Proceedings of the Biology and Characterization of Cultured Vertebrate Cell Lines Symposium, October 2-5, 1983, Bethesda, MD, Tissue Culture Association, Gaithersburg, MD, 1984 pp. 189-195

COMPARISON OF ORGANISMS AND CELL TYPES USING TWO-DIMENSIONAL ELECTROPHORESIS

N. LEIGH ANDERSON

Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory, Argonne, Illinois

SUMMARY

Various approaches to the use of two-dimensional electrophoretic patterns for the quantitation of protein-taxonomic distances are discussed. In comparisons of closely related organisms it is concluded that approximately one-third of all single-base substitutions may be detected in approximately 3×10^6 base pairs of coding DNA (assuming 1000 usable spots/pattern). This represents, by far, the highest data rate available in any genomic surveillance technique aimed at measuring single-base changes. In comparisons of more distantly related organisms, current 2-D gels allow exploitation of a range of "spot overlap" extending from about 2.6% (for completely nonhomologous organisms, i.e. spot patterns random with respect to one another) to $\geq 90\%$ (for samples from the same organism). Within this range, evolutionary distances characteristic of family, genus, and species differences may be measurable directly in terms of base substitutions per nucleotide averaged over very large amounts of coding DNA.

INTRODUCTION

Two-dimensional electrophoresis (1-3) is currently the highest resolution protein separation technique available. It yields a pattern of spots in which each protein occupies, and is identified by, a specific position relative to other protein spots. This pattern can be analyzed by computer to yield positional and quantitative data (4) that can be accurately intercompared between analyses (5). The chief characteristic of such separations in the context of organism and cell-type identification is the fact that they allow use of a large population of 100 to 2000 protein markers rather than the limited number (5 to 30 proteins) usually investigated in isozyme studies. In this paper I aim to examine some general aspects of 2-D pattern comparison, beginning with an assessment of the information obtainable in comparisons of (a) closely related and (b) very distantly related organisms. Such comparisons assume the use of analogous cell types from the different species examined or else that whole organisms are analyzed (the organisms having essentially equivalent proportions of major cell types). The problem of comparing different cell types from the same species, a problem which in fact provides the ultimate motivation for this investigation, is dealt with at the end.

At least two previous studies (6,7) have used two-dimensional electrophoresis to estimate genetic distances between species. Aquadro and Avise (6) compared rodent species using visual inspection of differences. They estimated the genetic similarity F as the total proportion of spots shared (2 x shared spots/spots in gels 1 and 2), and obtained results in which the estimates of two observers differ by 0 to 16% depending on the pair of species examined. The spots to be compared were chosen on the basis of "clarity and good resolution." Ohnishi et al. (7) used a similar approach in comparing the 2-D maps of different *Drosophila* species. Although these studies succeeded in pointing out the potential usefulness of 2-D patterns in taxonomy, the methods used lack the rigor necessary to support widespread use of 2-D gels as a serious numerical taxonomic tool. Neither the gel resolution nor the positional reproducibility between gels was examined quantitatively, with the result that spot differences and similarities could not be uniformly and rigorously defined. In addition the use of a simple measure of spot overlap (F) is not entirely satisfactory for two reasons: first, it is sensitive to the number of observed spots on each gel; comparison of two gels of identical samples, one with 1000 and one with 100 visible spots, would yield only 18% overlap (2 x 100/1000 + 100) rather than the *true* value of 100%. Second, this measure treats all proteins equally, whereas it may be necessary to consider subunit molecular weight (i.e. genetic target size) as a major factor influencing probability of variation. Further limitations of such a Jaccard-type similarity measure are discussed by Sneath and Sokal (8). In this and subsequent papers I will attempt to provide a more rigorous analysis of the factors involved in 2-D gel taxonomy and particularly the application of computers to allow a statistically meaningful approach to the problem.

2-D Pattern differences between closely related organisms. Proteins have physicochemical properties [among which are isoelectric point (pI) and subunit molecular weight (sMW)] determined by coding regions of DNA (exons in eukaryotes). A protein of N amino acids will, in general, be coded for by a gene having 3N nucleotides in its exons. During evolution, a gene may be subjected to several types of modification including single-base changes, insertions, deletions, frameshifts, et cetera. The most frequent, judging from observed mutations in human hemoglobin, are single-base changes. Given the genetic code, it is relatively straightforward to compute the probability that a random base change in a coding triplet will give rise to an altered amino acid having a charge different from the original. Table 1 shows the results of such a computation for two interesting cases: (a) equal codon usage and (b) more nearly typical eukaryotic codon usage (computed from a set of known gene sequences). For simplicity, it is assumed that arginine, lysine, and histidine have +1 charge, aspartic and glutamic acids have -1 charge, and all other amino acids are uncharged; this assumption is not entirely correct at all pH's since cysteine and tyrosine, for example, can be negatively charged at pH 9.

