the analysis. Gel resolution was 14 000 to 16 000. It is evident that dimethyl sulfoxide has little if any effect, while PMA produces substantial change, in accordance with its known physiological defects (15). The effect of phorbol myristate acetate appears not, however, to involve differentiation directly towards the expression pattern of the normal monocyte. A complete time course of the PMA-associated changes, and comparison of these with monocytes, macrophages, and activated macrophages will be required to fully characterize the differentiational changes occurring in this system.

Discussion

2-D protein patterns contain large amounts of quantitative data that directly reflect the functional status of cells. Although human observers are capable of searching such data for simple markers correlated with the available external information, global analysis—i.e., examination of the entire data—for complex patterns of change is extremely difficult. Differentiation, neoplastic transformation, and some drug effects are known to involve complex changes, and thus there is a requirement to develop an approach capable of dealing with data of this type. Ideally, one would like to use a method that could, by itself, discover the underlying logical structure of the gene expression control mechanisms. Such an approach, based on the methods of artificial intelligence and expert systems (*16*), is likely to become feasible within the next decade, if a large enough base of information is assembled. For the present, however, it is useful to explore the possibilities presented by the

application of statistical techniques for the detection of patterns of change and for evaluating the relationship between different patterns. This more-limited approach allows at least the measurement of similarities and differences between various complex patterns of change.

Using the statistical techniques of principal-component and cluster analysis, we have shown that a variety of human cultured cell types can be distinguished on the basis of complex patterns of protein expression. In the case examined, the cells could be distinguished almost as well by using quantitative differences in proteins expressed by all the cells as they could by using a full set of qualitative and quantitative markers. This result indicates that there is a wealth of useful information in purely quantitative cell-type differences.

We consider the concept of a multi-dimensional space of gene expression patterns very useful. The exact structure of such a space depends on the set of proteins used to define the coordinates, and hence can only be standardized by choosing representative, well-resolved, fixed groups of proteins. A great deal of work remains to be done before such groups can be chosen properly. Nevertheless, the possibility of using such a space to look at and measure "distances" between various cell types within an experiment is intriguing. Of special interest is the case of a cell line that moves in such a space over time (i.e., that differentiates). The simple prototype experiment reported here, involving the effect of a phorbol ester tumor promoter on the cell line U937, provides an example of an extensive change in the pattern of gene expression due to one widely used differentiation-inducing agent. Phorbol myristate acetate causes differentiation towards a monocytic form in HL-60 promyelocytic leukemia cells, and it causes the appearance of some monocyte markers in U937 as well (N.L.A. and A.G., unpublished). Another widely used agent, dimethyl sulfoxide (capable of inducing differentiation towards a granulocytic form in HL-60), is shown to have little effect on U937 under the same conditions. Principal-component analysis shows the gene expression changes brought about by phorbol myristate acetate to be almost orthogonal (i.e., generally unrelated) to those separating U937 cells from the peripheral blood monocyte, at least at the time point shown. Clearly, a detailed investigation of both the time course of the gene expression changes in U937 and HL-60—as well as the range of expression patterns associated with monocytes, macrophages, activated macrophages, and granulocytes—will be required to characterize adequately the phorbol myristate acetate-induced changes. In undertaking such an analysis, we hope to define the trajectories followed by these cells during in vitro differentiation. The ultimate goal of deducing the pathways of in vivo differentiation by analysis of a large series of in vitro systems, each exhibiting a part of the overall system, appears feasible if these statistical approaches are used.

References

1. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250, 4007–4021 (1975).

2. Anderson NL, Taylor J, Scandora Jr AE, et al. The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin Chem* 27, 1807–1820 (1981).

3. Aquadro CF, Avise JC. Genetic divergence between rodent species assessed by using two-dimensional electrophoresis. *Proc Natl Acad Sci USA* 78, 3784–3788 (1981).

4. Ohnishi S, Kawanishi M, Watanabe TK. Biochemical phylogenies of *Drosophila:* Protein differences detected by two-dimensional electrophoresis. *Genetics* 61, 55–63 (1983).

5. McLellan T, Ames GFL, Nikaido K. Genetic variation in proteins: Comparison of one-dimensional and two-dimensional gel electrophoresis. *Genetics* 104, 381–390 (1983).

6. Tarroux P. Analysis of protein patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. *Electrophoresis* 4, 63–70 (1983).

7. Westerbrink K, Havsteen B, Groenier K. Numerical taxonomy of two-dimensional protein maps: A rational biochemical approach to tumor characterization. In *Electrophoresis '82,* D Stathakos, Ed., W de Gruyter, Berlin–New York, 1983, pp 423–433.

8. Moll R, Franke WW, Schiller DL, et al. The catalog of human cytokeratins: Patterns of expression in normal epithelia, tumors, and cultured cells. *Cell* 31, 11–24 (1982).

9. Giometti CS, Anderson NL. Tropomyosin heterogeneity in human cells. *J Biol Chem,* in press.

10. Sneath PHA, Sokal RR. *Numerical Taxonomy,* WH Freeman, San Francisco, CA, 1973.

11. Gemmell MA, Anderson NL. Lymphocyte, monocyte, and granulocyte proteins compared by use of two-dimensional electrophoresis. *Clin Chem* 28, 1062–1066 (1982).

12. Anderson NG, Anderson NL. Analytical techniques for cell fractions. XXI. Two-dimensional analysis of serum and tissue proteins: Multiple isoelectric focusing. *Anal Biochem* 85, 331–340 (1978).

13. Anderson NL, Anderson NG. Analytical techniques for cell fractions. XXII. Two-dimensional analysis of serum and tissue proteins: Multiple gradient–slab gel electrophoresis. *Anal Biochem* 85, 341–354 (1978).

14. Sunderstrom C, Nilsson K. Establishment and characterization of a human histocytic lymphoma cell line (U-937). *Int J Cancer* 17, 565-577 (1976).

15. Grunberger G, Zick Y, Taylor SI, Gorden P. Tumor-producing phorbol ester stimulates tyrosine phosphorylation in U-937 monocytes. *Proc Natl Acad Sci USA* 81, 2762–2766 (1984).

16. Barr A, Feigenbaum EA (Eds.) *The Handbook of Artificial Intelligence,* 2, William Kaufman, Inc., Los Altos, CA, 1982.

17. Taylor J, Anderson NL, Anderson NG. Numerical measures of 2-D gel resolution and positional reproducibility. *Electrophoresis* 4, 338–346 (1983).

18. Taylor J, Anderson NL, Anderson NG. A computerized system for matching and stretching two-dimensional gel patterns represented by parameter lists. In *Electrophoresis '81,* RA Allen, P Arnaud, Eds., W de Gruyter, Berlin–New York, 1981, pp 383–400.