CLIN. CHEM. 28/4, 739–748 (1982)

The Human Protein Index

Norman G. Anderson and Leigh Anderson

The role of the Human Protein Index System (*1*, *2*) in clinical chemistry and in pathology is best understood by considering how a molecular pathologist will use the system diagnostically. His problem, as he views a high-resolution color display (*3*) of several thousand individual proteins (as illustrated in Figure 1) from a few milligrams of human tissue (*4*), from a frozen section (*5*), or from several thousand cells from tissue culture (*6–8*), is to find those differences that may be the causes of disease, may be indicative of existing disease, or may suggest predisposition to future disease.

He is familiar with many of the details of the pattern before him, and recognizes cytoskeletal proteins (*9*), a group of proteins in the HL-A region, a number of surface proteins, the major mitochondrial proteins (*10*), and several of the glycolytic enzymes that are present in the soluble phase of the cells or tissue being examined. He is also aware that most of the proteins seen have never been isolated, and that their functions remain to be discovered. But he is reassured by the knowledge that all of the proteins he sees (unless he will this day discover a new one, or a new variant of an old one) have been assigned numbers and map coordinates, and that all existing information relating to each is available in the Human Protein Index (HPI) data base. Via satellite, he now begins to request information from the central data base relating to the image before him. Does the pattern from this patient have spots differing in integrated absorbance (i.e., in amount) by more than two standard deviations from the data-base norms? In response to this request some spots in the pattern change color and are thus identified. Next he asks whether there are any disease-related genetic variants present in the pattern. Again, spot color changes indicate the answer. In parallel, descriptions of the interesting variants are printed out for further inspection.

Do the results obtained thus far have diagnostic or prognostic meaning? Have these changes been seen before? What are the probabilities, based on experience to date as recorded in the HPI data base? Do these suggestions match the patient's history and present complaints? Is there a single diagnosis, or must several be considered? What is the best treatment available now?

Thus the practicing physician, pathologist, or clinical chemist is directly and interactively linked with a large store of previous analytical experience and diagnostic wisdom. This, however, is insufficient. It must be possible to probe deeper, and to do so rapidly. The diagnostician puts a cursor over one of the spots, which has been identified as *the* key marker in his patient's disease, as is shown in Figure 1. The Index number appears at the bottom of the screen, together with the amount of protein in that spot in terms of integrated pixel absorbance units. Immediately, a menu (as shown in Table 1) appears on an adjacent screen. Which of the descriptors of that spots does he wish to see? He asks for a list of cell types in which the marker is found (Descriptor No. 13), and then asks for known disease correlations (Descriptor No. 19). Still not satisfied, he asks for pertinent literature references (Descriptor No. 23). From these he chooses four or five and asks for their summaries to appear. One appears to him to be a key one, and he asks for the entire text in hard copy, to be read at leisure that evening. *A black box has not told him what to do.* Rather, he has used the most sophisticated tools that modern science provides to keenly search out meaning, correlations, probabilities, ideas, and detailed information from a vast array of previous experience—all to the benefit of his patient. He has also added his own case to that total, to the advantage of all.

The toxicologist faces his cathode-ray tube display with different interests. A new and promising drug has been added to human lymphocytes in culture. A set of proteins is found to disappear from the protein pattern. Is this important, and which set is the one affected? With a cursor, he marks one of the spots, and asks for an identification from the HPI. Quickly he learns that it is a mitochondrial protein. He then asks for all mitochondrial proteins affected by known mitochondrial inhibitors (*10*, *11*) to appear on a standard reference pattern in color. Using a flicker technique, he finds that the new drug in question turns off the synthesis of exactly the same set of proteins as do dinitrophenol, nonactin, or oligomycin (*11*). Months of tedious biochemical analyses have been avoided. He now knows at least one major effect of the experimental drug: it interrupts production of mature mitochondrial proteins. Note that he may ask many additional questions concerning mitochondrial sets. Does the set affected by his drug include only proteins coded for by nuclear genes, mitochondrial genes, or both? He quickly discovers that the answer is "both." Are any of them heat-shock proteins (*12*) or proteins affected by interferon, and are any of them known enzymes? Each question is answered by alterations in the display pattern, or by statements similar to these available in the descriptor list of Table 1 and, if necessary, bits of information can be tracked down to their original sources.

To the radiation geneticist the problem is to find new protein variants in human offspring that are found in neither parent, and hence must be the product of a new mutation. The protein patterns from parents and children are compared automatically and a simple answer presented: NO, there are no new charge-variant mutations present in the offspring; or

Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory, Argonne, IL 60439.

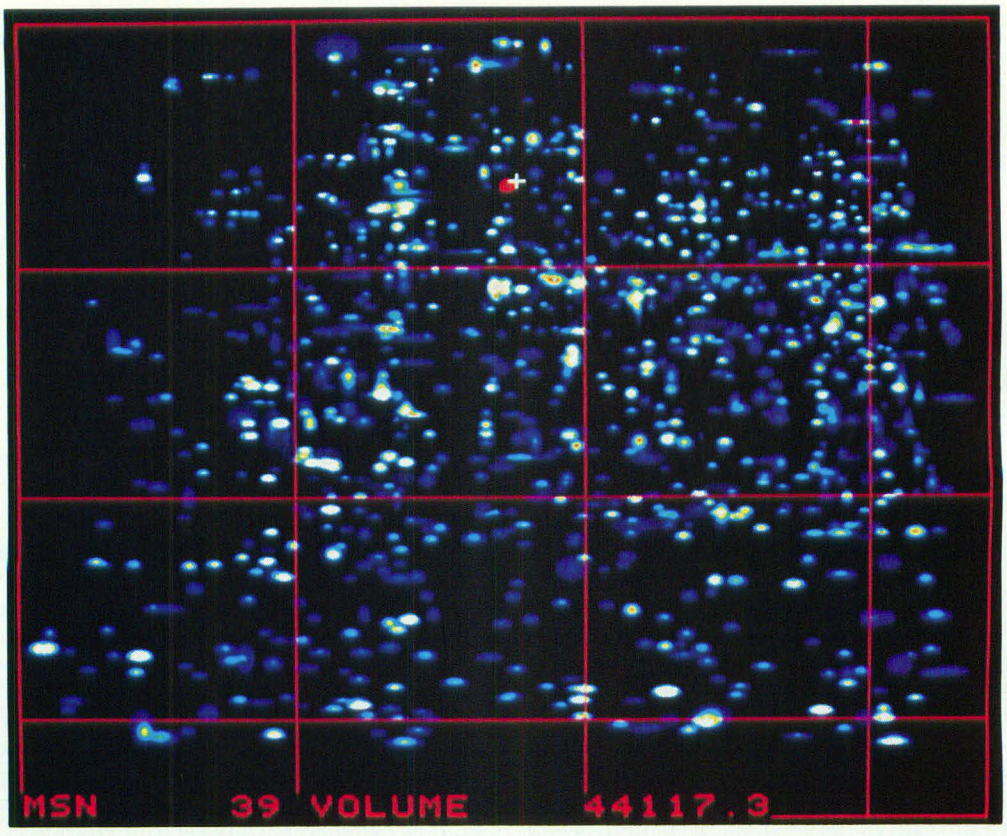

Fig. 1. Illustration of the method for accessing the Human Protein Index and data base from a pattern
Shown is a small section of a lymphocyte pattern after a cursor has been placed over one spot. The master spot number (MSN No. 38) immediately appears at the bottom of the screen, followed by the volume measurement, which is proportional to the integrated spot density (TYCHO system)