Nevertheless, in the pH ranges normally examined on 2-D gels (~ 3.5 to ~ 7.5) it is clear that approximately 28% of single-base changes should give rise to a single- or double-charge change in the polypetide produced and a further ~ 4% should yield a change in size (through removal or addition of a termination codon). Thus, in comparing closely related but noninterbreeding species, a spot movement should occur after about one-third of random base substitutions. A situation involving comparison of individuals from a population is slightly more complex as it is more likely that only one of the two (for diploid organisms) copies of a gene will be altered. In such cases there is a necessity to detect a 50% decrease in abundance of the original spot (assuming equal expression of two genes) or the correlated appearance of a new spot.

How detectable is a charge change? To measure the frequency of such an event reliably, a charge change must be rather rigorously defined. Figure 1 shows the general appearance of charge changes in the major proteins of the human lymphoblastoid cell line GM607 (Genetic Mutant Cell Repository, Camden, NJ). To produce a series of charge changes in all the proteins simultaneously the urea-containing sample has been heated to produce carbamylation of lysine residues (10). Each lysine modified changes the protein's charge by -1, shifting it through a characteristic distance to the left (toward the acid end of the focusing dimension). The three key requirements for systematic detection of charge shifts in a particular protein are (a) the characteristic distance must be significantly greater than the size of the protein spot (i.e. charge isomers must be resolved), (b) the characteristic distance must be at least roughly predictable, and (c) the charge shift must be in a well-defined direction. The first requirement depends heavily on the 2-D gel system used: for the ISO-DALT system (2.3) using wide-range ampholytes, the size of a single-charge shift ranges from ~ 1.5 mm near the top (high sMW end) of the gel up to ~ 12 mm at the bottom (low sMW proteins). Typical full widths at half maximum for spots at the top and bottom of gels are 0.4 and 1.2 mm, respectively, indicating that the charge shift distances are generally 4 to 10 times the spot widths and thus that charge-shifted spots are easily resolved. As to the second requirement, although the characteristic distance associated with a charge shift can vary for different proteins of the same size (due to differences in the slopes of their titration curves at the pI), nevertheless a strong relationship exists between protein

TABLE 1

PROBABILISTIC EFFECTS OF RANDOM SINGLE-BASE SUBSTITUTIONS

Codon Usage[a]	Amino Acid Changes, %	Single Charge Changes, %	Double Charge Changes, %	Length Changes, %	Charge and Length Changes, %	Shift Frequencies, %				
						-2	-1	0	+1	+2
Random codon usage	76	24.3	1.4	7.9	33.6	0.7	12.1	66.3	12.1	0.7
Approximate eukaryotic codon usage	76	25.8	2.3	4.2	32.3	1.1	12.1	67.6	13.7	1.2

[a]The table shows the probability of various results for a random base substitution averaged over all 64 possible starting codons. In the first row, all codons (even termination codons) are assumed to occur at equal frequency. In the second row, each possible result has been weighted by the frequency of the starting codon in a collection of 27 relatively distinct eukaryotic gene sequences (including a yeast mitochondrial ATPase subunit, ADH1, iso-1-cytochrome C, GPD, dictyostelium actin, silk fibroin, *Ps.miliaris* histones, chicken ovalbumin, mouse immunoglobulins and beta lipotropin, rat growth hormone, prolactin and preproinsulin, rabbit α and β globin, bovine corticotropin-β-lipotropin, preparethyroid hormones, α-chorionic gonadotropin, human α-chorionic gonadotropin, pregrowth hormone, and chorionic somatommotropin) constituting about 27,000 coding nucleotides [taken from (9)]. All results are expressed as percentage probabilities of the indicated outcome.

size (sMW) and charge shift length. This consistency should allow prediction of characteristic distance for a given protein based on its position on the gel to within ± 30%. The third requirement, that charge shifts occur along a well-defined direction, is evidently well satisfied: in general a charge-shifted protein will appear at the same sodium dodecyl sulfate molecular weight as the present spot, allowing specification of shift direction as ± 5° of the horizontal (pI) axis.

