N. G. ANDERSON, N. L. ANDERSON

The Human Protein Index Project and the Molecular Pathology Data Base

Although numerous proteins, the major functional units of human cells, have been identified, isolated, its structure determined, and changes in either structure, location, or amount correlated with disease, less than 2% of human proteins appear to have been described. The possibility of preparing a comprehensive Human Protein Index (HPI) by isolating one protein at a time by classical means is rather small. However, using high resolution two-dimensional gel separation techniques, it is possible to resolve 100 to 150 proteins in various human cells and fluids in health and disease in one analysis. Separation in the first dimension depends on isoelectric focusing whereby proteins separate on the basis of their isoelectric points. In the second dimension, sodium dodecylsulphate electrophoresis separates proteins on the basis of their molecular weights. The two-dimensional maps can be displayed on high resolution colour cathode ray tubes (CRTs), providing details for the Human Protein Index Data Base and the Pathology Data Base. It is the first attempt to provide a mechanism for organizing a large mass of knowledge relating to cell function and molecular biology in an accessible and interactive form. It is expected to find markers for single immunospecific clinical tests. With the use of these markers and the enlarging Index Data Base complete analysis of complex protein mixtures will become possible.

The most complex physical system thus far discovered in the known physical universe is man himself. The problem of his complete description would therefore appear to defy solution, and the prudent course of action, especially at the molecular level, has been to attempt to break larger problems down into discrete soluble questions and problems, and to solve each in turn. Thus at the level of proteins, one protein after another has been identified, isolated, its structure determined, and changes in either structure, location, or amount correlated with disease. Given sufficient time, on overwhelming mass of data on a large fraction, and ultimately on all, human proteins might accrue. The accumulated massive body of information would be spread through tens of thousands of journal articles. Integrating from it useful concepts and generalizations would be an impossibly difficult task. In fact, the literature of biochemistry is now unmanageably large, and few can comprehend more than a very small fraction of it.

As we have previously discussed[1], two classically different approaches have been taken to the problem of understanding the structure and function of man, and both have been limited by available techniques. The first, and most classical one, based on an hierarchical series of increasingly refined dissections, started with gross anatomy and proceeded with the description of smaller and smaller parts, as methods for seeing those parts became available. The end result of this progression has been ultrastructural analysis which has aimed at (and is actually close to achieving) a description of all different cell types found in man at all stages of development. *Completeness* in this effort is more than an academic predisposition. All

other branches of science have realized very early the fundamental necessity for classifying, listing, or indexing *all* of the units of each discipline. Thus from the very beginning of astronomy there have been star charts and atlases, and refined versions of these continue to be made. Without these maps, celestial objects could not be located, and work on the classification and evolution of stars and other objects would not be feasible. Similarly, chemistry relies on the atomic table, zoology on the classification system of Linneaus, and a central theme of nuclear physics is the discovery and characterization of as many subnuclear particles as possible, with the hope of finding all that either preexist, or can be made.

The driving forces behind these continuing exercises in separation or delineation, identification, mapping, and classification are not the sterile interests of taxonomists, lexicographers and »handbuch« editors. Fundamental theory in science always involves postulates regarding the interactions and interrelationships of fundamental units. Generalizations emerge which apply to many elements in the superset of units fundamental to each discipline. These generalizations or theories are destroyed or must be modified after the discovery of exceptions. This is true in nuclear physics, in chemistry, in the study of evolution, and in astronomy. To point up this fact, it is sufficient to note that it is a simple task to describe an hypothetical subnuclear particle, a chemical element, an animal, or a star which would destroy a present theory (examples would be a vertebrate with a ventral nerve chord, and element with a valence of ten, or a blue star of earth dimensions, etc).

It is difficult to do the same for the major functional units of human cells (proteins) because less than 2% of them appear to have been described[2], and hence there exists insufficient knowledge on which to develop predictive theories of cell function, of the groupings of proteins, and of cellular organization and development. This lack of underlying theory in cell biology is so complete that its lack is rarely noticed except by observers from other disciplines.

In fact, the view has been expressed that living cells are too complex to allow an underlying body of theory to be developed by human minds.

We therefore address two questions in this lecture. The first is whether the information required to formulate a molecular anatomy of human cells, to write a Human Protein Index (HPI), and to construct the data base of information on which to found a human molecular pathology can in fact be organized, managed, and used in actual practice. The second question relates to technical feasibility, to the question of whether the work required to provide the detailed content of the HPI can actually be done.

Given the very large existing literature on one human protein such as hemoglobin, and the large number of proteins thought to exist, the scope of the problem can be grasped by simply multiplying pages of hemo-

globin scientific literature by the number of estimated proteins to see that a new solution to the current literature explosion (or detonation) is now required. The question of technical feasibility is therefore not separate from that of data management. In fact the analytical systems devised must interface naturally, and be designed in parallel with, the data reductions systems developed, i.e. new data must go directly into the new data base system.

We address the question of data management first, and leave the problem of data generation to a subsequent section.

Lest the listening clinical laboratory scientist think, perchance, that we have strayed from our subject, we note that if indeed clinical laboratory medicine is concerned with man, then it ultimately must deal with all molecular (chemical) species which may be causally or diagnostically related to disease. With respect to the functional machinery of cells, which is almost all protein in nature, this science has hardly been founded.

The TYCHO System

Only electronic data storage and processing can offer a solution to the problem of storing, searching, and making accessible very large amounts of data. While the data may be organized and searched from many different viewpoints, some pivotal mode of access must be provided, especially if the information is to be readily available to the ultimate user, such as the pathologist or the clinical chemist. Typed lists are not a convenient mode of display when several thousand entities must be searched and intercompared. The human mind is better able to deal with two-dimensional maps where position in each dimensions conveys information, and where the size or color spots on the map confers information on amount, or indicates some unique property. Such maps including one or more thousand proteins can be conveniently displayed on high resolution color cathode ray tubes (CRTs) such as is illustrated in Figure 1. Electronically such maps may be compared with large numbers of other maps, and the differences marked either by color, or by intermittent display of a spot (i.e., by flickering it).

For proteins, the parameters used to construct the two-dimensional map, and which define the map coordinates, should be based on properties which are unrelated, and which can be actually used in practice to make high resolution separations. The two highest resolution separation methods currently available, and which resolve proteins as subunits where they are multimeric, are isoelectric focusing in the presence of urea and detergents to separate on the basis of isoelectric point, and acrylamide gel electrophoresis in the present of sodium dodecyl sulfate to separate the proteins on the basis of molecular mass. These techniques are used in the mapping systems described here. Since the human eye is more sensitive to differences in color than to differences in intensity, shades of gray (or of one color)

can be changed into different colors, and spots presented in »pseudocolor« to more easily allow differences in amount to be estimated and intercompared visually.