YES, there are. If there is one, the geneticist must arrange for additional samples, and for confirmatory studies (*13*) to demonstrate to his satisfaction that the variant is indeed related to the wild-type proteins seen in both parents, and is therefore a new mutation. He may also sequence the variant to find the exact amino acid substitution that has taken place. Systematically, he searches thousands of additional trios to determine the background mutation rate in man, and continues on in subsequent years to find out if radiation, chemical mutagens in the environment, or other factors are changing the human mutation rate (*14*).

The mind of the research clinical chemist is directed at a different problem. Can he find—in patterns of the proteins of plasma (*15*), urine (*16–18*), tears, lymph, seminal plasma (*19*), milk (*20*), saliva (*21*), spinal fluid (*22*), or cells and tissues (*23–26*)—new indicators or markers of disease? Sifting through patterns of urinary proteins, he discovers a spot that appears to be associated with prostatic cancer (*27*). Interacting with the HPI data base network, he asks what other proteins may occur in the same positions on his patterns, and discovers several other candidates. These he rules out one at a time in separate experiments. Only then does he know whether he has indeed found something new and worth further study and clinical exploration. Possibly he has found the basis for a new and useful screening test for cancer.

Tired of tedious and imprecise assays for lymphokines or interleukins, the experimental immunologist tests a new factor on freshly isolated lymphocytes, and discovers that some proteins in the maps displayed cease to be made after treatment, while others are present in greater abundance (*28*, *29*). How do these results compare with pattern changes caused by previously isolated factors? The new experimental results are carefully compared with patterns in the data base. Which proteins affected by the new candidate factor are affected by known lymphokines, and which are not? The immediate question is: has a new effector been discovered, an old one rediscovered, or is the effect due to a mixture of two or more in the preparation? Possibly a new name will be added to the effector list, and a new protein set to the HPI. Ultimately the question arises: has the set affected by the new factor been observed to undergo similar alterations in lymphocytes from patients with various diseases? After all, finding the underlying cause of a major human disease . . . ("Would you kindly leave me alone while I do this search; I would prefer to be the sole discoverer if you don't mind . . .").

The purpose of this meeting is to find out where we are in our attempts to make the preceding paragraphs real. We have already concluded that no insoluble technical problems stand between present achievements and the future we have described. But an extraordinarily large amount of work remains to be done. To grasp both our present position, the rate of present progress, and the magnitude of the HPI System project, it is important to review briefly the history of the work, the status of present systems, and then to outline what remains to be done.

Table 2 lists the accomplishments that appear to us to have been central to arriving at our present level of research. Such

Table 1. Partial List of Descriptors Used in the Human Protein Index Data Base

1. Human Protein Index Number (to be assigned after the majority of human cell types have been mapped).
2. Preparation-specific master number (lymphocyte, fibroblast, etc.).
3. Provisional map number(s) (derived from one type of experiment and may include a mixed cell type tissue preparation).
4. Cross indexing to other numbering systems used by other investigators.
5. Corresponding master human gene list number (assuming genes are ultimately numbered along individual chromosomes. Initial entries will be by chromosome number and possibly relative location on chromosomes.) Ultimately serves as the key to human genomic clone libraries.
6. Protein of which this protein is a genetic variant (i.e., allelic proteins). Proteins to which this protein is evolutionarily related.
7. Associated subunits, if multimeric.
8. All names in common use (where the protein has an existing name).
9. Function(s) (where known).
10. Six-letter group names and numbers (useful keys [aliases] referring to regulatory and other properties).
11. Enzyme Commission numbers (where applicable).
12. Corregulated sets to which the protein belongs.
13. Cell type(s) where it is found.
14. Subcellular localization.
15. Amount of protein present in relevant cell types.
16. Spot coordinates in standard reference map (SDS M_r and urea pI), also molecular mass and pI as determined by other methods.
17. Exact analytical data (amino acid composition, sequence etc.).
18. Post-translational modifications and pointers to associated spots.
19. All correlations with disease.
20. Source, description, and location of antibodies to this protein.
21. Sources of this protein, and location of samples in storage.
22. Position of protein in other analytical or preparative separations schemes.
23. Pertinent literature references.
24. All working notes, including responses to experimental variables.

Comments may be attached to each entry. Proteins in the index may be relisted for each investigator from his point of view, i.e., all entries may be relisted alphabetically, by molecular mass or pI, by disease correlations, by subcellular location, by number in one of the HPI numbering systems, or by numbering systems used by other investigators. This data base is not set up to search the scientific literature directly, but it includes information abstracted by specialists from it. The literature references included are key ones. Information on the number of references available in other data base systems for each enzyme may be included, and the installed operator-interface systems may include direct on-line access to existing large biomedical literature data bases allowing branching from the HPI data base directly into such literature data bases.

Table 2. Major Steps in the Development of High-Resolution Two-Dimensional Electrophoresis (Adapted from ref. *26*)

	Ref. no.
Initial experiments in electrofocusing	(*30*, *31*)
Recognized effect of sieving or molecular filtration in electrophoresis in gels	(*32*)
First two-dimensional electrophoretic separation	(*33*)
Use of acrylamide gel as an electrophoretic support medium	(*34*, *35*)
Liquid gradient systems for isoelectric focusing	(*36*, *37*, *38*)
Development of stacking-gel concept and suitable buffers	(*39*)
Synthesis of ampholytes	(*40*)
Use of two unassociated parameters for separation (mobility and molecular mass)	(*41*)
Isoelectric focusing followed by electrophoresis	(*42*)
Mapping of tissue proteins for genetic studies (IEF followed by PAGE)	(*43*)
Relationship between SDS electrophoretic mobility and molecular mass	(*44*)
Use of concentrated urea in gels and development of multiple slab-gel systems	(*45*)
Introduction of SDS stacking gels	(*46*)
Combination of IEF with SDS-PAGE	(*47*)
Electrophoresis followed by SDS-PAGE	(*48*)
IEF-SDS-PAGE of non-histone nuclear proteins	(*49*)
Acid urea electrophoresis-SDS PAGE of nuclear proteins	(*50*)
Beginning of mechanization and automation—system for casting, centrifugally and simultaneously, 500 tube gels	(*14*)
Discovery that SDS reacts rapidly with proteins in urea without heating	(*51*)
High-resolution mapping: IEF followed by SDS-PAGE	(*52–55*)
Optimization of the system for use of very small samples and autoradiography	(*52*)
High-resolution analysis of human serum	(*15*)
Development of non-equilibrium pH gradient two-dimensional electrophoresis	(*56*, *57*)
Development of the semi-automated ISO-DALT system	(*58*, *59*)
Development of internal charge standards	(*60*)
Development of internal high-resolution molecular mass standards	(*61*)
Use of monoclonal antibodies to dissect complex mixtures	(*62*)
Development of silver staining to increase sensitivity	(*63*)
Development of computerized data-reduction systems	(*3*, *26*, *64–67*)
Development of nitrocellulose transfers	(*68*)

compilations are not without bias, and we can only offer sincere apologies for any item left out.