In contrast to the single (or less frequent double) charge changes produced by ~ 28% of the base substitutions, the polypeptide size changes produced by ~ 4% of substitutions yield spot shifts that are not easily predicted. Changes in termination codons can lead to a mutated protein that is larger or smaller and of widely different pI. Thus a protein that undergoes such a "size-change" substitution during the divergence of two species will probably appear as distinct spots that cannot be related by any a priori procedure. Because other types of variation (frameshifts, mutations affecting mRNA splicing, neutral substitutions altering thermostability, et cetera) also produce such alterations in proteins, and because regulatory shutoff and induction of new genes cannot be factored out reliably, it is unlikely that an appreciable and known fraction of size-change, single-base substitution variants can be detected in species comparisons. The problems mentioned are decreased in comparisons of individuals from a population if strong assumptions can be made regarding the constancy of regulation of the proteins. If most proteins are synthesized at equal rates in two diploid individuals, then a spot present at 50% abundance may represent a size-change substitution in one copy of the gene. However, the instrusion of the other factors mentioned above still limits the detection of size changes in the absence of a pedigree analysis demonstrating anticorrelated expression of two spots.

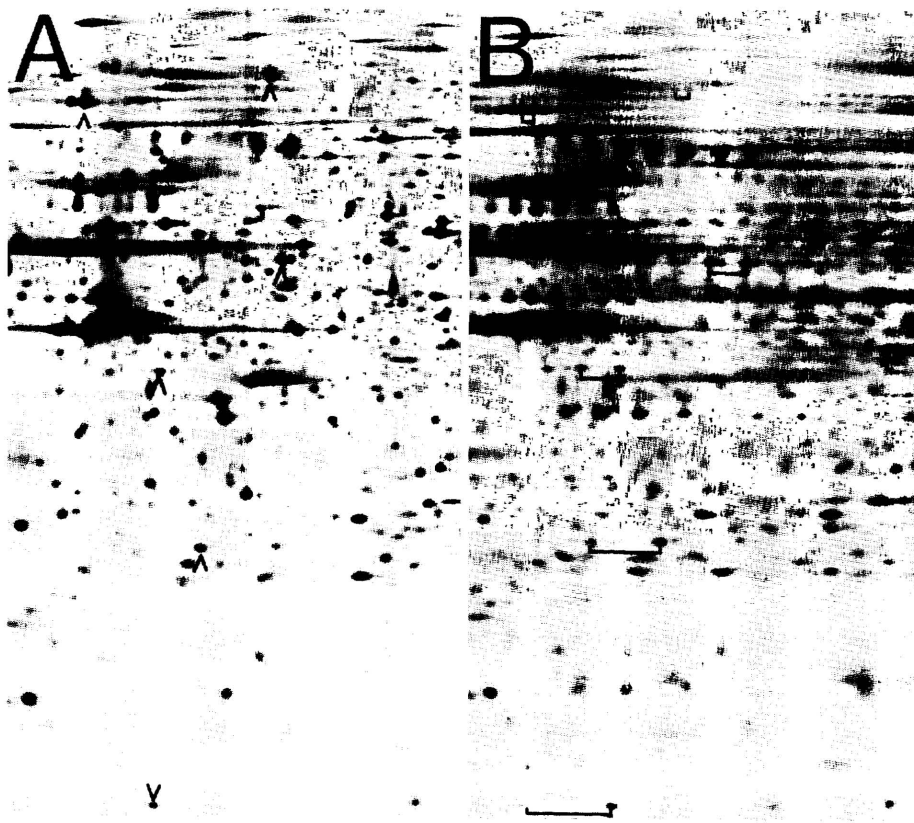

FIG. 1. Analogous portions of two 2-D protein patterns produced with the 7 x 7″ ISO-DALT system: *A*, Control (untreated) patterns of [^{35}S] methionine labeled proteins of the human lymphoblastoid cell GM607. *B*, The same proteins after a 5 min incubation at 95° C in a solubilization solution containing 9 *M* urea. Carbamylation by heating in urea produces numerous charge-shifts, revealing the size and direction (leftwards, towards the acidic end of the gel) of single-charge shifts for numerous distinct proteins.

