Protein Variants in Human Cells: Enumeration by Protein Indexing[a]

N. LEIGH ANDERSON,[b] CAROL S. GIOMETTI,[b]
M. ANNE GEMMELL,[b] AND MARVIN MACY[c]

[b]*Molecular Anatomy Program*
Division of Biological and Medical Research
Argonne National Laboratory
Argonne, Illinois 60439
[c]*American Type Culture Collection*
Rockville, Maryland

INTRODUCTION

In any attempt to construct a catalog of the proteins of a given species, the genetic heterogeneity of natural plant and animal populations makes it necessary to consider variants of each protein. Thus in compiling a Human Protein Index[2] using high-resolution two-dimensional electrophoresis as a separating technique, protein variants differing from the wild type by charge (about one third of amino acid variants) or by polypeptide length (a smaller fraction of variants) will be recognized as different and must be accounted for. A high level of variation among individuals would clearly make the task of protein indexing much more difficult, since the system would have to cope with numbers of protein spots much larger than the number of structural genes. To some extent the feasibility of cataloging human proteins thus depends on limited genetic variability, and measurements of this variability are consequently of great interest. Results of several recent investigations[3-5] have argued for average heterozygosities of approximately 1% for human cellular proteins examined by two-dimensional electrophoresis, while the results of classical biochemical genetics (using one-dimensional gels of serum or soluble cellular enzymes) lead to higher results (about 6%).[6] Average heterozygosities above a few percent yield proportions of "new" spots in individual patterns that might be difficult to deal with in large studies.

On the other hand, the ability to observe genetic variation in large numbers of proteins can be valuable in several contexts. More than 2,000 human genetic diseases have been identified,[7] the vast majority of which have not as yet been associated with a defect in a particular protein. Correlation of the appearance of protein variants with the presence of disease could serve to identify the protein products of the mutated genes, leading to better diagnosis and the possibility of protein replacement. A second useful context is the monitoring of experimentally (or naturally) induced new mutations. Any tolerable level of mutation will produce alterations in only a small number of proteins in an individual or cloned population, so that screening of individuals or clones is most efficient if a large number of proteins (loci) is examined in each. A third area of widespread usefulness concerns the technology of protein cataloging itself and the ability to generate for inclusion in the catalog a library of

[a]Supported by the March of Dimes (at the ATCC) and the U.S. Department of Energy under contract No. W-31-109-ENG-38.

TECHNOLOGY IMPACT:
POTENTIAL DIRECTIONS FOR LABORATORY MEDICINE Vol. 428
Reprinted from
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

THE NEW YORK ACADEMY OF SCIENCES · 1817 ·
PER CŒLESTIS PER ÆREM PER MARIA PER TERRAS

electrophoretic variants of many proteins, preserved in permanent cell lines or cloned DNA. Such variants can be used to identify their wild-type form on two-dimensional analyses; to help identify nucleic acid sequences coding for a protein, to investigate possible physiological effects of alterations in a protein, and to provide additional genetic markers for cell fusion and other studies. The availability of a library of variant hemoglobins, for example, has proved extremely useful in establishing the molecular mechanism of hemoglobin action, which in turn allowed an understanding of the pathophysiology of the variants.[8] Ultimately such libraries must be constructed for each variable protein in man.

The present study of 63 human fibroblast cell lines was initiated in order to determine the level of genetic variation at many loci and to see whether known genetic diseases were associated with any obvious variant proteins that might be expressed in fibroblasts in culture. The main result is a group of ten new putative variants available in permanent cell lines.

MATERIALS AND METHODS

Growth and Labeling of Cells

Human fibroblast cell lines obtained from the cataloged stocks of the American Type Culture Collection, Rockville, MD (CCL and CRL lines) or from the Genetic Mutant Cell Repository, Camden, NJ (GMR lines) were grown up and labeled with [^{35}S]methionine at the American Type Culture Collection using a protocol developed at Argonne.[1] Briefly, the cells were grown in 24-well tissue culture plates in complete MEM medium and then labeled for about 18 hr in medium containing approximately 60 μCi/ml [^{35}S]methionine and no nonradioactive methionine. After labeling, medium was aspirated off and cells were solubilized directly in 9 M urea, 4% NP-40, 2% mercaptoethanol, 2% ampholyte. Samples were then frozen and sent to Argonne on dry ice.