The present status of work in two-dimensional electrophoresis is the subject of our meeting, and is recorded in the following papers. Even if we discount the effects of the great enthusiasm manifested there, it still appears that we are now seeing the publication of a variety of "founding" papers that will give rise to whole new areas of research and development; each may lead to entire symposia in less than a decade.

We initially attempted to include here a review of current research on or with two-dimensional electrophoresis, including areas outside of clinical chemistry. Space limitations, time, and the phenomenal rate of increase in publication (more than 2000 papers were recovered in a literature data-base search based on the phrase "two-dimensional electrophoresis") make this impossible.

The Future: Where the Technology Is Taking Us

Data-base technology, large disc-storage capability, and wide ranging telecommunication nets already exist. Fast scanners, image analysis, normalization and intercomparison programs, and computing systems to execute the programs rapidly also exist and are undergoing constant improvement. Many of these are being designed to lead linearly to small desk-top systems for gel-pattern analysis that can come into general use in small hospitals within a decade.

As shown by the gel-running capacity of our own small laboratory, it is not difficult to run 10 000 or more two-dimensional electrophoretic analyses per year (*58*, *59*). Within a year, we expect to be able to handle the data in all of these gels electronically in real time. We must admit, however, that two-dimensional (2D) analysis is still a virtuoso technique, with very large differences among laboratories in the quality of analyses. Part of this is due to variations in the quality of reagents, to the lack of standardization of methods, and to differences in procedures. Further, there is not yet complete agreement on how the gels should be conventionally shown, whether acid side to the right or to the left, and—as was discovered during this meeting—on whether to put the high molecular mass end of the sodium dodecyl sulfate (SDS) dimension on the top or bottom. (Thus far, we have not seen the last remaining possibility, namely, running the first dimension either up or down.) In our laboratory, we have standardized on Cartesian coordinates (*15*), as are universally used in science, and put the small numbers for pI and molecular mass in the lower left-hand corner.

Reagent Standardization

Standardization of reagents, especially SDS and ampholytes, is critical to reproducible work and must be done systematically. To allow positions of spots on gels to be intercompared between and among laboratories it is essential to have narrow-range internal standards for charge (*60*, *69*, *70*) and for molecular mass (*61*). A series of different charge standards will be required, which will include the entire pH range accessible to isoelectric focusing. In addition, it is desirable to have standards which will produce spots at several different SDS molecular mass levels on the gel. The problem of defining spot positions in non-equilibrium pH gradient electrophoresis (NEPHGE or BASO) gels (*56*, *57*) is a difficult one, because migration distance through the pH gradient depends on *both* charge and molecular mass; hence, spots may be at their proper positions in the SDS dimension in such gels, but charge standards at one SDS-mass level on the gel do not apply to those at a different SDS-mass level. To properly specify position in such gels, a range of charge standards of differing mass is required.

The question of precisely specifying the pI of each spot in a charge-standard train is also a difficult one to answer satisfactorily because of the large concentrations of urea or other denaturing agent present. (We have proposed a collaborative study with the U.S. Bureau of Standards to solve this problem.)

The objective of our program for the development of internal standards is to be able to specify map position to the limits of system resolution, and in such a way that map positions from different experiments agree, to the same resolution limits.

Spot Detection in Patterns

New, more sensitive, and specific protein detection and quantitation methods for 2D gels are urgently needed. Some improvement may be obtained by the synthesis of improved stains, including fluorescent ones, and by optimizing staining and destaining procedures. Silver staining offers a great advantage in sensitivity (*63*), and the development of methods for producing spots with different colors offers promise (*71*). Autoradiography and fluorography are the most sensitive methods currently available, but even they can be further improved: commercially available radioactivity standards and more sensitive films that have been well calibrated would be appreciated.

Methods for Protein Identification Involving Pre-Electrophoresis Procedures

Although two-dimensional electrophoretic patterns can be analyzed empirically, and spots or pattern changes can be correlated with disease without knowing the identity of any of the proteins seen, it is obvious that the value of the information obtained in these analyses increases as some function of the number of spots identified. (Our own view is that it is at least a square and possibly a cubic function.) Hence it is important to consider seriously how large numbers of spots may be identified in a systematic manner.

Protein isolation: One solution to the identification problem is to isolate every known enzyme (or non-enzyme protein) one after another and run them alone and in the presence of small samples of the starting mixture to identify the protein's position in the pattern. We have used this approach in some of our studies (*72*), as have others. It is a most tedious approach because of the effort involved and because many of the products of published procedures are far from pure. A second approach is to arrange collaborative studies with other investigators who are purifying proteins to obtain samples. This approach has been used successfully in the development of maps for *Escherichia coli* proteins (*73*), and that example should be followed with human proteins. One sometimes discovers that a purified protein is damaged during isolation, especially when the purification procedure involves heating, and consequently no spot appears in the pattern of the original mixture that corresponds to the one (or those) of the purified material.

Use of literature values for biophysical properties: In some cases, reported values for SDS molecular mass and for pI could be used for identification. One problem is the wide range of values often reported for one protein. Reported data are valuable, however, when they include data on some unusual property such as a change in apparent SDS molecular weight in the presence of urea, a missing amino acid residue, well identified post-translational modifications, unusual solubility properties, or extremes of pI. In the case of mixtures that contain relatively few proteins in high abundance, such as milk (*20*), use of literature values for identification is feasible now.

Cell selection: Some identifications derive naturally from clever sample selection. For example, one may compare a cell line known to be deficient in a given enzyme with one that has the enzyme. The difference may be the result of mutation, or viral infection, or the presence of an identified additional chromosome [for example, a single human chromosome in a Chinese hamster ovary cell (*7*)]. The problem, of course, is that more than one spot difference may be observed. In any case, however, one has a smaller set of candidate spots to choose from, as compared with the entire pattern. Actually, as the HPI is built up, confirmations of identifications should appear, because many enzymes are known that are absent from one or more cell types, or that differ greatly in concentration in different tissues.