Thus electronically processed two-dimensional color video maps[3] provide the key to the data management problem in several ways. They may be used for intercomparison, learning, and ultimately as a means of searching the sum total of available knowledge related to a study being done by an investigator on a tissue or cell type. The operator may ask for the major mitochondrial proteins in a map of human lymphocytes to be color marked, as is shown in Fig. 2, the proteins which are markedly diminished by exposure to a calcium ionophore (Fig. 3), or affected by phorbol esters (Fig. 4), or by an interferon (Fig. 5) to be highlighted. Further, he may ask for those proteins which are common to two affected protein sets (i.e.,

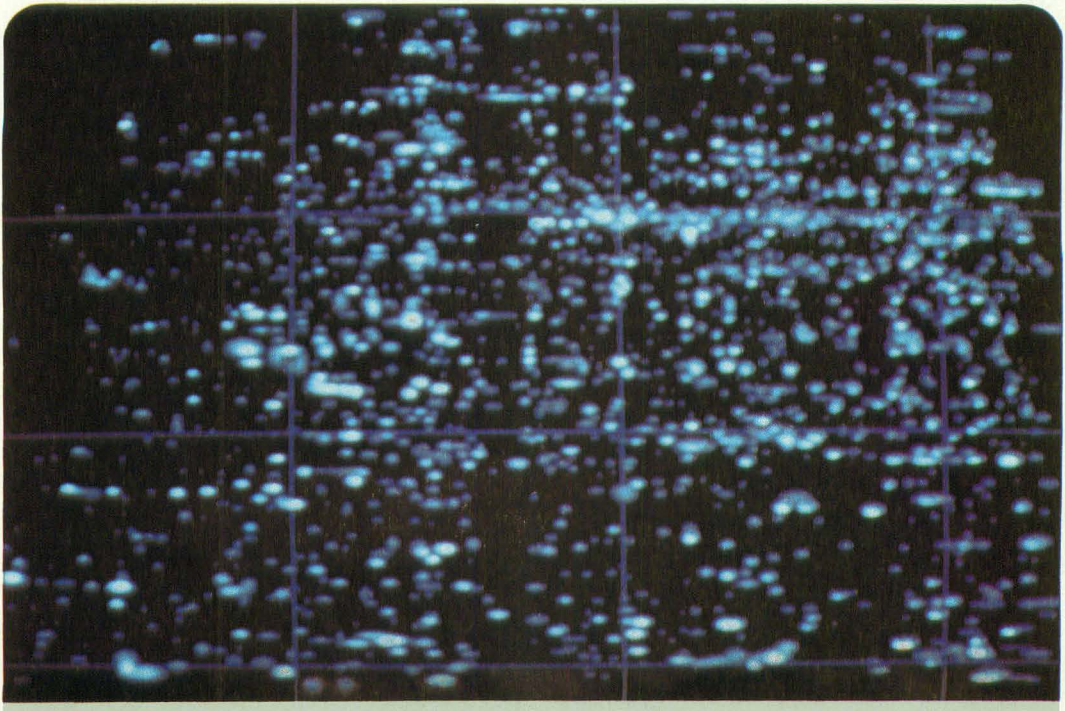

Fig. 1 Two-dimensional electrophoresis pattern of proteins of human lymphocytes processed by the TYCHO image analysis system, and displayed on a color CRT. (From J. Amer. Med. Assoc. 246: 2621, 1981, with permission).

the intersection of two sets) to be displayed (Fig. 6). These are ways of asking whether changes seen in an experimental pattern have been seen before, and can be at least partially explained. Note that the TYCHO system contains routines for plotting information relating to each of many different spots from different analyses (in an experimental series), as a function of time or some experimental variable, and for doing statistical analyses on the results obtained.

In this way results from many different samples, from many patients, from diagnostic or experimental studies, and from the same individuals as a function of time may be stored, intercompared, and analyzed.

The result common to all analyses, and accessible to the diagnostician or research investigator, is a two-dimensional pattern, and a variety of studies can lead to its acquisition. Thus results from many different

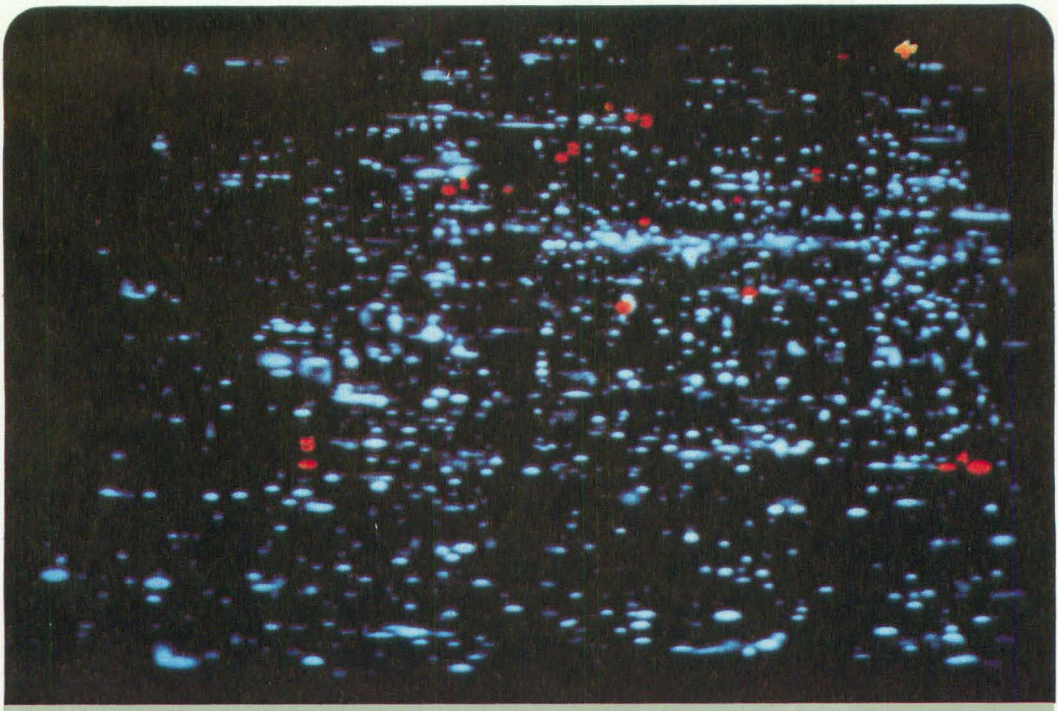

Fig. 2 Two-dimensional pattern of human lymphocytes with major mitochondrial proteins described in reference (20) highlighted in red.