Two-dimensional Analysis

Samples were analyzed at least twice using the 7″ × 7″ ISO-DALT system I.[9,10] One set of gels was fluorographed, the others autoradiographed. For determination of the presence of likely variants, a series of 96 candidate proteins was marked on a copy of a representative gel. This image was copied again (on direct-duplicating film [Kodak]) to yield two sets of clear copies. Using these templates, two observers independently examined the original patterns of each cell line for likely variants of each of the designated proteins and recorded the results on the corresponding template (one for each cell line). A likely heterozygous variant is a "new" spot near which there is a "common" spot of nearly equal abundance and very similar SDS-molecular weight, but different pI (shift consistent with about one normal charge difference or less). Results were correlated, merged, and re-examined (in some cases the samples were re-run) in order to obtain a final list of variants. The nomenclature of variants used here (V1, 2, etc.) is a provisional one, referring to putative fibroblast variant proteins in a series found by the Molecular Anatomy Program. Variants of a named protein, such as human tropomyosin:3,[11] are designated by numbers after a decimal point (in order of discovery), as in Tm:3.1, elaborating on the nomenclature for protein sets described earlier.[12]

TABLE 1. Diseases and Cell Lines

Disease	Line	Individual (Family)
Acute intermittent porphyria	GM1622	proband
	GM1623	affected brother
Argininosuccinicaciduria	GM2830	proband– –
	GM2831	mother + –
	GM2832	father + –
Basal cell nevus syndrome	CRL1175	proband
Citrullinemia	CCL76	proband
Cockayne syndrome	CRL1336	proband
Cri du chat	CCL90	proband
	CCL123	proband
Cutis laxa	CRL1396	proband
Cyclops	CRL1246	proband
Cystic fibrosis	GM1013	proband (93)
	GM1708	father + – (93)
	GM1957	proband (94)
	GM1959	affected brother (94)
	GM2803	unaffected brother (336)
	GM2827	proband (336)
Dermatomyositis	CRL1244	proband
Down's syndrome	CCL54	proband
	CCL66	proband
	CCL84	proband
Ehlers-Danlos syndrome	CRL1131	proband
	CRL1138	proband
	CRL1144	proband
	CRL1243	proband
	CRL1326	proband
	CRL1327	proband
	CRL1332	proband
Eosinophilic fascitis	CRL1389	proband
Epidermolysis bullosa	CRL1331	
Familial hypercholesterolemia	GM1354	mother + –
	GM1355	proband – ?
	GM1385	father + –
	GM1386	mother + +
Fibrodysplasia ossificans progressiva	CRL1241	proband
Galactosemia, symptomatic	CCL132	proband
Hereditary adenomatosis	CRL1533	proband
Klinefelter's syndrome (XXXXY)	CCL28	proband
Langer-Giedion syndrome	CRL1400	proband
Leiomyoma uteri	CRL1309	proband
Lesch-Nyhan syndrome	CRL1111	proband
	CRL1112	affected brother
Marfan's syndrome	CRL1174	proband ?
Melorheostosis leri	CRL1345	proband
Menke's kinky-hair syndrome	CRL1230	proband
Methylmalonicaciduria	CCL124	proband
Multiple congenital abnormalities	CCL117	proband
"Normal" (MRC-5)	CCL171	
Osteogenesis imperfecta	CRL1286	
	CRL1293	
Poikiloderma	CRL1351	proband
Porokeratosis	CRL1310	proband
Pseudoachondroplasia	CRL1231	proband

TABLE 1. *Continued*

Disease	Line	Individual (Family)
Pseudoxanthoma elasticum	CRL1374	proband
Stiff-skin syndrome	CRL1388	
Systemic sclerosis	CRL1108	proband
Turner's syndrome	CCL65	proband
Xeroderma pigmentosum	CRL1161	
	CRL1165	
	CRL1167	
	CRL1168	
	CRL1223	
	CRL1254	

The column headed Individual (Family) indicates the genetic status of the person from whom the cell line was derived; the proband is the principal affected family member. Pluses and minuses indicate presence of wild-type (+) or variant (−) disease-associated genes (where known). Numbers in parentheses indicate family number where more than one family was examined for a given disease.