Correlation with fractionation procedures: The amount of protein in the spot corresponding to a given enzyme should correlate with the activity of that enzyme in the sample analyzed. Hence preliminary identifications may be done by comparative analysis of corresponding spots on different gels, or of runs done after any preparative or separative procedure. One could, for example, map all of the fractions obtained when one initial sample mixture is fractionated by a variety of different methods such as ammonium sulfate precipitation, gel filtration, column isoelectric focusing, or progressive thermal denaturation. Note that in each case one starts with the whole mixture, not with fractions from previous ones. All fractions obtained are analyzed for a large number of different enzymes by using an automated enzyme analyzer such as a centrifugal analyzer (*74*). In theory, if a pure protein can be isolated by some combination of all of the methods used, then only one spot (or spots corresponding to all of the enzyme's subunits) should always correlate with activity when all of the maps of all of the fractions are compared. This approach is much too laborious for the identification of one enzyme, but is feasible when many are to be identified concurrently. In addition, the fundamental data obtained should facilitate the design of an isolation procedure for *any* spot of interest. In fact, there does not appear to be any other way to approach protein fractionation rationally. Ultimately, subcellular fractionation (*75*) should be included in these studies. A model study, in which thermal denaturation was used for identification, has been published (*76*).

Use of specific affinities: In addition to classical fractionation procedures, specific affinities of antibodies, dyes, cofactors, inhibitors, or targets of specific protein binding sites may be exploited for identification purposes. This is much too large a field to be reviewed in detail here. We merely point out that where specific antibodies are available they may be used to prepare purified protein for analysis, or to selectively remove one protein spot from a pattern. Dyes or cofactors may be immobilized to isolate all proteins reacting with one cofactor (NAD^+ for example), or vitamins, neurotransmitters, or hormones may be immobilized and used to identify the protein that binds them. In addition, where tightly or irreversibly bound radiolabeled enzyme inhibitors are available that remain attached during mapping, identifications may be done by comparing such "stained" gels with autoradiographed ones.

Manipulation of rates of protein synthesis: Gene expression may be manipulated at many different levels by drugs, hormones, heat shock, virus infection, and a variety of other methods, and the differences in rate of protein synthesis between maps of experimental and control cells can be used for identifications. This approach has already been used to identify mitochondrial proteins whose synthesis is affected by mitochondrial inhibitors (*10*) and to locate proteins whose rate of synthesis is altered by various drugs (*11*).

Protein Identification by Use of Post-Electrophoresis Analysis

A major challenge is to develop new micro-analytical methods that are suited to, and in a sense extensions of, high-resolution two-dimensional electrophoresis.

Renaturation of proteins from electrophoresis patterns: Surprisingly, some proteins can be renatured and their enzyme activity demonstrated after isoelectric focusing in urea and electrophoresis in SDS, as shown by Scheele (*77*). Adaptation of this method to histochemical methods of enzyme detection appears promising.

Specific staining: Staining for polysaccharides (*78*), phosphate residues (*79*), and other groups can contribute to identifications and should be more extensively explored.

Compositional differences: If different aliquots of cells are radiolabeled with different amino acids and all such samples are mapped, approximate amino acid compositions may be deduced by comparison with one protein in the maps for which the amino acid composition is accurately known—for example, actin. An extension of this approach for identification purposes is to discover proteins that are completely missing one amino acid (*10*, *78*). These are sufficiently rare to be unambiguously useful, when combined with map position information, for identification.

Electrophoretic transfers combined with antibodies: Entire unstained or only partly-stained protein patterns may be transferred electrophoretically to nitrocellulose (*68*) or to other suitable supports (*80*) on which the proteins renature sufficiently to bind antibodies (*68*, *19*). In our experience, about 95% of all plasma proteins renature sufficiently to bind antibodies from polyvalent rabbit antisera. Monoclonal antibodies rarely bind, however. This suggests the use of sandwich techniques in which polyvalent antisera antibodies are used to locate a protein on the immobilizing support and to bind, at the proper locations, a layer of undenatured antigen. The undenatured antigen should then bind monoclonal antibodies, and should also be enzymically active in histochemical tests. Further, if the secondarily bound enzyme is radiolabeled and is composed of subunits, only one of which is in the underlying spot, then the other subunits may be identified by eluting the radiolabeled native enzyme and analyzing it under the denaturing conditions used for high-resolution two-dimensional electrophoresis.

Use of eluted protein for immunization: Several investigators have eluted the protein from spots cored out of two-dimensional gels, for use as antigens in immunizations (B. Dunbar, personal communication). This appears to be successful in many cases, but little information is available concerning what proportions of attempts are successful. Obviously, this approach is the one of choice for the production of monoclonal antibodies.

Sequencing of eluted proteins—a link to recombinant DNA technology: Recent advances in amino acid sequencing technology allow at least partial sequencing of proteins eluted from single spots on 2D gels (*81*). This is not only important for identification of proteins, but it provides the information required to synthesize probes for the isolation of the gene for that protein as well. Once the gene for an interesting new human protein has been cloned, then it is possible to produce that protein in quantity, either for use in a sensitive immunoassay or for replacement therapy in man, should that be indicated.

Overlapping Spots

The possibility that a given spot may in fact represent two or more proteins must always be considered. The first solution to this problem is to maximize resolution, and use comput-

erized image analysis to identify and resolve overlapping gaussian or pseudo-gaussian image peaks (*3*). A more general approach was developed by O'Farrell (*52*) and can be applied to large numbers of spots simultaneously. In this method, two aliquots of the same sample are prepared (one radiolabeled), and a mixture of the two (containing a large excess of the unlabeled sample) is run on the same gel. If only one protein is present in a spot, then the spot will have the same dimensions in both stained and autoradiographed patterns, because the proteins dilute each other. If, however, the spot is made up of different proteins in the two samples, then the radiolabeled second protein in the minor sample will not be completely diluted by the major sample protein, and the minor sample spots will be smaller, and displaced relative to the stained protein spot. Note that if a spot is followed through a variety of separation methods as outlined above, and is precipitated by a monospecific antisera, and can be isolated using one monoclonal antibody, then there is a high probability that it is only one protein gene product.

The Antibody Library

A library of polyvalent and monoclonal antibodies against as many human proteins as possible is essential to the Human Protein Index project. Such antibodies are essential for re-identification of proteins previously described, for protein isolation, for the demonstration of identity of proteins in patterns obtained from different tissues, and for ongoing studies on cellular and subcellular localization. They will be especially important in studies on regulation, to find out when specific proteins appear and disappear during embryogenesis and in disease. Furthermore, such antibodies provide the principal ingredient for a widely useful immunoassay in cases where the protein in question turns out to have diagnostic significance.