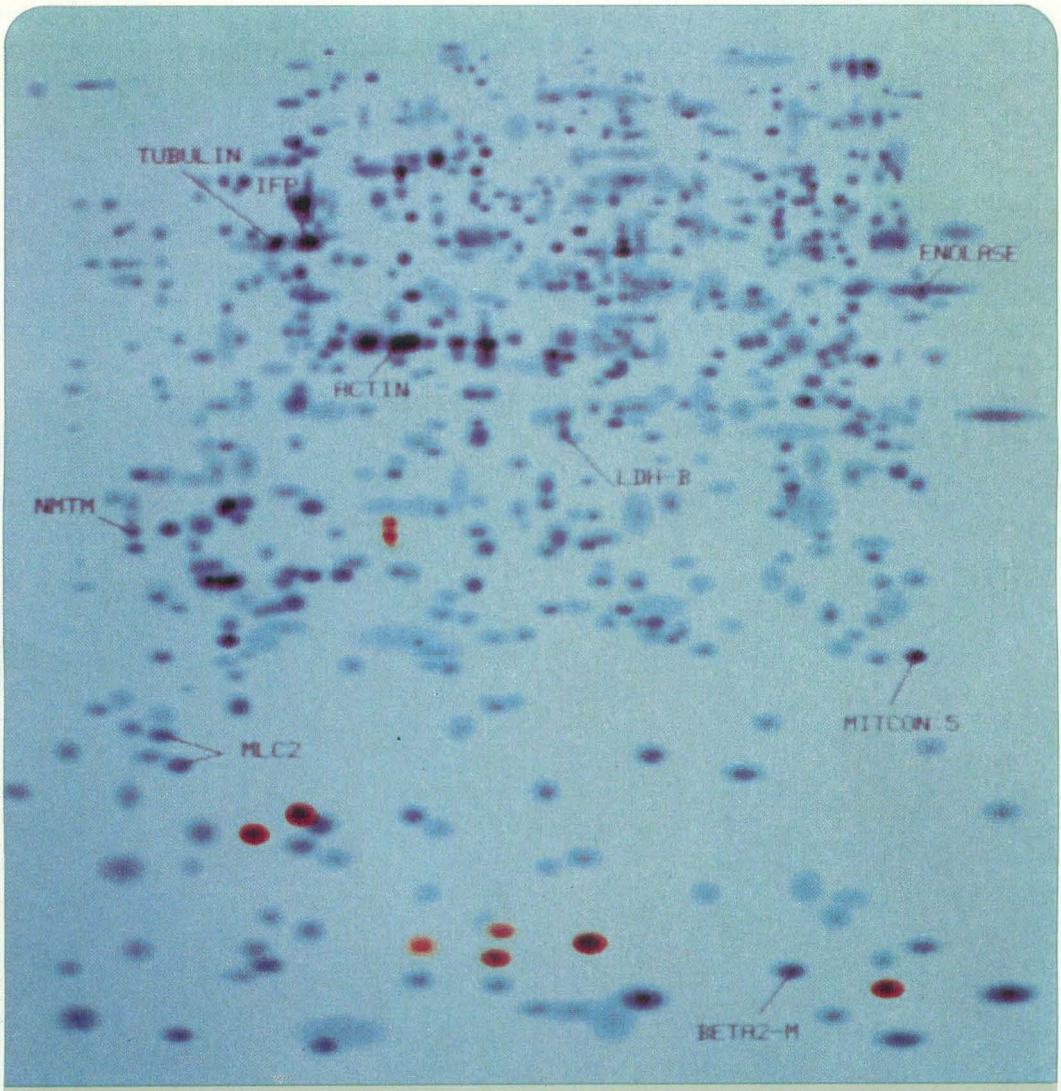

Fig. 3 Proteins in a mixture of peripheral human blood lymphocytes affected in a major way by the calcium ionophore A 23187. Note that a different color lookup table has been used to convert shades of gray to color, as compared with Figs. 1 and 2.

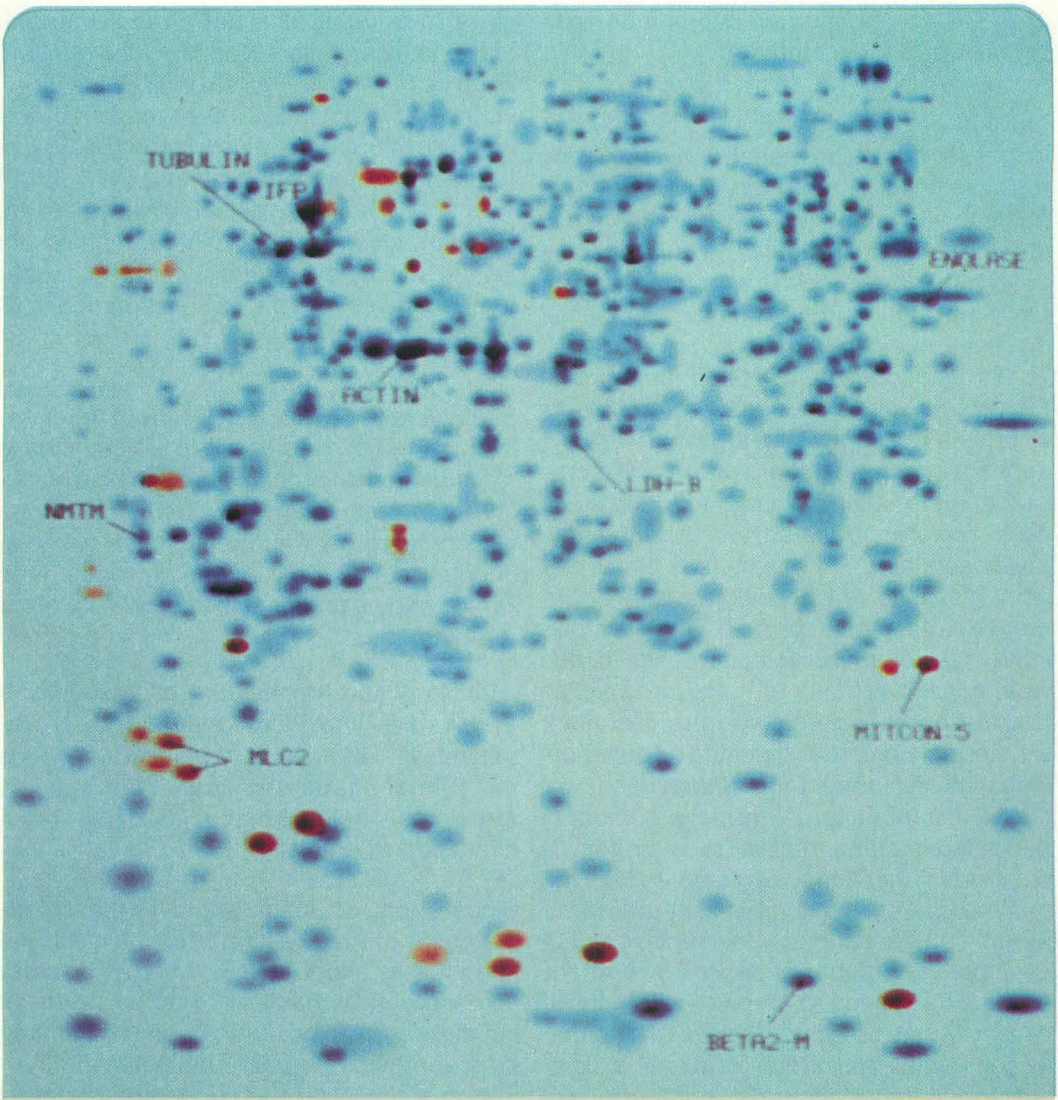

Fig. 4 Proteins in a mixture of human peripheral blood lymphocytes affected in a major way by phorbol esters.

analyses may be merged into one or a few patterns and intercompared.