RESULTS

The 63 cell lines examined in the study are listed in TABLE 1. Probable genetic variants of 10 proteins were detected (TABLE 2; FIGURES 1 and 2). In each case further analysis including partial proteolytic cleavage product comparisons and partial amino acid composition comparisons[1] will be required to confirm the association of each variant with a wild-type protein.

Association of Variants with Specific Diseases

TABLE 2 indicates that variants 1, 3, and 13 do not correlate with the diseases associated with the cell lines in which they were found, and that variants 14 and 17 appear unlikely to correlate. Variants 11, 12, 15, and 16 may be associated with specific diseases, but as only a single line was examined in each case a broader base of lines with these diseases (Cockayne's syndrome, melorheostosis leri, basal cell nevus syndrome) must be investigated to eliminate the possibility of accidental association. Basal cell nevus syndrome is an autosomal dominant disease[7] and thus might not be expected to occur in an apparently homozygous form (as V16 in CRL 1175). Melorheostosis leri is not known with certainty to be a genetic disease.[7] The Cockayne syndrome is recessive,[7] and hence the appearance of the two variants (of different proteins) in apparently heterozygous forms in line CRL 1336 makes it uncertain whether either is related to the disease.

Variant V18 appeared in the lines examined here to be a good candidate for the low-density lipoprotein (LDL) receptor, a defect in which is responsible for familial hypercholesterolemia.[15] However, an examination of additional cell lines (GM3040A, 3064, 3065, 701A, and 486) indicates that if V18 is a variant of the LDL receptor, other non–charge-change variants are also involved. The receptor has been isolated from bovine tissue[16] and has a single-chain molecular weight of approximately 164,000, which is very close to the SDS-molecular weight of V18 (about 165,000 based on rabbit muscle molecular weight standards).

TABLE 2. Association of Variants with Specific Diseases

Fibroblast Variant	Probable Wild Type (PWT) of the Protein (Spot #)	Disease	Cell Line	Presence in Cell Line		Presence in Gene	
				Variant	PWT	Disease	Normal
1	F26	Familial hypercholesterolemia	GM1355	–	+	+	+
			GM1385	+	+	+	+
			GM1386	+	+	–	+
			GM1354	–	+	+	+
3	F312	Trisomy 21	CCL54	+	+	+	?
			CCL66	–	+	+	?
			CCL84	–	+	+	?
11	F102	Cockayne syndrome	CRL1336	+	+	+	?
12	F7	Melorheostosis leri	CRL1345	+	+	?	?
13	F99	Argininosucciniacidurea	GM2830	+	+	+	–
			GM2831	+	+	+	+
			GM2832	–	+	+	+
14	F19	Xeroderma pigmentosum	CRL1223	+	+	+	?
			(similar variant also present in normal line GM607)				
15	F17	Cockayne syndrome	CRL1336	+	+	+	?
16	F54	Basal cell nevus syndrome	CRL1175	+	–	+	?
17.1, .2	F4	Symptomatic galactosemia	CCL132	1	+	+	?
		Methylmalonicacidemia	CCL124	2	+	+	–
		Systemic sclerosis	CRL1108	2	–	?	?
18	F14	Familial hypercholesterolemia					
		(Pro)	GM1355	+	+[a]	+	+
		(Fa)	GM1385	+	+[a]	+	+
		(Bro)	GM1386	–	+	–	+
		(Mo)	GM1354	–	+	+	+

[a] Indicates probable presence of a second variant indistinguishable electrophoretically from the normal. +, – indicate presence or absence. Pro: proband, Fa: father of proband, Bro: brother of proband, Mo: mother of proband.