The Rationale of the HPI

We believe that the research reviewed here and presented during this symposium demonstrates that it is technically feasible to construct the Human Protein Index and the data base associated with it, and systematically to explore the molecular anatomy and pathology of human cells.

The Number of Different Human Proteins

We must examine here the classical approach to biochemical analysis, in which activities or functions are first discovered, then the protein exhibiting that function, property, or structure is isolated and studied exhaustively—this cycle being repeated again and again. The implication is that if this is done long enough, then human cells will eventually be completely described in terms of the proteins they contain.

However, the sudden emergence of high-resolution methods for seeing large numbers of proteins make possible a quite different, more general, approach. Do we in fact have methods which could resolve most or all of the human proteins? How many different proteins are thought to exist in human cells at some stage in human development?

The number of known enzymes described in reports of the Enzyme Commission since 1961 is given in Table 3. Whether the present rates of increase in number will continue is not known. However, many more spots than can be accounted for by known enzymes are seen in 2D patterns.

Estimates for the number of human structural genes vary, and are given in Table 4. A reasonable estimate appears to be 50 000, but we must stress that the values in the table are only informed guesses. The problem is illustrated diagrammatically in Figure 2.

Table 3. Cumulative Estimates of the Number of Known Enzymes (from ref. *82*)

Report of the Enzyme Commission (1961)	712
Enzyme Nomenclature (*1964*)	875
Enzyme Nomenclature (*1972*)	1770
Enzyme Nomenclature plus supplement (*1975*)	1974
Enzyme Nomenclature (*1978*)	2122

Of the 2122 (in 1978) listed enzymes, 879 are reported (*83*) to have been isolated from the tissues of an animal.

For many proteins that have been carefully studied, some disease-associated alteration has been discovered. In some instances a variant of a protein *is* the cause of the disease. How many diseases may we expect to see? Estimates of the number of different human genetic diseases described to date is given in Table 5, and is dismayingly large from the point of view of contemporary clinical chemistry, but reassuringly small from the point of view of this symposium. Most mutations will probably be found to be lethal, and only a relatively few to be compatible with cell or organism survival. We hope there is not a disease corresponding to each protein!

It is difficult to obtain estimates of either the number of proteins that are causally related to disease or are secondarily and reproducibly altered in amount as a result of disease. The numbers will probably be small, however, compared with the total number of proteins present in man.

Medical Benefits

We cannot forecast the benefits that may ultimately accrue from the possession of a "parts list for man," and the technology to explore its use. Quite obviously, the search for molecular differences in cells and tissues that are linked to known genetic diseases should have a high priority. In some cases single genetic variants may be discovered that are the cause of disease symptoms. In other instances, pleomorphic effects will be observed, and many quantitative—and possibly qualitative—differences will be seen. In the latter cases the effects may result from defects in control mechanisms regulating gene expression, or may be secondary effects ascribable to one specific molecular defect, which is difficult to identify but which affects the synthesis or breakdown of other proteins. The end results of studies of genetic disease could be the discovery of a series of variants that could each become the basis of a different clinical test for genetic disease.

Table 4. Various Estimates of the Number of Human Structural Genes

Method	Estimated no.	Reference
1. From the complexity of mRNA	50 000–100 000	(*84, 85*)
2. Estimation of tolerable mutational load	30 000	(*86*)
3. Assuming same gene density as the mouse	60 000	(*87*)

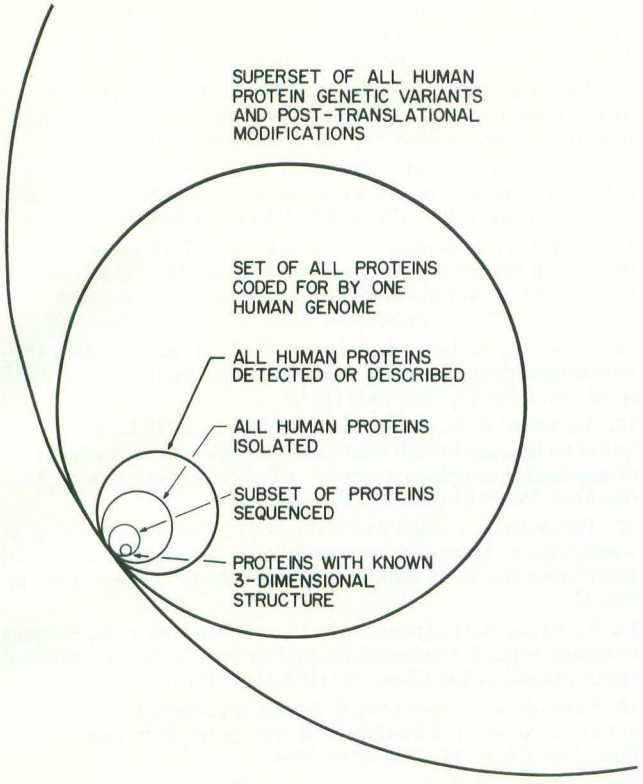

Fig. 2. Venn diagram showing relative relationships among all human proteins, including variants, all the proteins coded for by one genome, the set of human proteins thus far detected or described, the subset of proteins sequenced, and the very minor subset of human proteins for which the three-dimensional structure is known

The second area of interest is that of genetic predisposition. If several tissues and body fluids are mapped relatively early in life, and if several variants have been identified that predispose to disease later in life, then mapping followed by detailed comparison with the HPI data base should indicate diseases that may be expected to occur later. Medical counseling could then be undertaken on a more factual basis.

Intellectual Fallout

A major reason for compiling the Human Protein Index is that a new class of knowledge will emerge. It is unlikely that each gene is switched on or off separately; more probably, gene expression during embryogenesis is controlled in sets or groups. Still, we do not know with certainty the members of a single co-regulated gene set and can only suggest that large numbers of such sets exist. It is evident that we need the data required to examine this question. Such data are contained in maps prepared for many cell types, and from cells at many different stages of human development. Are there sets or batteries of proteins that always appear and disappear together? Is it possible to describe human development in terms of concatenated sets of events, a type of Markov chain, in which expression of one set is a prerequisite for switching on another? Could it be that one set includes a derepressor that turns on the next set, and may also contain a repressor that turns off a set previously expressed? In short, are the programming, sequencing, and control all built into the system so that it largely runs itself? These important and unanswered questions bear directly on the question of whether cancer, or some cancers, involve the re-expression of sets of embryonic or fetal genes, which then attempt to set in motion some part of normal development, leading to the plasticity and the so-called progression seen in human cancer.

Table 5. Cumulative Estimates of the Number of Genetic Diseases

Year	Confirmed genetic disorders	Suspected genetic disorders
1966	574	913
1968	692	853
1971	866	1010
1975	1142	1194
1978	1364	1447

Adapted from Table I (p xiv), ref. *87*.

