editorial

CLIN. CHEM. 30/12, 1898–1905 (1984)

Some Perspectives on Two-Dimensional Protein Mapping

Leigh Anderson and Norman Anderson

High-resolution two-dimensional electrophoresis is the only technique currently available which can resolve and map the very complex mixtures of proteins found in cells and body fluids. We review briefly the evolution of protein mapping by use of this technique and then discuss five major issues which will be important in the future. These are 1) development of standardized high-throughput computerized mapping systems, 2) application of mapping to the characterization and identification of cells and organisms, 3) development of a flexible standardized nomenclature, 4) the application of artificial intelligence to the analysis of protein data bases, and 5) the use of 2-D mapping for the detection of mutations. We conclude that new classes of information can and will be acquired that will be important for understanding both normal cell function and disease.

Systematic exploration of the proteins, the working parts of all living cells on Earth, seems to have entered a new and exciting phase. This appears to be due primarily to the widespread perception that the technology on which we rely for the separation and detection of large numbers of proteins, namely high-resolution two-dimensional (2-D) electrophoresis, is being conquered (perhaps beaten into submission) on a variety of fronts. This gradual mastery of a sometimes capricious and generally problematical technology has provided room for an increasing level of intellectual sophistication in the design of experiments and the analysis of data. Although major problems remain to be conclusively solved, we can at least begin to contemplate some issues arising from the present successes of mapping efforts, many of which are described in this symposium.

The Size of the Problem

The point of using 2-D analysis is, of course, to obtain information about a large number of proteins. We care, or at least should care, about many proteins because many proteins are almost invariably present in the biological samples, cells, tissues, etc. that we study. It is, generally speaking, unsafe to exclude any fraction of these blindly, since there is no convincing argument demonstrating that only a small number of known proteins participate in whatever phenomenon we might choose to investigate. Any such assumption is a pleasant fiction.

The importance of recognizing how much remains to be discovered is evident from Figure 1. If we estimate the number of known proteins of all types by using sequence data, genetic data (such as the number of heritable human diseases discovered by the medical profession), or enzyme nomenclature (the Enzyme Commission system) it is evident that a large majority (90 to 95%) of human proteins remain undiscovered. Further, it appears that extrapolations of past rates of discovery into the future yield very discouraging estimates of the time required to identify a major fraction of those molecules. This simple analysis provides the best argument for the importance of systematic protein mapping, and shows why, if 2-D gels did not exist, we would be forced at this point to invent them. While biochemistry has many tools of exquisite specificity (such as monoclonal antibodies, DNA probes, enzyme assays), 2-D electrophoresis currently is the only tool that allows us to observe from one self-consistent vantage point a large fraction of the gene expression and regulatory activity of the cell.

As a tool, then, 2-D electrophoresis appears to have the

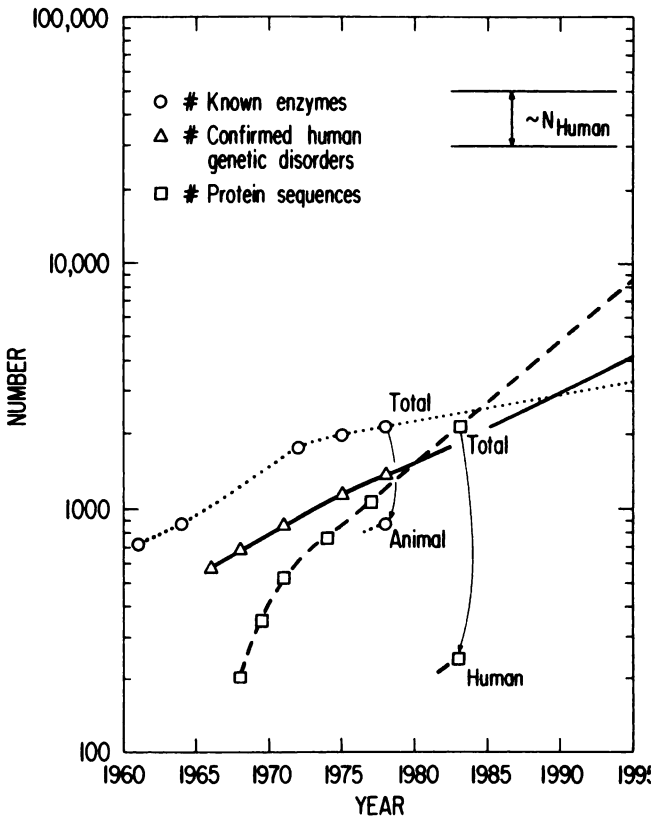

Fig. 1. The number of proteins (on a log scale) that have been characterized in various ways, plotted against time