The pattern also provides a mode of access to masses of data previously gained. Taking a simple case, one protein may have been found to alter in a major way in response to an experimental variable, or to always be associated in some way with one disease. Once the marker protein has been identified in a pattern, a cursor may be moved over it. The protein index number then appears on the screen[3], and a menu of search options or protein descriptors as shown in Table 1 appears on an adjacent screen. One may ask for information from memory relative to each protein displayed which relates to each descriptor. For example, one may ask what diseases are associated with an increase of the marker protein, what its molecular mass or amino acid sequence is, and what genetic variants of it are known. The information associated with the descriptor list, and linked to it, is digested information. One may ask further for a search of original literature information. Where the number of published papers relative to a protein is small, all of these may be listed. Where the number is large, the number of papers in different categories may be displayed. For example the number of papers by year, by disease, or related to one type of biophysical study may be shown, so that the operator can restrict the number of titles to be displayed. Of these, a smaller number of abstracts may be selected, displayed, and read. Ultimately, the operator may ask for complete hard copies of the research papers to be sent him, to be read at a later date. Thus access will be provided from the map to the Human Protein Index Data Base, to the Pathology Data Base, and to the relevant literature data base for each individual protein or for sets of proteins. Only part of this entire system exists in prototype form, developed for feasibility studies carried out at the Argonne National Laboratory under the auspices of the U.S. Department of Energy.

This represents the first attempt of which we are aware to provide a mechanism for organizing a large mass of knowledge relating to cell function and molecular biology in an accessible and interactive form, in a form which can grow with the advance of knowledge and of technology, and in a form which begins to match the complexity of the problem.

Details of the architecture of the systems required for the Index and data base problems will be described elsewhere. The point to be made here is that no insoluble problems in implementation have been identified to date.

Thus spots on maps stand at the crossroads. Spots can be related on the one hand to disease, injury, aging, therapy, or other experimental variables in a series of current studies, and on the other, by simply pointing to a spot with a cursor and indicating from the descriptor list the class of information desired, the investigator can interrogate the accreting body of biochemical knowledge. Integration of both experimental and dia-

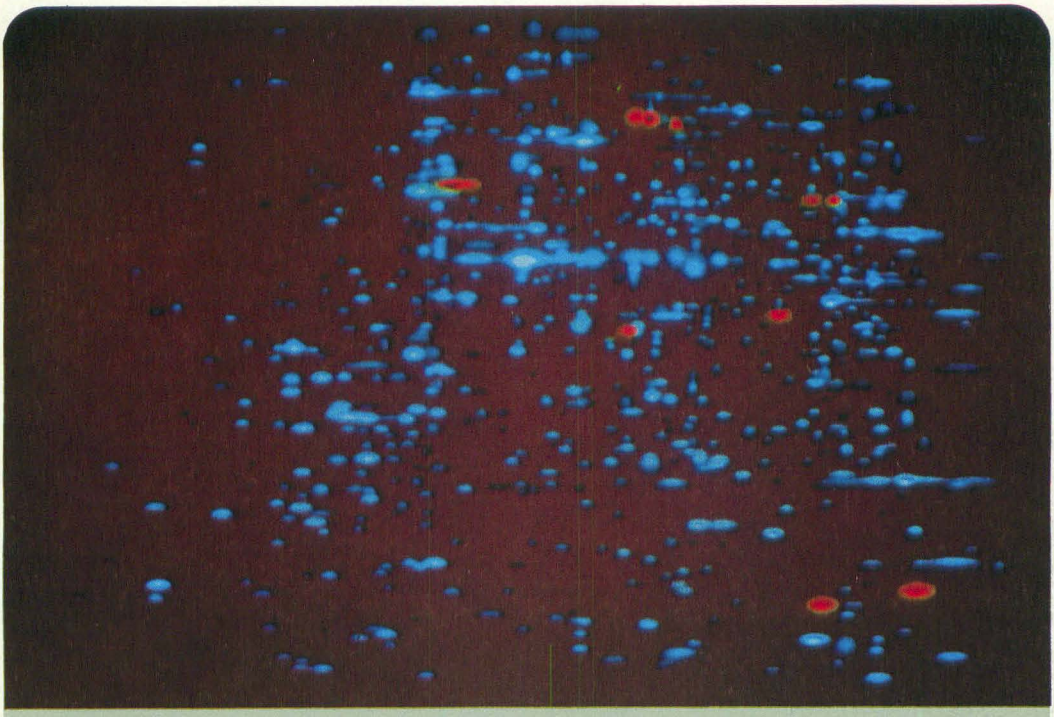

Fig. 5 Proteins in a mixture of human peripheral blood lymphocytes affected by treatment with interferon.

gnostic results can be through individual spots (proteins), or sets of spots. We know of no other way of allowing physicians, biochemists, or clinical laboratory scientists to interact with a wide range of data, to extract information from that data, to search the scientific literature in both its processed and original forms, and to use the information obtained interactively to formulate decisions and to draw conclusions.

The Number of Human Proteins

No accurate count of the number of different human proteins exists. The estimates which are available range from 50.000 to 100.000 based on the complexity of mRNA[4, 5], 60.000 based on the assumption that human DNA has the same gene density as the mouse[6], to 30.000 from estimates of the tolerable mutational load[7]. These estimates