Alternative Analytical Methods

The Human Protein Index and its application in clinical laboratory medicine is now closely tied to the only technique that provides the requisite resolution: two-dimensional electrophoresis. However, the entire project must be designed so that a smooth transition can be made to any new analytical method of higher resolution or convenience when and if such should appear. Hence the data base is designed to easily accommodate the use of alternative numbering systems and other separation parameters as this becomes necessary.

Implementation of the HPI Project

We now ask the most important questions of all: Should a serious effort be made to do the Human Protein Index project, and how should it be done (*88*)?

The answer to both questions depends in part on whether the Index and supporting systems will evolve with no direction and no special organization, funding, or effort. It is probable that, given a sufficiently long period of time—possibly half a century or so—a system to do what we have described will evolve from common unorganized effort. (There is, however, no guarantee that this is actually true.) To develop the Index and the ancillary systems described in the foreseeable future, and to reap the benefits anticipated, an integrated, standardized, and initially centralized effort will therefore be required. This is not the forum for a discussion of the organization of funding, of government vs private support, and of all possible variations on these two themes. Suffice it to say that it is quite unlikely that the project will lie fallow for any extended period, given the nature of international industrial competition.

It is necessary to mention briefly an additional problem raised by the HPI, the associated data base, and the gradual assembly of the analytical systems and staff to approach head-on many of the more difficult problems in medicine. Historically, animal experimentation, much zoological research, and a wide variety of ancillary studies have been supported in the belief that human cells and tissues could not be studied directly, that human samples were so variable that few valid conclusions could be drawn, and that more would continue to be known of animal models of disease than about human disease itself. The great triumphs of animal-model studies have been in infectious diseases, and some of these will continue. However, the notion of producing a complete and fully implemented protein index for mice, rats, dogs, and

monkeys as a prelude to human studies, and of then proceeding on to man (as some have suggested) is not feasible, because of the time and cost involved. Therefore we have aimed directly at the human protein indexing problem, and have not, in general, done the traditionally expected animal-model studies first.

If normal human cells, tissues, and body fluids are soon almost completely mapped, and if micro-biopsy samples of tissues from individuals with many different diseases are mapped, then human disease will soon begin to be systematically described at the protein level. It will no longer be acceptable to consider that an animal disease is a model of human disease unless it is shown that the molecular lesions are identical or very similar. *At this level, to demonstrate identity is to understand the cause of both.* For further therapeutic experimental studies, the validated animal model would be invaluable. But for studies on fundamental causal molecular lesions, the animal model is not needed. Hence, it is correctly seen that the HPI Project does jeopardize some established approaches to human disease.

In actual practice the implementation of the HPI project, the melding of present research efforts with clinical laboratory medicine, and the establishment of a library of antibodies to mapped human proteins will only widen the horizon of research opportunities available to the individual investigator, rather than narrow them. For the clinical chemist the prospect of being part of a new and far-reaching exploration is exciting indeed.

Work supported by the U.S. Dept. of Energy under contract No. W-31-109-ENG-38.

References

1. Anderson, N. G., and Anderson, N. L., Automatic chemistry and the Human Protein Index. *J. Autom. Chem.* **2,** 177–179 (1980).
2. Anderson, N. L., Toward a complete catalogue of human proteins. *Trends Anal. Chem.* (in press, 1981).
3. Anderson, N. L., Taylor, J., Scandora, A. E., et al., The TYCHO system for computerized analysis of two-dimensional gel protein mapping data. *Clin. Chem.* **27,** 1807–1820 (1981).
4. Giometti, C. S., Anderson, N. G., and Anderson, N. L., Muscle protein analysis. I. Development of high resolution two-dimensional electrophoresis of skeletal muscle proteins for analysis of small biopsy samples. *Clin. Chem.* **25,** 1877–1884 (1979).
5. Giometti, C. S., and Anderson, N. G., Muscle protein analysis. III. Analysis of solubilized frozen-tissue sections by two-dimensional electrophoresis. *Clin. Chem.* **27,** 1918–1921 (1981).
6. Bravo, R., Bellatin, J., and Celis, J. E., ^{35}S-Methionine labelled polypeptides from HeLa cells, coordinates, and percentage of some major polypeptides. *Cell Biol. Int. Rep.* **5,** 93–96 (1981).
7. McConkey, E. H., Identification of human gene products from hybrid cells. A new approach. *Somatic Cell Genet.* **6,** 139–147 (1980).
8. Ivarie, R. D., and O'Farrell, P. H., The glucocorticoid domain: Steroid-mediated changes in the rate of synthesis of rat hepatoma proteins. *Cell* **13,** 41–55 (1978).
9. Giometti, C. S., Willard, K. E., and Anderson, N. L., Cytoskeletal proteins from human skin fibroblasts, peripheral blood leukocytes, and a lymphoblastoid cell line compared by two-dimensional electrophoresis. *Clin. Chem.* **28,** 955–961 (1982).
10. Anderson, N. L., Identification of mitochondrial proteins and some of their precursors in two-dimensional electrophoretic maps of human cells. *Proc. Natl. Acad. Sci. USA* **78,** 2407–2411 (1981).
11. Anderson, N. L., Studies on gene expression in human lymphocytes using high-resolution two-dimensional electrophoresis. In *Electrophoresis '81*, W. de Gruyter, Berlin, 1981, pp 309–316.
12. Anderson, N. L., Giometti, C. S., Gemmell, M. A., et al., A two-dimensional electrophoretic analysis of the heat-shock-induced proteins of human cells. *Clin. Chem.* **28,** 1084–1092 (1982).
13. Giometti, C. S., and Anderson, N. L., A variant of human non-muscle tropomyosin found in fibroblasts by using two-dimensional electrophoresis. *J. Biol. Chem.* **256,** 11840–11846 (1981).
14. Neel, J. V., Anderson, N. G., and Tiffany, T. O., Approaches to monitoring human populations for mutation rates and genetic diseases. Chap. 28 in *Environmental Chemical Mutagens*, **3,** A. Hollaender, Ed., Plenum Press, New York, NY, 1973, pp 105–150.
15. Anderson, N. L., and Anderson, N. G., High resolution two-dimensional electrophoresis of human plasma proteins. *Proc. Natl. Acad. Sci. USA* **74,** 5421–5425 (1977).
16. Anderson, N. G., Anderson, N. L., Tollaksen, S. L., et al., Analytical techniques for cell fractions. XXV. Concentration and two-dimensional electrophoretic analysis of human urinary proteins. *Anal. Biochem.* **95,** 48–61 (1979).
17. Tollaksen, S. L., and Anderson, N. G., Two-dimensional electrophoresis of human urinary proteins in health and disease. In *Electrophoresis '79* B. Radola, Ed., W. de Gruyter, Berlin, 1980, pp 405–414.
18. Anderson, N. G., Anderson, N. L., and Tollaksen, S. L., Proteins in human urine. I. Concentration and analysis by two-dimensional electrophoresis. *Clin. Chem.* **25,** 1199–1210 (1979).
19. Edwards, J. J., Tollaksen, S. L., and Anderson, N. G., Proteins of human semen. I. Two-dimensional mapping of human seminal fluid. *Clin. Chem.* **27,** 1335–1340 (1981).
20. Anderson, N. G., Powers, M. T., and Tollaksen, S. L., Proteins of human milk. I. Identification of major components. Clin. Chem. **28,** 1045–1055 (1982).
21. Giometti, C. S., and Anderson, N. G., Two-dimensional electrophoresis of human saliva. In *Electrophoresis '79*, B. Radola, Ed., W. de Gruyter, Berlin, 1980, pp 395–404.
22. Merril, C. R., Switzer, R. C., and Van Keuren, M. L., Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. *Proc. Natl. Acad. Sci. USA* **76,** 4335–4339 (1979).
23. Comings, D. E., Pc1 Duarte—a common polymorphism of a human brain proteins, its relationship to depressive disease and multiple sclerosis. *Nature* (*London*) **277,** 28–32 (1979).
24. Fardeau, M., Godet-Guillain, J., Tome, F. M. S., et al., Congenital neuromuscular disorders: A critical review. In *Current Topics in Nerve and Muscle Research*, A. J. Aguayo and G. Karpati, Eds., Excerpta Medica, Amsterdam, 1977, pp 164–177.
25. Anderson, N. L., and Anderson, N. G., The potential of high resolution protein mapping as a method of monitoring the human immune system. In *Biological Relevance of Immune Suppression*, J. H. Dean and M. Padarathsingh, Eds., Van Nostrand-Reinhold Co., New York, NY, 1981 pp 136–147.
26. Anderson, N. G., and Anderson, N. L. Molecular anatomy. *Behring Inst. Mitt.* **63,** 169–210 (1979).
27. Edwards, J. J., Anderson, N. G., Tollaksen, S. L., et al., Proteins of human urine. II. Identification by two-dimensional electrophoresis of a new candidate marker for prostatic cancer. *Clin. Chem.* **28,** 160–163 (1982).
28. Willard, K. E., and Anderson, N. G., Alterations of gene expression in Novikoff hepatoma cells induced by a factor in human urine. *Biochem. Biophys. Res. Commun.* **91,** 1089–1094 (1979).
29. Willard, K. E., and Anderson, N. G., Two-dimensional analysis of human lymphocyte proteins. I. An assay for lymphocyte effectors. *Clin. Chem.* **27,** 1327–1334 (1981).
30. Kolin, A., Separation and concentration of proteins in a pH field combined with an electric field. *J. Chem. Phys.* **22,** 1628–1629 (1954).
31. Kolin, A., Isoelectric spectra and mobility spectra: A new approach to electrophoretic separation. *Proc. Natl. Acad. Sci. USA* **41,** 101 (1955).
32. Smithies, O., Zone electrophoresis in starch gels. *Biochem. J.* **61,** 629 (1955).
33. Smithies, O., and Poulik, M. D., Two-dimensional electrophoresis of serum proteins. *Nature* (*London*) **177,** 1033 (1956).