Estimated number of different human proteins is indicated at the top. The total number of enzymes is taken from ref. *12*, the total number of genetic disorders is from ref. *19*, and the total number of proteins sequenced from all sources and from man are from ref. *13* and Dr. W. Barker, personal communication

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

Received and accepted July 18, 1984.

power required to overcome the present comprehensive ignorance concerning human and other proteins, if it is systematically applied to the problem. The ability to see many proteins naturally leads to curiosity about many proteins, and this is the basis for the expectation that large databases of information on proteins, the so-called protein indexes (*1*), will eventually come into being.

The Evolution of Ideas about Mapping

It is now 10 years since the beginnings of really high-resolution two-dimensional protein mapping (*2–5*), and it is possible to detect some evolution in the prevailing views of this technology. Despite O'Farrell's elegant demonstration of the capability of the system to resolve almost all the proteins of *Escherichia coli,* most early users of the system exploited it as a kind of highly specific assay for specific proteins. More recently, there has been a shift towards the notion of complete indexes of the proteins of the various organisms and tissues. Why has it taken so long for this idea to be accepted as an attainable reality? One reason appears to be the lack of a pre-existing methodological and intellectual basis for the concept of really large catalogs of information in biochemistry. Such catalogs exist in taxonomy (catalogs of species) and in medicine (catalogs of diseases), but biochemistry has been oriented more toward the detailed analysis of individual proteins. Molecular biology has been similarly oriented; sequencing, X-ray crystallography, and molecular genetic analysis are generally applied to carefully selected genes and proteins.

A more important reason for the slow acceptance of comprehensive mapping may be the lack of information-handling systems capable of managing the wealth of rather disorganized information on proteins that is already available. The systems that did exist (obviously computer-based systems) were not directly useful to most biologists. Even now, there is little appreciation among either the information specialists or the biochemists of the degree to which 2-D gel technology and computerized databases of biochemical knowledge are directly interdependent. The future is likely to reinforce the notion that 2-D map-based protein databases and the simpler DNA sequence databases (*6*) will together constitute the principal gateways through which computers come into biology. With them will come the capacity to manage complexity and to solve problems of gene control that cannot now even be expressed for lack of a suitable language.

Old Issues That Have Been Retired

A striking reminder of the maturation of the 2-D field is the absence of several issues that once seemed important. These included particularly a series of arguments concerning the "correct" approach to the display of 2-D data. O'Farrell originally presented his patterns with the acid end to the right and high molecular mass at the top, in order that the direction of protein migration be from left to right, and from top to bottom. However, later techniques for the separation of basic proteins (the NEPHGE approach; ref. *5*) used a reversed orientation (direction of migration towards the left) in order to maintain consistency with the results of the original system (acid end on the right). Unfortunately, this method of presentation (showing low pH values on the right) is not a conventional Cartesian coordinate system with regard to pH and molecular mass: the conventional X coordinate is reversed. A second school of thought presented results with acid proteins (low pI) on the left and high molecular mass at the top, thus adhering to the Cartesian convention. The latter system has the practical advantage that shifts of protein spots to the left correctly indicate modifications yielding more negative charge, and vice versa.

The controversy over this issue, reminiscent of the war between "little enders" and "big enders" in *Gulliver's Travels,* has now almost disappeared (at least in the core 2-D community) as a result of general acceptance of the Cartesian system. The emergence of such a consensus is indicative of a serious commitment to the exchange of data, since the conversion from one system to the other requires substantial effort on the part of a number of workers.

A further series of debates once raged concerning the best way to label unknown proteins on 2-D maps. Various schemes were devised for dividing up the map into quadrants, landmark areas, or other defined regions and then for numbering or lettering the proteins in each area. While landmark areas may be useful for indicating groups of proteins for discussion, they do not provide a suitable basis for a general identification scheme. Other investigators rejected the use of exact physical coordinates on the grounds that these were not sufficiently well standardized to form the basis for a permanent map reference system.