How much information on DNA changes is obtained in a 2-D gel? There are approximately equal numbers of unique spots above and below actin (mol wt 42,000), but since typical 2-D gels show proteins from 8 to 10,000 daltons up to >200,000 daltons, the *average* molecular weight is likely to be at least 50,000 daltons. Thus, the average protein contains approximately 500 amino acids and is coded for by two copies (in diploid organisms) of a gene having 1500 base pairs (bp) in its exons. The average spot shows the result of the expression of perhaps 3000 bp of genomic DNA and a pattern showing 1000 unique spots (unrelated by posttranslational modification) would thus contain information corresponding to 3×10^6 bp of DNA. In comparisons of individuals from a population, approximately one-third of all single-base substitutions will result in a detected change; thus a pattern of 1000 unique spots is equivalent, in terms of genomic surveillance, to a detection system capable of seeing any single-base substitution in 1×10^6 bp of coding DNA. This represents by far the highest data rate per unit of experimental effort (i.e. gel) obtainable at present, and serves as the principal justification of the development of 2-D gel technology as a genetic monitoring tool.

It is also apparent from Fig. 1 that the higher molecular weight proteins show longer carbamylation trains, or in other words that they have experienced more modifications than smaller proteins under the same conditions. This is precisely what would be expected inasmuch as larger proteins generally contain more lysines than smaller ones and hence are likely to experience more carbamylation events; the number of modified lysines is proportional to "target size," in this case the total lysine content of the protein. An analogous situation exists with respect to the accumulation of random single-base substitutions. Because the DNA target size of a 100,000-dalton polypeptide is about 10 times as large as that for a 10,000-dalton one, it would be expected to accumulate 10 times as many random-base substitutions. More charge-change variants should therefore be observed at the top of the pattern than at the bottom, unless some selection process opposes this trend. A test of the postulated linear relationship between polypetide size and frequency of charge-change variation should, in fact, demonstrate whether any differential selection pressure exists on proteins of different molecular weight. If a simple linear relationship holds, it should be possible to extract a mutation frequency in total substitutions per nucleotide (and hence a useful evolutionary distance) directly from the data. Such a measurement, based on molecular changes averaged over a large number of proteins (and hence a large amount of DNA) may be superior to previous "molecular clock" data obtained from one or a few protein sequences.

Comparisons of distantly related organisms. What is the result of comparing the proteins of organisms so distant that little, if any, homology remains? Addressing this question requires exploring the relationship between patterns of spots that are random, or nearly random, with respect to each other. The answer determines the extent to which 2-D gel technology can be useful in establishing distant relationships. As might be expected, it depends heavily on the resolution of the gels used because we are asking about the capacity to differentiate between large nonidentical sets of proteins.

A general approach to the problem of comparing nonidentical patterns must take account of the fact that no a priori knowledge regarding spot homologies will be available; it is not known at the outset which spot, if any, in Pattern B is the homologue of a spot in Pattern A. Even given this limitation, at least two interesting statistics can be calculated. The first, a general overlap G, can be defined as the correlation coefficient describing the similarity of the overall distributions of spots. Functions representing the number density of spots (for G_N) or the total quantity of protein (for G_Q) over the gel area can be computed essentially by "smearing" or locally averaging the pattern, and the product of these functions interpreted over the area of the gel to yield a value of G. G is likely to be a useful measure primarily in cases of great evolutionary distance, where almost no proteins comigrate and only the general distribution remains.

A second useful measure is the histogram of vector lengths between spots in a Pattern A and those in a second Pattern B. The general approach is to generate vectors from each of N spots in A to each of M spots in B and from these to calculate histograms of number of vectors versus total length, length along the focusing dimension, et cetera. The resulting vector length histogram (VLH; Fig. 2) of the N x M vectors can be used to derive the number of vectors shorter than a given matching length (and thus the number of pairs of spots, one from A and one from B, which have positions close enough to indicate probable identity). The general form of the VLH relating two random patterns is a linearly increasing function of vector length: for reasonably short lengths, twice as many spots will lie at distance 2 x l as at distance l, due to the probability of their random occurrence being given by the differential of area ($2\pi l dl$), and the histogram will give zero vectors of zero length.

The primary questions in such an analysis are the following: (a) what distance (vector length) is to be taken as a threshold Δ within which we presume spots to be related? (b) What percentage of real matches (same spot on replicate gels) are collected within this threshold? (c) How many vectors fall within the threshold by chance (and thus constitute the irreducible statistical background in measurement)?