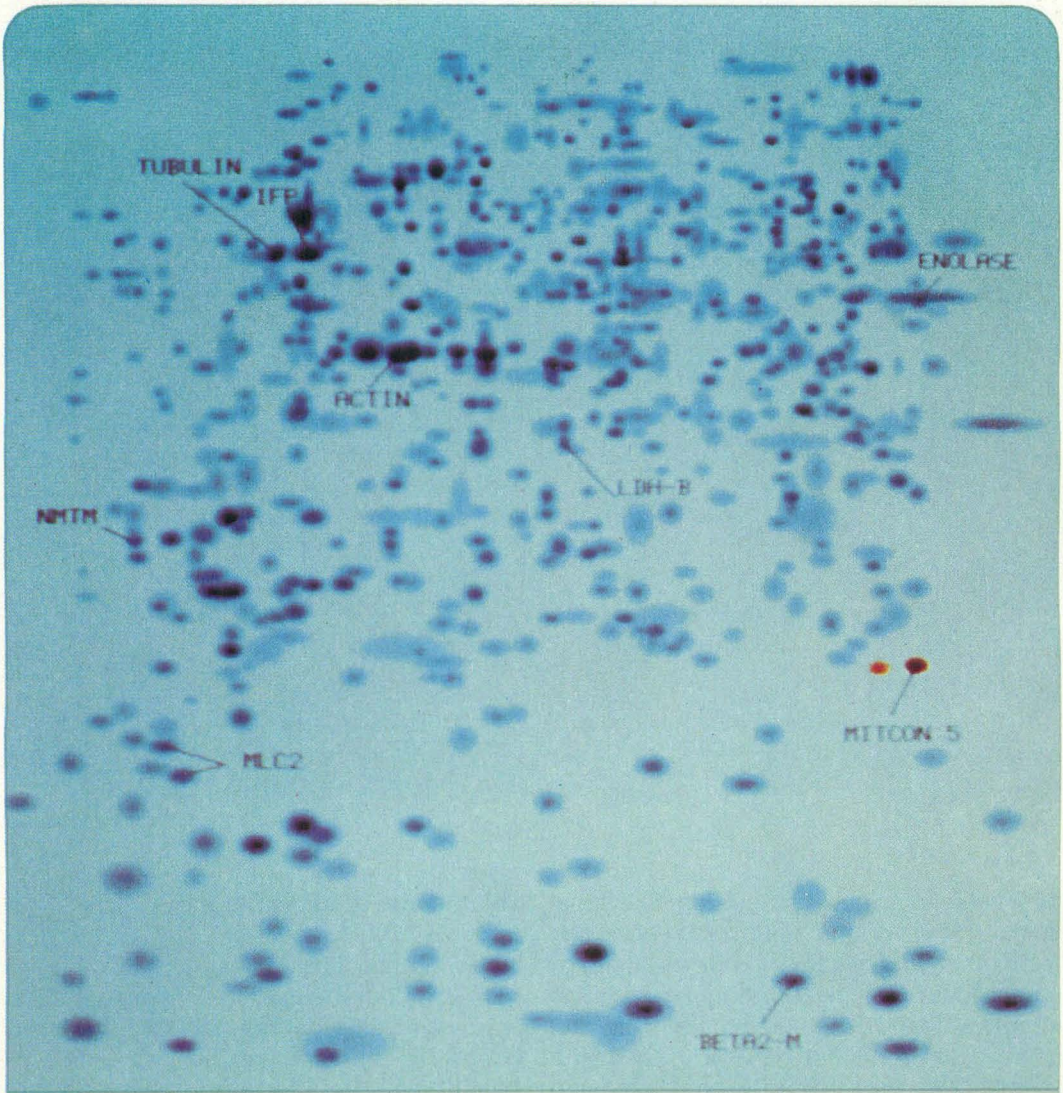

Fig. 6 Illustration of display of proteins which are members of two different groups, i.e., are the intersection of two sets. In this case the sets are the major mitochondrial proteins of peripheral blood lymphocytes, and the set of proteins affected by phorbol esters. Cells are peripheral blood lymphocytes.

do not include estimates of the total number of different antibody molecules which might be produced by splicing existing gene segments to produce new somatic antibody genes limited to single clones. Also not included are all of the post-translational modifications which occur during intracellular processing, heteropolysaccharide addition (as a marker for export from the cell), sialation, phosphorylation or sulfation. Fortunately extensive modification, which is characteristic of the proteins of body fluids[8] and cell surface proteins, occurs much less frequently in intracellular proteins.

The total number of proteins to be dealt with is very large relative to the resolution of classical separations and analytical systems such as ultracentrifugation, moving boundary electrophoresis, ion exchange chromatography, gel filtration, and ammonium sulfate precipitation. It is also large relative to the number of proteins studies clinically in the average patient. The total number of proteins is *not* large, however, relative to the number of identified, discrete, numbered, and available parts of a large jet aircraft. Nor are the data base ultimately required, the communications links, display, access, image analysis, and data processing systems needed large relative to those now in existence for business, research in chemistry and physics, the census, internal revenue, weather forcasting, or military purposes.

An additional complicating factor is the presence in the human population of numerous genetic variants of many proteins. In fact, the criticism has been made that genetic variability will make the Human Protein Index Project infeasible. Earlier studies on the degree of genetic polymorphism in man were largely based on the study of plasma proteins. It now appears that the structural variability compatible with survival is much greater for plasma proteins than for intracellular proteins. Indeed, recent studies on human cells suggest that when the cells from two unrelated individuals are intercompared by high-resolution two-dimensional electrophoresis, 1% or less of the proteins will be found to occupy different positions in the intercompared patterns[9, 10]. These differences are almost invariable differences in pI, and one charge difference between a wild type protein and a variant is readily observed in all but very high molecular weight proteins. In approximately one third of amino acid substitutions due to point mutations, a charge change occurs. Hence approximately one third of all genetic variants should be detected using 2-D electrophoresis. Where more than one amino acid is substituted, or where more than one amino acid is added or deleted in a peptide chain, a larger fraction of mutants would be detected.

Experimentally, one of us (N.L.A.) has found a number of intracellular proteins to be highly conserved during evolution, and to occupy the same positions in a very wide range of vertebrates. Thus we do not expect to see an equal number of variants of all proteins, and may, for some proteins, find none.

An additional simplifying factor is the selectivity of gene expression of single cell types. It is estimated that 10% or less of all structural genes in one human genome are expressed in one cell type, making the analysis of a uniform cell preparation, or even a mixed cell tissue, a simpler task than separating the totality of protein gene products from one preparation.

We cannot leave the problem of the relationship of the estimated numbers of human proteins to the structure of present research without dealing with the view often expressed by individuals who isolate proteins one after the other by classical methods, that cells contain too large a mixture of proteins for all to be resolved into discrete and different protein species. If the whole cannot be resolved, how can it be that single protein species can be plucked out one at a time.

A number of answers suggest themselves. The first, the most obvious, and least acceptable answer is that few if any proteins have been actually isolated in pure form, and that the low resolution of our analytical methods has prevented us from observing the heterogeneity actually present. A second answer is that the proteins chosen for isolation are those which are the most abundant and most easily isolated because of some unique properties. Hence fractionating the remainder may still present insoluble problems.