Over time it has become evident that very little information has actually been transmitted between laboratories on the basis of physical map coordinates; more often, photographs are exchanged and notations made indicating analogous proteins recognized by eye. The real solution to this problem—comprehensive standardization of gel technology—remains a goal for the future. In general, investigators have observed that content matters more than form; systems that contain or allow access to information are used and others are not. Thus it has become clear that none of the primitive systems in present use is likely to be dogmatically adhered to, and that a new system, perhaps heavily influenced by the evolution of computerized databases, is likely to evolve. In this case, experience has taught that the laborious development of a body of self-consistent results represents a greater contribution than the premature declaration of a standard.

New Issues

Of the range of major issues which appear to be central to the long-term future of protein mapping efforts, we wish to touch on aspects of five:

- the development of 2-D electrophoresis into a highly-reproducible, standardized, high-throughput data generation system
- the increased importance of the identification of organisms, from patent and other perspectives
- the necessity for a flexible, computer-compatible nomenclature system for both known and unknown proteins
- the possible utility of artificial-intelligence approaches in the task of understanding the relationships between large numbers of proteins
- the proper role of 2-D mapping in the detection of mutations

2-D Technology

The technical requirements that have been repeatedly stressed at this symposium are high resolution, reproducibility, sensitivity, and fast scanning with a wide dynamic range. Unless a technical advance produces a startling improvement, some objective measurement is needed. For example, it is common experience that larger gels give improved resolution. However, quantitative data on resolution to support this conclusion are rarely provided. One cluster or constellation of spots may be used for a visual resolution test, but such a test applies only to a very limited region of a gel. We have proposed an objective mathematical

test (7) applicable to whole gels, and we hope that such measures will soon be much more widely used.

Reproducibility requires that both isoelectric focusing and sodium dodecyl sulfate gels be prepared and run under well-defined conditions. We have developed methods for simultaneously casting sets of gels for both dimensions (*8, 9*), but we recognize the need for extreme reproducibility over even larger sets of gels. If precast gels are to become an item of commerce, sequential machine casting of gels may be required. With Immobilines (*10*) the possibility exists of both standardizing and calibrating the first-dimension separations, and work to this end should be encouraged.

If most such analyses ultimately are performed for toxicological, pharmacological, mutational, and protein-indexing purposes, then one or more large centralized laboratories will be required. For such facilities, automatic or robotic systems for casting, running, and analyzing gels will be developed. A major advantage of such operations would be that of running continuously, with all the benefits in long-term consistency that this entails.

If, in contrast, most analyses and protein identifications are done by individual investigators, then every effort must be made either to develop very stable precast gels or to make it possible to cast such gels reliably in small batches. In all likelihood, research and development along both lines will proceed simultaneously, with centralized laboratories providing much of the basic standardization and most of the protein identifications and database management, while individual investigators provide new leads, and some identifications, in cooperation with central facilities. If 2-D analyses should ever become routinely used for diagnostic purposes, the analyses probably will be done in only a few centralized laboratories. Similarly, if mapping of various human-cell types and identification of protein spots on such maps are to be carried close to completion, it will, in our view, not be done exclusively or even primarily by the accretion of theses or grant-supported studies. The Human Protein Index Project, and parallel projects for animal, plant, or bacterial cells may therefore require special organizational arrangements and support, especially if the underlying technology is to be optimized.

Sensitivity of detection continues to improve as new modifications of silver staining and improved methods for autoradiography and fluorography are developed. The mechanisms underlying silver staining are not fully understood, suggesting that if they are more fully explored, some increase in sensitivity may be obtained. Alternative physicochemical development systems, such as the Eastman nickel-based stain, warrant intensive development.

The problem of developing satisfactory scanners appears to be approaching a solution. The core problems have been dynamic range and speed, because the required resolution (50–100 μm) seems to be generally attainable with all except television systems. Because of dynamic range problems originating in lens flare, scanners that depend on lens optical systems (mainly photodiode arrays and television systems) have not been satisfactory for most applications. Existing mechanical scanners with a wide dynamic range have been too slow for routine work. Hence there is substantial interest in laser-based flying-spot scanners such as the one described at this meeting.