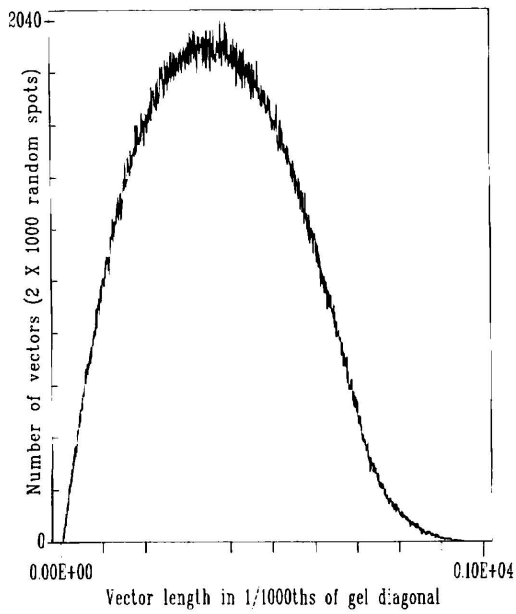

FIG. 2. Complete VLH comparing two square patterns of 1000 spots each of which are random with respect to each other. The number of vectors rises linearly with length for short vectors, allowing accurate prediction of background due to random close association of spots.

Questions a and b are closely related and indeed seem to define each other. Provisionally, we take the identity threshold Δ to be the distance within which lie 90% of the known matches within a set of gels of identical samples. In a study of reproducibility in the ISO-DALT system, Taylor et al. (5) found a value of $\Delta = 0.54$ mm for 7 x 7" gels, which is equivalent to 0.54/251 = 0.0022 of the gel diagonal (the longest possible interspot vector). A Monte Carlo calculation of the VLH between sets of random 1000 spot images gives ~ 26 vectors expected by chance within a Δ of 0.54 mm. Therefore, 26 of 1000 spots in a random pattern would be expected to overlap spots on another random pattern by chance, using this matching criterion. Improving the matching accuracy to yield a Δ of 0.25 mm would result in a decrease of chance matches to 4/1000. Nevertheless, a background of 2.6% chance matches is low enough to allow reliable detection of 5 to 10% real matches by this technique (Fig. 3).

DISCUSSION

The use of 2-D protein electrophoretic patterns for identification and classification of organisms is currently in its infancy. Nevertheless, two peculiar advantages justify considerable attention to the development of the method. First, inasmuch as all known organisms synthesize a collection of proteins, protein comparison methods such as those outlined should be uniformly applicable across all kingdoms. This wide applicability of the approach makes it possible to imagine the construction of a general taxonomy based on a uniform method, rather than use of characters specific to each group of organisms classified. Such a uniform approach might also facilitate some standardization of genus- and species-level

TABLE 2

CUMULATIVE NUMBER OF INTERSPOT VECTORS LESS THAN A GIVEN LENGTH FOR RANDOM PATTERNS EACH HAVING 1000 SPOTS. RESULTS OF A MONTE CARLO SIMULATION.

Length in 10^{-4} Gel Diagonal	Equivalent Length on 7 x 7" Gel, mm	Cumulative Number of Vectors Less Than Given Length
0	0.	4
10	0.25	20
20	0.50	49
30	0.75	98
40	1.00	170

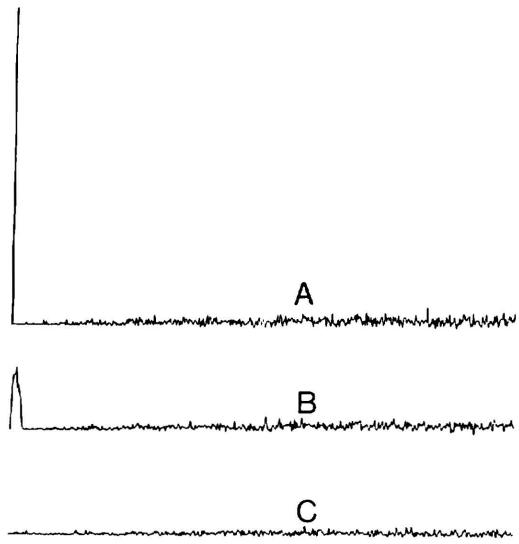

FIG. 3. Vector length histograms showing only the very short vectors (0 to 5 mm) for cases of: *A*, a pattern of 483 spots compared with itself (483 vectors of zero length). *B*, Two patterns of 483 spots in which identical spots are related by gaussian matching errors with half width 0.8 mm. *C*, Comparison of two random 483-spot patterns.

differences across phyla in terms of base substitutions per nucleotide. Second, the fact that the method allows the examination of many proteins (and hence a great deal of coding DNA) in a single procedure makes it inherently superior, from a statistical viewpoint, to molecular methods that examine one or a few proteins or genes in great detail. Individual proteins may vary enormously in their rates of change during evolution, and hence an average over many genes should be a more reliable measure than change in any individual gene.