34. Raymond, S., and Weintraub, L., Acrylamide gel as a supporting medium for zone electrophoresis. *Science* **130,** 711 (1959).

35. Davis, B. J., Ornstein, L., Taleporos, P., and Koulish, S., Discussion following paper entitled, "Simultaneous preservation of intracellular morphology and enzymatic or antigenic activities in frozen tissues for high resolution histochemistry." *J. Histochem. Cytochem.* **7,** 291 (1959).

36. Svensson, H., Isoelectric fractionation analysis and characterization of ampholytes in natural pH gradients. I. The differential equation of solute concentrations at a steady state and its solution for simple cases. *Acta Chem. Scand.* **15,** 325–341 (1961).

37. Svensson, H., Isoelectric fractionation, analysis, and characterization of ampholytes in natural pH gradients. III. Description of apparatus for electrolysis in columns stabilized by density gradients and direct determination of isoelectric points. *Arch. Biochem. Biophys.*, Supplement **1,** 132–138 (1962).

38. Vesterberg, O., and Svensson, H., Isoelectric fractionation, analysis, and characterization of ampholytes in natural pH gradients. IV. Further studies on the resolving power in connection with separation of myoglobins. *Acta Chem. Scand.* **20,** 820–834 (1966).

39. Ornstein, L., Disc electrophoresis. I. Background and theory. *Ann. N.Y. Acad. Sci.* **121,** 321–349 (1964).

40. Vesterberg, O., Synthesis and isoelectric fractionation of carrier ampholytes. *Acta Chem. Scand.* **23,** 2653–2666 (1969).

41. Margolis, J., and Kendrick, K. G., Two-dimensional resolution of plasma proteins by combination of polyacrylamide disc and gradient electrophoresis. *Nature (London)* **221,** 1056–1057 (1969).

42. Dale, G., and Latner, A. L., Isoelectric focusing of serum proteins in acrylamide gels followed by electrophoresis. *Clin. Chim. Acta* **24,** 61–68 (1969).

43. Macko, V., and Stegemann, H., Mapping of potato proteins by combined isoelectric focusing and electrophoresis. *Hoppe-Seylers Z. Physiol. Chem.* **350,** 917–919 (1969).

44. Weber, K., and Osborn, M., The reliability of molecular weight determinations by dodecyl sulfate–polyacrylamide gel electrophoresis. *J. Biol. Chem.* **244,** 4406–4412 (1969).

45. Kaltschmidt, E., and Wittman, H. G., Ribosomal proteins. VII. Two dimensional polyacrylamide gel electrophoresis for fingerprinting of ribosomal proteins. *Anal. Biochem.* **36,** 401–412 (1970).

46. Laemmli, U. K., Cleavage of the structural proteins during the assembly of the head of bacteriophage T_4. *Nature (London)* **227,** 680–685 (1970).

47. Stegemann, H., Proteinfraktionierungen in Polyacrylamid und die Anwendung auf die genetische Analyse bei Pflanzen. *Angew. Chem.* **82,** 640 (1970).

48. Martini, O. H. W., and Gould, H. J., Enumeration of rabbit reticulocyte ribosomal proteins. *J. Mol. Biol.* **62,** 403–405 (1971).

49. Barrett, T., and Gould, H. J., Tissue and species specificity of non-histone chromatin proteins. *Biochim. Biophys. Acta* **294,** 165–170 (1973).

50. Orrick, L. R., Olson, M. O., and Busch, H., Comparison of nucleolar proteins of normal rat liver and Novikoff hepatoma ascites cells by two-dimensional polyacrylamide gel electrophoresis. *Proc. Natl. Acad. Sci. USA* **70,** 1316–1320 (1973).