How (and how fast) 2-D technology develops will depend in large part on which applications are found to be most useful. Applications in basic research, clinical chemistry, industrial research and development, agriculture, or in government-sponsored population or environmental monitoring could each support the required developments, though at very different levels and in different styles.

2-D Maps, Patents, and the FDA

Clinical chemistry stands at a four-way intersection between patients, physicians, basic research, and industry. For research to reach the patient, industrially-produced products are usually required, and many of these are based on highly purified, sometimes genetically engineered proteins. These products require patent protection, which in turn involves conclusive identification. In addition, methods for certifying purity of protein products will become increasingly important. As patent and quality-control aspects of protein production gain in importance, 2-D analyses will be more frequently used in communications between industry and government agencies.

For proteins and the cells that produce them, the patent situation is unclear (*11*). The case of *Chakrabarty vs Diamond, Commissioner of Patents and Trademarks,* established that living organisms could be protected by patents, provided that the standard criteria of usefulness, novelty, and unobviousness were met. In addition the "invention" must be described in such detail that a person skilled in the art could reproduce it. The question of what constitutes an adequate description of a cell or organism and of what constitutes "novelty" have not been resolved. A single base substitution in DNA may give rise to an amino acid substitution that dramatically alters the behavior of a cell. Since such a single-base substitution is the irreducible minimum genetic alteration, it is difficult to see how the concept of a "minimum distance" proposed for discussion by the Internation Union for the Protection of New Varieties of Plants (UPOV) can be anything more than such a single-base substitution (*11*). For genetically engineered cells where the objective is to introduce one new gene, that gene is the difference for which protection is sought. However, what is patented is the whole cell, or in the case of plants, the whole plant. The question of patenting a whole animal does not appear to have been raised. Cells or plants are rarely homozygous, and the questions of other genetic differences, or of new mutations which may subsequently appear, have only begun to be considered. The problems of description and reproducibility have been partly solved by the requirement that samples of patented organisms be made available—for example, through the American Type Culture Collection. Electrophoretic analysis may show these to be mixtures of very slightly different cells, leading to additional problems.

The only conclusion we can draw is that, in part because of the legal problems that are now raised, the description and definition of cells and of the proteins from them will increasingly depend on high-resolution separations, and perhaps mainly on 2-D electrophoresis. It is, in fact, very likely that a major driving force for improved resolution and standardization will come from the plant breeding, genetic engineering, reagent, and biologics manufacturers, and from the patent and regulational problems that they encounter.

Nomenclature

The development of some form of practical protein nomenclature is an important issue, given the large number of proteins that can now be seen on two-dimensional maps. Since the functions of a vast majority of these molecules are unknown, they cannot yet be given names of the type generally favored in biochemistry (such as enzyme names, organized according to catalytic properties into an internationally recognized system; reference *12*). Instead, we are faced with the necessity of devising a system of flexible, perhaps temporary, nomenclatural constructs, formulated

in such a way as to give the maximum immediate usefulness and the minimum long-term inconvenience. Given computers to handle the work of connecting a nomenclature to images, maps, protein databases, and to the research literature, a flexible user-oriented approach seems feasible.

Purpose. The primary purpose of a protein nomenclature must be to serve the needs of people working on proteins. Thus a practical approach to the generation and handling of names is required. In particular, names should be of a type easily remembered and dealt with by people. A second purpose is the facilitation of information transfer to and from the computerized databases in which data concerning proteins will be increasingly concentrated. A third purpose, and perhaps the most difficult to address, is the making of a system that can be logically extended to meet the needs of the long-term future.

Philosophy. Several considerations beyond the purely practical enter into the design of a useful nomenclatural system. In the present case there are, to begin with, several boundary conditions that limit the elegance of what can be achieved. It will be some time before the major function(s) of most proteins are known, and thus the preferred functional names will emerge in most cases only after extensive future work. In addition, it is likely that each protein will be given multiple names as it is detected in different cell types, species, clinical samples, or in different experimental systems. Thus a protein defined in a particular two-dimensional mapping system may be simultaneously defined by a monoclonal antibody, by genetic work establishing a new genetic disease, by a cloned and sequenced stretch of DNA, or by 2-D mapping of tissue from a related species. By such routes a series of different names may be generated by workers in a range of areas, all denoting the same gene product. It is thus a requirement that any successful system be able to handle synonyms gracefully and in a dynamically expandable way.