Overall Frequency of Variants

Two observers independently screened all 63 lines for variants of 96 easily located protein spots. Seven of the 96 proteins were subsequently disallowed because of either their peripheral location in the pattern or evidence that they showed large quantitative variations from line to line, probably associated with differences in growth conditions between batches of lines.

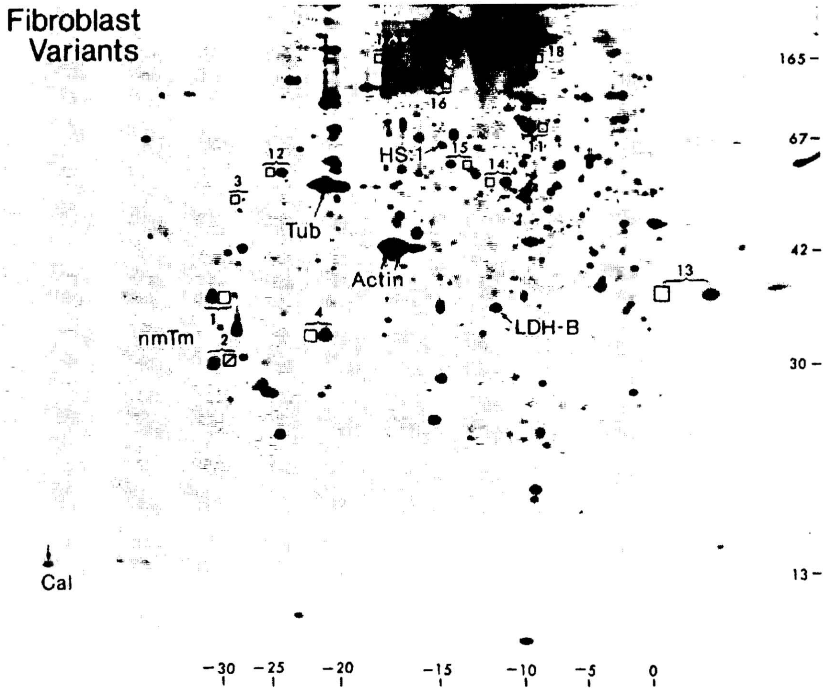

FIGURE 1. Fluorogram of a two-dimensional pattern of a representative human fibroblast line (CRL 1309). Square boxes indicate the positions of putative variant forms of the adjacent protein spots enclosed in brackets. Numbers associated with brackets are the numbers of these variants (V1, etc.) in a provisional nomenclature system as described in Materials and Methods. Labels indicate actin (β and γ forms), tubulin (Tub), LDH-B chain, calmodulin (Cal), the major heat-shock protein (HS:1, Ref. 13), and a major non-muscle tropomyosin (nmTm). Acid pI is to the left and high molecular weight at the top in accordance with Cartesian convention; scales along the bottom and right side mark positions of creatine phosphokinase pI standards[14] and approximate SDS-molecular weight in kilodaltons, respectively.

Among the 89 apparently "reliable" proteins, we found nine presumed variants in the 63 lines (the putative LDL receptor was not among the spots in the screen). Each of them occurred only once or in one family, with the exception of a V17, which occurred in three individuals. We could thus conclude from this very limited sample that only one of the examined proteins shows a common charge-change genetic polymorphism, while 9% appears to show a "rare" variant. The average chance that a given protein