A third answer is that by using in sequence a series of separations methods (for example »n« different methods) each based on a different parameter, and each able to resolve up to ten proteins, then a theoretical resolution of 10^n could be achieved. Hence five such methods in series might give a theoretical resolution of 100.000 proteins. It does not appear that we indeed possess five such methods based on five unique and different parameters. Rather it appears that we possess a number of methods of rather low resolution and only two methods with high ($>$ 100 proteins resolved) resolution. Low resolution methods in series may be successful for preparative purposes, but it is difficult to combine them in such a way that they can be used analytically for all components resolved.

It may be that all of the answers suggested above are partially true. It is also probable that with some methods, especially those involving precipitation, that when fractionation has proceeded to the point where the desired protein is the most abundant one, minor contaminants are lost in succeeding steps at a disproportionate rate. This is true with precipitation, for example, because the percentage of a protein precipitated is usually some function of its concentration.

It is difficult to escape the conclusion that the possibility of preparing a comprehensive Human Protein Index including the majority, and ultimately all human proteins by isolating them one at a time by classical means is vanishingly small, and would probably require more than a century of work. Note that less than 1.000 enzymes have

thus far been characterized from animal tissues, and that only a small fraction of these were obtained from man. Hence less than 2% of human proteins appear to have been described, and these are most often those with easily identified functions, which provide assays to guide isolation. Clearly new approaches are required to solve the Human Protein Index Problem.

To assemble the resources required to do feasibility studies for the HPI on the proper scale, it was important to assess possible benefits, noting both the scientific and the practical advantages in so far as they can be seen.

New Knowledge

Quite obviously large amounts of new information will be, and can be obtained relative to each new protein by available methods and by new methods now under development. Ultimately this information will be stored in, and retrieved through, systems such as the one described here. It is less obvious, however, that new previously unexpected and now unobtainable types of knowledge will result from the global nature of the information obtained. Since it is our view that the HPI and associated systems will eventually yield new types of information which will be of great importance, we speculate here briefly on their possible natures.

Proteins are the means for effecting the plans stored in DNA, and very little that goes on in cells is mediated other than through proteins. Imprisoned in a transparent nuclear globe, DNA can only govern indirectly, and express its plans by making or not making specific proteins. (To a very much lesser extent some cell functions are mediated directly by specific RNAs.) Mitochondria represent a simpler and only partially self-contained system with a small amount of DNA in each mitochondrion coding for only a fraction of mitochondrial proteins. The concept that genes are switched on and off in sets, batteries, or groups has much experimental data to support it; however, for man we do not know for certain all of the members of a single set. We do not know how gene set expression is sequenced during development, and whether one set can be the cause of the expression of the next in order to appear, and whether a set in an interconcatenated chain of sets can turn off previous ones in a series. We also do not know whether elements of one set can belong to another, i.e., whether sets are separate, unique, and independent, or whether they intersect and can have common members. Quite obviously we have insufficient information to be able to write out the sequence of set expression which, in a sense, is the program of human development. Further, we cannot grasp the details of how the organization of gene expression changes in cancer, although theories are not lakking.

In addition, we cannot as yet define cells as being comprised of sets of sets, and thus know what sets (remembering that a set

may have only one member, i.e., one protein) are characteristic of a germ layer, a specific phase of development, or a specific cell type.

Further, we cannot now systematically group proteins by intracellular location, function, chromosomal location of corresponding genes, or evolutionary relationships, although many individual details are known.

More importantly, we cannot, without a wealth of additional information, understand how cells work, and delineate the control circuits underlying cellular responses to experimental variables. These variables may include those which cells may ordinarily experience such as nutritional status changes, diurnal rhythms, exposure to hormones, or changes in oxygen tension or pH. Or they may involve responses to drugs, carcinogens, or toxic agents. It is not just that it is important to see as many of the alterations produced by one agent as possible, it is that only by a global analysis is it possible to attribute similar properties to different agents. For example, if one is attempting to produce a drug mimicking interferon, it is important to show that it affects exactly the set of proteins that interferon affects. Similarly, if a drug affects one protein in a desired way, but as a side reaction produces in addition the effects seen with a tumor promoter such as a phorbol ester, then the investigator would be well advised to delay human trials until much further study.

It is, in fact, difficult to see how pharmacology, toxicology, and a large fraction of pathology can be considered to be at grips with the real world of human cells without the possibility of a global analysis of cellular molecular responses. The knowledge gained in studying such responses will, in our view, give rise to several new medical specialities and disciplines, whose content of detailed and useful knowledge will be large indeed.

We conclude that the information now required to build the Human Protein Index and associated data bases cannot be gained and integrated in future years through totally uncoordinated efforts, with bits of data scattered widely in the literature, and that a new central data management and interactive access system will be required to make the Index and data base information useful to physicians and clinical laboratory scientists.

Present Status

The Molecular Anatomy Program has been concerned with feasibility studies to determine whether there are any parts of the model for the integration of pathology and biochemistry outlined above which contained problems which are not now soluble, and to carry out a small demonstration project to show that the model is a feasible one, albeit on a very small scale. This work divides into three interrelated areas which are first, the development of methods for doing large number of analyses in a reproducible manner, second, the exploration of the use of the

system developed to map human body fluids and a limited number of cell types; and third, the development of image analysis, data reduction, and data base systems to acquire, store, intercompare, and manage the large amounts of data obtained and to convert it into useful information.

High-Resolution Two-Dimensional Electrophoresis

The work we are describing here is based on high-resolution two-dimensional electrophoresis as developed by O'Farrell[11]. The history of the development of this technique has been reviewed briefly elsewhere[1, 2]. Two points deserve emphasis. The first is that two-dimensional methods, derived in concept from the very early combination of electrophoresis and chromatography on paper, should, in theory, yield a resolution which is the product of that of each separation individually. Isoelectric focusing in urea and electrophoresis in the presence of sodium dodecyl sulfate are the two highest resolution separation systems currently available, each being capable of resolving in one analysis of a suitable preparation 100 to 150 proteins. In combination the theoretical resolution should be over 10.000 proteins or protein subunits.

The second important point is that the Human Protein Index and data base are designed to be *independent* of two-dimensional electrophoresis in the sense that if better, higher resolution analytical systems, either single or multidimensional, become available, all information can be readily transferred to the new analytical system, and all data keyed to it. This flexibility is essential to future progress.

High-resolution two-dimensional electrophoresis is an exploding experimental field. A review of the studies in which it has been employed is unfortunately already beyond the range of this discussion.

The ISO-DALT System

We have adapted the general concept of multiple-parallel analysis which we exploited previously in the centrifugal fast analyzer[12] to the problem of high-resolution two-dimensional electrophoresis, and have developed rather simple systems[13, 14] which allow us to run up to 100 analyses per day, with a total of over 45.000 during the course of this work. Present efforts are directed toward the automation of these devices. However, the systems in their present form have been adequate for initial exploratory studies on human samples.