Once methods are in place for measuring molecular-taxonomic distances using 2-D patterns routinely, the methods of numerical taxonomy (8) should be directly usuable for the construction of relatedness trees and tentative phylogenies. Alternatively, a cladistic-type approach could be explored if electrophoretically altered homologues of numerous proteins could be systematically identified. Such an approach might involve considerably more effort than the methods described here (because additional experiments or techniques of physicochemical characterization might be required to establish spot homologies) but the results could offer more direct insights into the pathways of evolutionary change.

An understanding of the mechanics of evolutionary change as observed in protein patterns is necessary to allow exploration of two additional questions of great interest. First, can the cell types of a single species be compared and classified by a similar approach to yield an "ontongenetic tree" compatible with cell lineages established in embryology? The development of a cell-type taxonomy based on protein expression patterns could allow the solution of many problems in differentiation and differentiation-related diseases, such as cancer. Second, assuming that homologous proteins of related species can be identified, will the quantitative aspects of gene regulation turn out to be generally conserved through evolution? Put another way, are homologous genes similarly regulated in different species? Data on this question covering a variety of proteins would allow an approach to the question of the relative roles of structural and regulatory changes in evolution.

Finally, it is worth considering the impact of a well-developed protein-pattern taxonomy on the problem of standardization of cell cultures. An experienced user of 2-D gels can distinguish human and mouse patterns by inspection. Using computer analysis, comparison of data from a test cell line with a sufficiently large data base should allow identification of both species and general cell type over a broad evolutionary range. Pushed to its limits, such a pattern identification system might be able to identify any known cell type, assuming the cells of all important species or groups had been previously mapped. Inasmuch as the effort involved in such an analysis would be reasonably small and the amount of useful and interesting data obtained rather large, the use of protein mapping methods could become routine and cell line misidentifications eliminated.

This work is supported by the U.S. Department of Energy, Washington, D.C., under Contract W-31-109-ENG-38.

References

1. O'Farrel, P. H. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 250:4007–4021; 1975.
2. Anderson, N. G.; Anderson, N. L. Analytical techniques for cell fractions. XXI. Two-dimensional analysis of serum and tissue proteins: Multiple isoelectric focusing. Anal. Biochem. 85:331–340; 1978.
3. Anderson, N. L.; Anderson N. G. 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.
4. Anderson, N. L.; Taylor, J.; Scandora, A. E.; Coulter, B. P.; Anderson, N. G. The TYCHO system for computerized analysis of two-dimensional gel protein mapping data. Clin. Chem. 27:1807–1820; 1981.
5. Taylor, J.; Anderson, N. L.; Anderson, N. G. Numerical measures of 2-D gel resolution and positional reproducibility. Electrophoresis. 4:338–346; 1983.
6. Aquadro, C. F.; Avise, J. C. Genetic divergence between rodent species assessed by using two-dimensional electrophoresis. Proc. Natl. Acad. Sci. USA 78:3784–3788; 1981.
7. Ohnishi, S.; Kawanishi, M.; Watanabe, T. K. Biochemical phylogenies of *Drosophila*: protein differences detected by two-dimensional electrophoresis. Genetica 61:55–63; 1983.
8. Sneath, P. H.; Sokal, R. R. Numerical taxonomy. San Francisco, CA: W. H. Freeman; 1973.
9. Grantham, R.; Gautier, C.; Gouy, M. Codon frequencies in 119 individual genes confirm choices of degenerate bases according to genome type. Nucleic Acids Res. 8:1893–1912; 1980.
10. Anderson, N. L.; Hickman, B. J. Analytical techniques for cell fractions. XXIV. Isoelectric point standards for two-dimensional electrophoresis. Anal. Biochem. 93:312–320; 1979.

Discussion

Dr. Siciliano: What are the mutation results? One slide indicated that you had exposed cells to mutagens. You said you had looked at some 70.

Dr. Anderson: I used a weaker mutagen than you did and have not found any mutations.