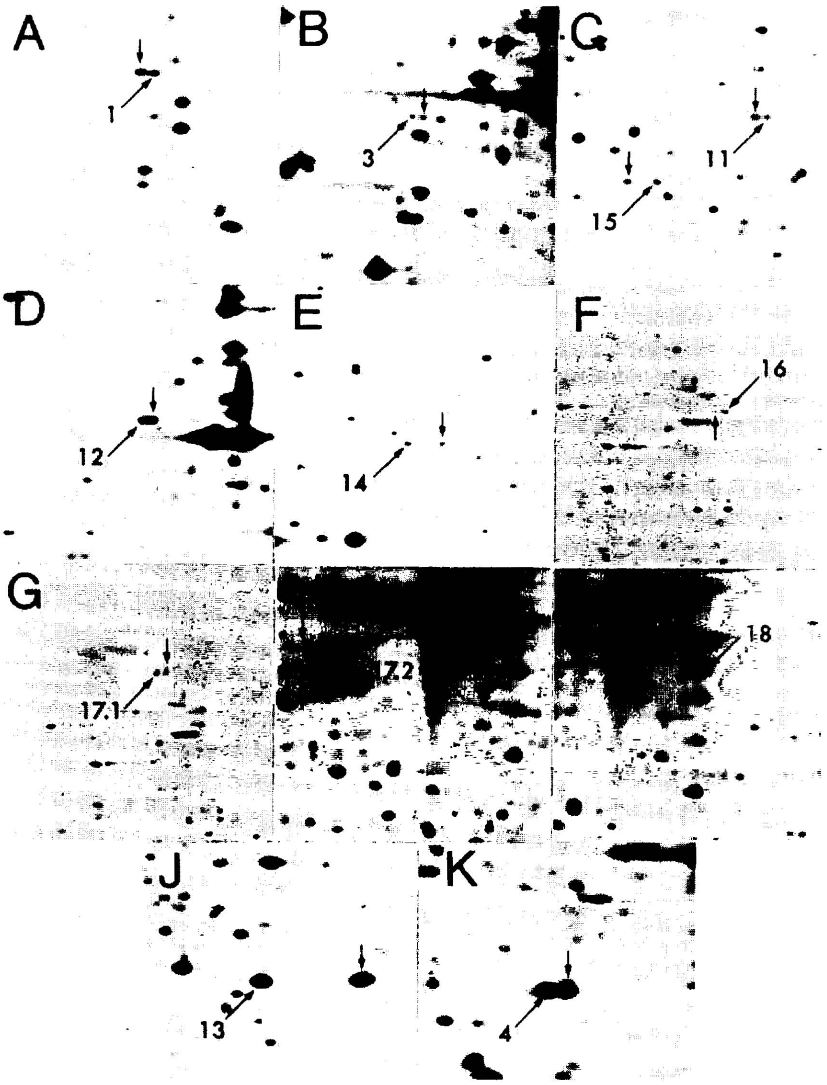

FIGURE 2. Sections of fluorograms or autoradiograms showing the appearance of the variants described here. Small downward-pointing arrows point to positions of putative wild-type spots, while longer slanted arrows indicate variant proteins numbered as in FIGURE 1 and TABLE 2. Patterns are from the following cell lines: (A) GM1386, (B) CCL54, (C) CRL1336, (D) CRL1345, (E) CRL1223, (F) CRL1175, (G) CCL132, (H) CRL1108, (I) GM1355, (J) GM2831, and (K) CCL124.

(from among this group) would be detected as charge heterozygous in a given individual is approximately 0.2%.

DISCUSSION

The level of genetic variation among human individuals, assessed using two-dimensional electrophoresis of major cellular proteins, appears to be quite low. This conclusion is in general agreement with results of previous studies[3-5] using two-dimensional gels to examine much smaller numbers of cell lines, in that the level of variation detected is much lower than would have been expected from classical biochemical genetics of human plasma proteins and soluble enzymes.[6] This discrepancy is not attributable to a lack of sensitivity of the two-dimensional system to electrophoretic variants, since the two-dimensional system (ISO-DALT) used in this stüdy reveals a large majority of plasma and red cell protein variants detected by the widely used one-dimensional techniques.[17,18] New polymorphisms detected by a two-dimensional system (such as that in α_2HS glycoprotein[17]) have also been confirmed by classical methods.[19] Therefore it seems necessary to conclude that major cellular proteins display much lower genetic variability in human populations than the restricted sets of proteins studied previously by biochemical geneticists. Since the population of proteins observed on two-dimensional gels can be regarded as essentially unselected (including soluble and cytoskeletal/nuclear proteins, enzymes, and structural proteins, etc.), the results of this and other two-dimensional studies appear to be more representative of the "average locus" than classical results. The fact that the average heterozygosity of cellular protein loci is quite low substantially improves the prospects for building a useful Human Protein Index.[2]