Initial Mapping Studies

An updated map of human plasma proteins is shown in Fig. 7. The positions of the majority of the proteins identified have been confirmed by transferring the entire patterns to nitrocellulose, and reacting the pattern with specific antibodies[15]. Despite the fact that the 2-D separation involves denaturing

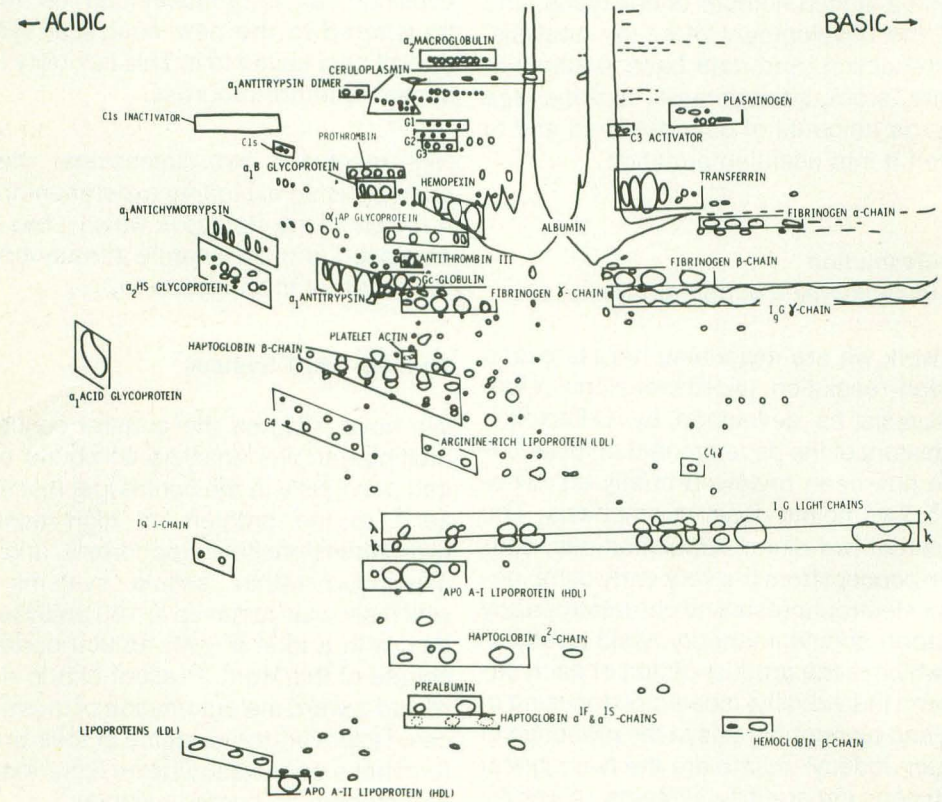

Fig. 7 Updated diagram of the major human plasma proteins seen on two-dimensional electrophoretic patterns using the ISO-DALT system. (From reference 15, with permission). Variation in the number of sialic acid groups added to a protein account for the multiplicity of spots shown for one protein species in the horizontal direction, while variations in mass due to large variations in the amount of heteropolysaccharide added account for the vertical displacements. Proteins of one type (for example α_1-antichymotrypsin) are evidently synthesized with wide variations in the amount of carbohydrate added. The amount of sialic acid added to carbohydrate residues appears to be roughly proportional to the amount of carbohydrate present on any individual protein molecule. Hence molecules which are heavier because they bear more sugars would also tend to have more acid charges in the form of sialic acid residues. The net result is the formation of diagonal charge shift trains which almost invariably tilt up and to the left, as shown.

conditions, over 95% of the proteins we have examined renature sufficiently on nitrocellulose transfers to react with polyvalent antisera. This is not the case with monoclonal antibodies; in our experience these rarely react to a detectable extent[15]. We have shown how the microheterogeneity of serum transferrin, haptoglobin, and a_2HS glycoprotein can be resolved, and new variants found using these methods[16].

A large number of protein spots may be resolved in the IgG light chain region of both human and animal plasmas[1, 17], and the isoelectric point of multiple myeloma light chains determined[18].

Using human peripheral blood lymphocytes, the alterations induced by concanavalin A have been examined[19], the map positions of the major mitochondrial proteins determined[20], and the discovery made that mitochondrial inhibitors specifically stop completion of their post-synthetic processing, which appears to be highly integrated. The β and γ cytoplasmic actins of lymphocytes are differentially thermostabilized by MgADP, which is bound more strongly to the γ form[21]. The effect of drugs and hormones on gene expression is especially interesting when the effects on any one of several thousand proteins can be seen in one experiment. To date, some 18 different multiprotein effects have been described in studies of the effects of drugs, antibiotics, and hormones on lymphocyte proteins[22]. A factor has been discovered in human urine which reproducibly increases the synthesis of some proteins in Novikoff hepatoma cells[23], or human lymphocytes[24], providing a model for studies on lymphokines (interleukins) and leukotrienes.

Characteristic pattern differences between normal and leukemic cells[25], after lymphocyte heat shock[26], in mononucleosis[27], and in rheumatoid arthritis[28] have been discovered. In addition the differences between lymphocytes and granulocytes, and between different subtypes of types of lymphocytes have been described[29].

The contractile proteins of rabbit and human muscle, and a number of the major enzymes have been mapped in rabbit and in human muscle[30]. Interesting alterations have been described in these patterns in several muscle diseases[31]. Using sensitive silver staining[32] or postsynthetic radiolabelling, it has been found possible to produce maps of major muscle protein from single frozen section of muscle[33]. A variant of human muscle non-muscle tropomyosin has been found in the cultured fibroblasts from one member of a set of non-identical twins which was absent from the fibroblasts of the other[34].

We do not review here the variety of ancillary methods which are now being developed and exploited in many laboratories, many of which we have reviewed elsewhere[2]. These include methods for cell separation and fractionation, group separations based on affinity methods, a variety of approaches to spot identification, microme-

thods for the determination of amino acid composition and amino acid sequence, and approaches to the problem of producing specific antibodies to proteins in spots, and to isolating and cloning genes for specific proteins.

Relationship of the HPI Project to Ongoing Research

We must deal explicitly with the now too widely held view that implementation of the HPI Project would in some way complete with, supercede, and render less important traditional biomedical research carried out by small groups with relatively modest grant support. This view is fortunately incorrect. The feasibility and demonstration studies for the HPI project are now complete. The changes and integrations envisioned if the HPI project is indeed carried out on the next larger logical scale are necessary extensions of developments which exist, and involve changes and integrations which we believe are inevitable in biomedical research.

Simple statistics suggest that there is no guarantee that the protein underlying a given type of cancer (or other disease) may not be the very last protein to be isolated and studied. If new proteins are isolated and characterized at the same rate as in the recent past, then a thorough study of the role of each protein separately in the several thousand known diseases will be an extraordinarily lengthy and costly procedure. Loss of public interest and support for such protracted research would be quite understandable.