51. Mets, L. J., and Bogorad, L., Two-dimensional polyacrylamide gel electrophoresis: An improved method for ribosomal proteins. *Anal. Biochem.* **57,** 200–210 (1974).

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

53. Klose, J., Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26,** 231–243 (1975).

54. Scheele, G. A., Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *J. Biol. Chem.* **250,** 5375–5385 (1975).

55. Iborra, G., and Buhler, J. M., Protein subunit mapping. A sensitive high resolution method. *Anal. Biochem.* **74,** 503–511 (1976).

56. O'Farrell, P. Z., Goodman, H. M., and O'Farrell, P. H., High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* **12,** 1133–1142 (1977).

57. Willard, K. E., Giometti, C. S., Anderson, N. L., et al., Analytical techniques for cell fractions. XXVI. Two-dimensional electrophoretic analysis of basic proteins using phosphatidyl choline urea solubilization. *Anal. Biochem.* **100,** 289–298 (1979).

58. Anderson, N. G., and 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).

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

60. Anderson, N. L., and Hickman, B. J., Analytical techniques for cell fractions. XXIV. Isoelectric point standards for two-dimensional electrophoresis. *Anal. Biochem.* **93,** 312–320 (1979).

61. Giometti, C. S., Anderson, N. G., Tollaksen, S. L., et al., Analytical techniques for cell fractions. XXVII. Use of heart proteins as reference standards in two-dimensional electrophoresis. *Anal. Biochem.* **102,** 47–58 (1980).

62. Pearson, T., and Anderson, N. L., Analytical techniques for cell fractionations. XXVIII. Dissection of complex antigenic mixtures using monoclonal antibodies and two-dimensional gel electrophoresis. *Anal. Biochem.* **101,** 377–386 (1980).

63. Switzer, R. C., Merril, C. R., and Shifrin, S., A highly sensitive silver stain for detecting proteins and peptides in polyacrylamide gels. *Anal. Biochem.* **98,** 231–237 (1979).

64. Lutin, W. A., Kyle, C. F., and Freeman, J. A., Quantitation of brain proteins by computer analyzed two-dimensional electrophoresis. In *Electrophoresis '78*, N. Catsimpoolas, Ed., Elsevier/North Holland, Amsterdam, 1978, pp 93–106.

65. Garrels, J. I., Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by cloned cell lines. *J. Biol. Chem.* **254,** 1971–1977 (1979).

66. Bossinger, J., Miller, M. J., Vo, K. P., et al., Quantitative analysis of two-dimensional electropherograms. *J. Biol. Chem.* **254,** 7986–7998 (1979).

67. Lipkin, L. E., and Lemkin, P. F., Data base techniques for multiple PAGE (2-D gel) analysis. *Clin. Chem.* **26,** 1403–1412 (1980).

68. Towbin, H., Staehelin, T., and Gordon, J., Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proc. Natl. Acad. Sci. USA* **76,** 4350–4354 (1979).

69. Hickman, B. J., Anderson, N. L., Willard, K. E., and Anderson, N. G., Internal charge standardization for two-dimensional electrophoresis. In *Electrophoresis '79*, B. Radola, Ed., W. de Gruyter, Berlin, 1980, pp 341–350.

70. Tollaksen, S. L., Edwards, J. J., and Anderson, N. G., The use of carbamylated charge standards for testing batches of ampholyte used in two-dimensional electrophoresis. *Electrophoresis* **2,** 155–160 (1981).

71. Sammons, D. W., Adams, L. D., and Nishizawa, E. E., Ultrasensitive silver-based color staining of polypeptides in polyacrylamide gels. *Electrophoresis* **3,** 135–141 (1981).

72. Edwards, J. J., Anderson, N. G., Nance, S. L., and Anderson, N. L., Red cell proteins. I. Two-dimensional mapping of human erythrocyte lysate proteins. *Blood* **53,** 1121–1132 (1979).

73. Bloch, P. L., Phillips, T. A., and Neidhardt, F. C., Protein identification on O'Farrell two-dimensional gels: Location of 81 *Escherichia coli* proteins. *J. Bacteriol.* **141,** 1409–1420 (1980).

74. Anderson, N. G., The development of fast analyzers. *Z. Anal. Chem.* **261,** 257–271 (1972).

75. Anderson, N. G., Ed., *The Development of Zonal Centrifuges and Ancillary Systems for Tissue Fractionation and Analyses.* National Cancer Institute Monograph, No. 21, NIH, Bethesda, MD, 1966.

76. Nance, S. L., Hickman, B. J., and Anderson, N. L., A method for studying proteins in 2-D gels using thermal denaturation analysis. In *Electrophoresis '79*, B. Radola, Ed., W. de Gruyter, Berlin, 1980 pp 351–360.

77. Scheele, G., Pash, J., and Bieger, W., Identification of proteins according to biological activity following separation by two-dimensional isoelectric focusing/sodium dodecyl sulfate gel electrophoresis: Analysis of human exocrine pancreatic proteins. *Anal. Biochem.* **112,** 304–313 (1981).

78. Anderson, N. L., Edwards, J. J., Giometti, C. S., et al., High-resolution two-dimensional electrophoretic mapping of human proteins. In *Electrophoresis '79*, B. J. Radola, Ed., Walter de Gruyter, Berlin, 1980, pp 313–328.

79. Green, M. R., and Pastewka, J. V., Characterization of major milk proteins from BALB/c and C_3H mice. *J. Dairy Sci.* **59,** 207–215 (1975).

80. Alwine, J. C., Kemp, D., and Stark, G., Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA* **74,** 5350–5354 (1977).

81. Hunkapiller, M. W., and Hood, L. E., New protein sequenator with increased sensitivity. *Science* **207,** 523–525 (1980).

82. *Enzyme Nomenclature 1978. Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes*, Academic Press, New York, NY, 1979, p 5.

83. Dixon, M., and Webb, E. C., *Enzymes*, Longman Group Ltd., London, 1979, pp 683–974.

84. O'Brien, S. J., On estimating functional gene number in eukaryotes. *Nature (London) New Biol.* **242,** 52–54 (1973).

85. Bishop, J. O., The numbers game. *Cell* **2,** 81–86 (1974).

86. Ohta, T., and Kimura, M., Functional organization of genetic material as a product of molecular evolution. *Nature (London)* **233,** 118–119 (1971).

87. McKusick, J. A., *Mendelian Inheritance in Man*, 5th ed., Johns Hopkins University Press, Baltimore, MD, 1978, p xiv.

88. *Report of the Human Protein Index Task Force*, N. G. Anderson, Chairman, 1980, available from Dr. R. E. Stevenson, American Type Culture Collection, Rockville, MD 20852.