Although ten strong candidate variants were found in this study, none appears very likely to be diagnostic for a genetic disease. A possible variant of the LDL receptor was found, but the identification is equivocal because of the possibility that multiple alleles are involved (or that V18 is a variant of an entirely different protein). Familial hypercholesterolemia is believed to be the most prevalent genetic disorder in man (heterozygous in 1 in 500), being more frequent than either sickle cell anemia or cystic fibrosis (p. 211).[7]

Less than 0.5% of the estimated 30,000 to 50,000 human protein gene products coded for by the human genome has been examined in these studies. Given the small fraction of the total examined, the fact that expression of proteins characteristic of highly differentiated cells may not occur in cultured fibroblasts, that the more basic proteins seen on non-equilibrium gradient electrophoresis gels[20,21] were excluded, and that only about one third of single amino acid substitutions are detectable as charge shifts in these experiments, it is not surprising that a spot polymorphism correlating with cystic fibrosis was not observed.

Considered as the beginnings of a protein-variant library, however, the results of the present study can have great value. The putative variants found are reproducible markers, present in cell lines that are permanently available at low cost from major cell repositories. They can be used to identify the adjacent (presently putative) wild-type spot in a manner that is independent of gel systems. In addition, once functional information is available concerning the associated wild-type spot, such variants may be used to search for correlated physiological abnormalities in the affected cell lines or to help in the positive identification of RNA or DNA sequences corresponding to the protein in question.

As an example, the present collection of variants shown in FIGURE 1 contains mutant forms of two different tropomyosins. One, here called V2 but discovered in a previous study,[1] has been shown using a variety of techniques to be a variant of the tropomyosin designated Tm:4. The other, V1, discovered in the present study, occurs in two members of one family (and is thus almost certainly a genuine mutation) and appears to be a variant of Tm:3.[11] Only one of the several tropomyosin spots present in fibroblasts[23] is doubled by the presence of either variant, indicating that at least Tm:3 and Tm:4 are the products of different genes, not derived from each other or other tropomyosins by post-translational processing. Since a spot like Tm:3 is expressed in smooth muscle and fibroblasts, it would be possible, if smooth muscle could be obtained from the donors of lines GM1385 or GM1386, to determine whether the same gene was being expressed to make this polypeptide in each cell type. If it was, then the variant should appear also in smooth muscle; if not, then two genes with different expression patterns exist for "Tm:3-like" molecules. Likewise for Tm:4, which appears to be expressed in almost all non-muscle cells (though at variable levels),[22] and is highly conserved during evolution.[23] Both variants offer excellent opportunities for the investigation of possible cytoskeletal abnormalities associated with small alterations in one of the major cytoskeletal proteins. In addition, the availability of the variants will facilitate assignment of Tm:3 and Tm:4 as the products of particular cloned genomic DNA sequences. Finally, the availability of the charge variants will make possible the mapping of tropomyosin genes to chromosomes using cell hybrids; this was not previously possible, especially for Tm:4, since these proteins are highly charge-conserving in the mammals (unpublished studies) and the human gene product normally seen is electrophoretically indistinguishable from that produced in a mouse or hamster parental line.

Recently, a third tropomyosin variant has been found by cloning cells of line 1386.[24] This variant appears to be a further altered version of variant V1, thus suggesting the likelihood of double mutation in the smooth muscle tropomyosin gene and the possibility that the first mutation (leading to V1) generated a state of increased susceptibility to subsequent mutation.

Altogether, the results described here support the notion that large numbers of human cell lines (particularly all those contained in the major permanent collections) can and should be examined by two-dimensional electrophoresis. An expanding library of genetic variants, with disease correlations when available, will constitute a major resource for many areas of human biology.

ACKNOWLEDGMENTS

We wish to thank our colleagues in the Molecular Anatomy Program and the American Type Culture Collection for help and useful discussions.