It should not be thought, therefore, that there is a simple choice between only two approaches to the molecular basis of human disease, one based on accident and empirical findings, and the other on an impossibly protracted research schedule.

Rather the third alternative is to apply very high technology to the separation and characterization of human proteins (and later to the sequencing of human DNA), and raise biochemical research to an entirely new plane. Regarding accidental discovery, all sincerely hope for new therapies which relieve human suffering and cure disease even if the underlying mechanisms of action are not understood, and that the protein variants underlying major genetic diseases are found by simple means in the near future. Prudence dictates, however, that a systematic alternative to chance, luck, and accident exist, at least in this part of biomedical research.

We conclude that the HPI project, if carried near completion, will provide startling new research possibilities to the majority of clinicians, clinical laboratory scientists, and biologists. These include the discovery of the causes of many human genetic diseases, a systematic analysis of the programming of gene expression during human development, and the global analysis of the effects of drugs, toxic agents and pollutants, hormones, lymphokines, infectious agents,

age, and physical insults on cells and tissues. The objective is to study what might be termed »digital« effects such as the switching on and off of genes, and analogue effects including effects on all post-translational synthetic processes, membrane effects which may produce cell leakage, and effects on specific cellular control circuits.

If the expectation that a large number of corregulated gene batteries or sets exists is correct, then it should be possible to identify at least one marker protein for each set, and to use fluorescence or otherwise tagged antibodies to follow the switching on and off of sets during human development. In addition the set markers and post-transcriptional control circuit markers may be used to study the effects of a variety of experimental variables.

Provision for the production of a large battery of specific antibodies is an integral part of the HPI project. These should become widely available, and be exploited for research and diagnostic purposes by a wide variety of investigators.

We suspect that for the next five or ten years, the high resolution separations systems developed for the HPI will be used largely to find markers which will then be made the basis of single immunospecific clinical tests. As use of these markers, and of tests based on them proliferates, and as the Index data base is enlarged to include increasing numbers of disease-related correlations, then it is probable that complete analysis of complex protein mixtures will supercede the use of a large battery of single tests.

Author's address:
N. G. Anderson, N. L. Anderson, Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory, Argonne, Illinois 60439, USA.

References

[1]*Anderson, N. G., N. L. Anderson:* Molecular anatomy. Behring Inst. Mitt. *63:* 169 (1979). – [2]*Anderson, N. G.,* and *N. L. Anderson:* The human protein index. Clin. Chem. *28:* 739 (1982). – [3]*Anderson, N. L.* et al.: The TYCHO system for computerized analysis of two-dimensional gel electrophoresis patterns. Clin. Chem. *27:* 1807 (1981). – [4]*O'Brien, S. J.:* Nature (London) New Biol. *242:* 52 (1973). – [5]*Bishop, J. O.:* The numbers game. Cell *2:* 81–86 (1974). – [6]*McKusick, J. A:* Mendelian Inheritance in Man. 5th Ed., Johns Hopkins University Press, Baltimore, MD, 1978, p xiv. – [7]*Ohta, T., M. Kimura:* Nature (London) *233:* 118 (1971). – [8]*Anderson, N. L., N. G. Anderson:* Proc. Nat. Acad. Sci. USA *74:* 5421 (1977). – [9]*McConkey, E. H., B. J. Taylor, H. DucPhan:* Proc. Nat. Acad. Sci. USA *76:* 6500. – [10]*Walton, K. E., D. Styer, E. I. Gruenstein:* J. Biol. Chem. *254:* 7951. – [11]*O'Farrell, P. H.:* J. Biol. Chem. *250:* 4007 (1975). – [12]*Anderson, N. G.:* Z. Anal. Chem. *261:* 257 (1972). – [13]*Anderson, N. G., N. L. Anderson:* Anal. Biochem. *85:* 331 (1978). – [14]*Anderson, N. L., N. G. Anderson:* Anal. Biochem. *85:* 341 (1978). – [15]*Anderson, N. L.* et al.: Electrophoresis, in press (1982). – [16]*Anderson, N. L., N. G. Anderson:* Biochem. Biophys. Res. Commun. *88:* 258 (1979). – [17]*Anderson, N. L.:* Immunology Letters *2:* 195 (1981). – [18]*Anderson, N. L.* et al.: High resolution two-dimensional electrophoretic mapping of human proteins (*B. Radola,* Ed.). Electrophoresis '79, *W. deGruyter,* pp. 313 (1980). – [19]*Willard, K. E., N. L. Anderson:* Alterations of two-dimensional electrophoretic maps of human peripheral blood lymphocytes induced by concanavalin A (*B. Radola,* Ed.). Electrophoresis '79, *W. de Gruyter,* pp. 415 (1980). – [20]*Anderson, L.:* Proc. Nat. Acad. Sci. *78:* 2407 (1981). – [21]*Anderson, N. L.:* Biochem. Biophys.

Res. Commun. *89:* 486–490 (1979). – [22]*Anderson, N. L.:* Studies of gene expression in human lymphocytes using high-resolution two-dimensional electrophoresis (*R. C. Allen, P. Arnaud,* Eds.): Electrophoresis '81, *W. de Gruyter,* pp. 309 (1981). – [23]*Willard, K. E., N. G. Anderson:* Biochem. Biophys. Res. Comm. *91:* 1089 (1979). – [24]*Willard, K. E., N. G. Anderson:* Clin. Chem. *27:* 1327 (1981). – [25]*Willard, K. E.:* Potential of two-dimensional electrophoresis of human leukocyte proteins for clinical diagnosis in rheumatoid arthritis, infectious mononucleosis, and leukemia (*Ebert, M. H.,* Ed.) Applications of two-dimensional electrophoresis in clinical research. Ann. Intern. Med., in press (1982). – [26]*Anderson, N. L.* et al.: Clin. Chem. *28:* 1089 (1982). – [27]*Willard, K. E.:* Clin. Chem. *28:* 1031(1982). – [28]*Willard, K. E.* et al.: Clin. Chem. *28:* 1067 (1982). – [29]*Willard, K. E., D. Haugh, M. R. Loken:* Analysis of human leukocyte subpopulations by fluorescence-activated cell sorting with subsequent two-dimensional gel analysis. In preparation (1982). – [30]*Giometti, C. S., N. G. Anderson, N. L. Anderson:* Clin. Chem. *25:* 1877 (1979). – [31]*Giometti, C. S.* et al.: Clin. Chem. *26:* 1152 (1980). – [32]*Merril, C. R., R. C. Switzer, M. L. Van Keuren:* Proc. Nat. Acad. Sci. USA, *76:* 4335–4339 (1979). – [33]*Giometti, C. S., N. G. Anderson:* Clin. Chem. *27:* 1918 (1981). – [34]*Giometti, C. S., N. L. Anderson:* J. Biol. Chem. *256:* 11869 (1981).