REFERENCES

1. GIOMETTI, C. S. & N. L. ANDERSON. 1981. J. Biol. Chem. **256:** 11840–11846.
2. ANDERSON, N. G. & N. L. ANDERSON. 1982. Clin. Chem. **28:** 739–748.
3. WALTON, K. E., D. SLYER & E. I. GRUENSTEIN. 1979. J. Biol. Chem. **254:** 7951–7960.
4. MCCONKEY, E. H., B. J. TAYLOR & D. PHAN. 1979. Proc. Natl. Acad. Sci. USA **76:** 6500–6504.
5. RACINE, R. R. & C. H. LANGLEY. 1980. Nature **283:** 855–857.

6. HARRIS, H., D. A. HOPKINSON, Y. H. EDWARDS. 1977. Proc. Natl. Acad. Sci. USA 1977, **74**, 698–701.
7. MCKUSIK, V. A. 1978. Mendelian Inheritance in Man. Fifth edit. Johns Hopkins University Press. Baltimore, MD.
8. PERUTZ, M. F. & H. LEHMANN. 1968. Nature **219**: 902–909.
9. ANDERSON, N. G. & N. L. ANDERSON. 1978. Anal. Biochem. **85**: 331–340.
10. ANDERSON, N. L. & N. G. ANDERSON. 1978. Anal. Biochem. **85**: 341–354.
11. GIOMETTI, C. S. & N. L. ANDERSON. 1983. J. Mol. Biol. **173**: 109–123.
12. ANDERSON, N. L. 1981. Proc. Natl. Acad. Sci. USA **78**: 2407–2411.
13. ANDERSON, N. L., C. S. GIOMETTI, M. A. GEMMELL, S. L. NANCE & N. G. ANDERSON. 1982. Clin. Chem. **28**: 1084–1092.
14. ANDERSON, N. L. & B. J. HICKMAN. 1979. Anal. Biochem. **93**: 312–320.
15. BROWN, M. S. & J. L. GOLDSTEIN. 1974. Proc. Natl. Acad. Sci. USA **71**: 788–792.
16. SCHNEIDER, W. J., V. BEISIEGEL, J. L. GOLDSTEIN & M. S. BROWN. 1982. J. Biol. Chem. **257**: 2664–2673.
17. ANDERSON, N. L. & N. G. ANDERSON. 1977. Proc. Natl. Acad. Sci. USA **74**: 5421–5425.
18. WANNER, L. A., J. V. NEEL & M. H. MEISLER. Am. J. Hum. Genet. **34**: 209–215.
19. COX, D. W. & B. J. ANDREWS. 1983. *In* Electrophoresis '82, D. STATHAKOS, Ed.: 243–247. W. de Gruyter. New York.
20. O FARRELL, P. Z., H. M. GOODMAN & P. H. O'FARRELL. 1977. Cell **12**: 1133–1142.
21. WILLARD, K. E., C. S. GIOMETTI, N. L. ANDERSON, T. E. O'CONNOR & N. G. ANDERSON. 1979. Anal. Biochem. **100**: 289–298.
22. GIOMETTI, C. S. & N. L. ANDERSON. (Manuscript in preparation.)
23. GIOMETTI, C. S., K. E. WILLARD & N. L. ANDERSON. 1982. Clin. Chem. **28**: 955–961.
24. ANDERSON, N. L., M. A. GEMMELL & C. S. GIOMETTI. (Manuscript in preparation.)

DISCUSSION OF THE PAPER

R. TRACY (*Mayo Clinic, Rochester, MN*): In doing the mixing experiments with the hybrid cell lines, you indicated that the proteins that were expressed in the hybrids were homologous proteins, the high molecular weight acidic proteins. Is there any way to pin that down besides location on the gel?

N. L. ANDERSON: Yes, one could cut the spots out and do partial proteolytic digestions and run the fragments. This is the normal way to do such things, but I'm not sure how much conservation you would have between species of that pattern.

The only rigorous way to do it is to cut the proteins out and do microsequencing to show they have the same terminal sequence or else to use antibodies to that protein to specifically react to it. In these cases we didn't have either those processes or those reagents available. However, you and I both know that you can look at the gels and know that those are the same protein. In the case of those two little runs of spots, those both happen to be spots for multiply-phosphorylated proteins with several other properties that correlate between the two species so it is hinting but not proving that they are the same thing.