From a logical point of view, it also seems necessary to require that names refer only to molecules rather than to positions in a particular mapping system. There is a practical motivation for this requirement, namely, that 2-D mapping systems are as yet insufficiently standardized to allow the use of map position alone for conclusive identification. The deeper motivation arises from the desire to develop a nomenclature that expresses something more fundamental than a pattern of spots. Since the nomenclature is ultimately supposed to provide a basis for study of the relationships between the proteins, it must name proteins and not spots *per se*. Some more-detailed characterization is therefore required, and if the protein is related directly (as through post-translational modification) to another in the system, then this should be explicitly noted.

Any system of names should take into account the preferences of the human mind and tongue; neither appears to prefer numbers to words. Indeed there are few if any systems of human language which use numbers; the telephone system constitutes the most numerical system we all use, and most people have difficulty remembering more than 10 or 100 numbers of 10 digits. Words generally have complex meanings and are more easily remembered. We each know and use thousands, and we can learn thousands more when knowledge of a new language is required. Further, the species have successfully been named by using (mainly made up) words; and the enzymes, which are named by both word and number systems, are referred to almost exclusively by word names. Since it is to be hoped that a language of some sort evolves to express the relationships between proteins (covering chemical, differentiational, and evolutionary relationships), it may prove wise to start with a system that is relatively similar to a true language.

A final point of design philosophy is the role of computers in protein nomenclature. It is in fact the availability of computing machinery that makes it possible to develop and manage a flexible nomenclature on the scale required by protein indexes. Only computerized systems can handle, reliably and without complaint, a nomenclature allowing many synonyms and thousands of entries. This does not seem to be a particularly serious drawback, as personal and other computers continue to proliferate at a rate of millions of units per year. In addition, it seems clear that most other information resources available in biology or biotechnology are increasingly computer-oriented; major examples include the nucleic acid and protein sequence databanks (*6, 13*), the hybridoma databank (*14*), as well as the literature search services. Hence it appears justifiable to include the computer from the outset, and to build a nomenclature appropriate for use as the basis of a computer-interpretable language of protein characteristics and functions.

Approach. We wish to propose the following general approach to the development of protein nomenclature, based on practical experience with various protein indexes that are in an embryonic stage and on discussions occurring at this and previous Argonne symposia on protein-mapping technology. It is intended that such a nomenclatural language grow to resemble in some respects a high-level computer language such as LISP PROLOG, or PASCAL.

1) Protein names should be words (maximum length 256 characters), and should include numbers only as secondary identifiers. For practical reasons, only the commonly recognized ASCII characters (excluding punctuation marks) should be used; i.e., no greek letters, subscripts, or superscripts. In general, memorable words without other confusing meanings should be selected; otherwise all words, whether familiar (George, Ford, Philadelphia, or rollerskate) or newly coined (zligzin, brankthup, trigsum, etc.) should be allowed. The use of a vague term combined with a rough measurement (molecular mass, pI, etc.), as in p60-src or 100K-heat-shock protein, should not be encouraged since the resulting name appears to provide a specification for a protein but does not, in fact, exclude the existence of two or more different proteins with similar characteristics.

2) Names may be applied to (*a*) a set of proteins, (*b*) a generic protein, or (*c*) a precisely specified version of a generic protein. Examples of these three kinds of name as *currently* applied (referring to Figure 2) are the use of the class name "plasma proteins" to denote the set of proteins found in the non-cellular fraction of human blood, use of the term "plasminogen" to indicate a generic protein (in this case a circulating protease precursor), and use of the more specific term "*glu*-plasminogen" to indicate one of two predominant circulating forms of the protein (this one having an N-terminal glutamic acid residue). It is clear that in fact *glu*-plasminogen is itself composed of a train of spots of approximately constant molecular mass but varying in charge. Thus a more precise specific name would refer to the fact that a series of at least eight charge isomers (differing presumably in sialic acid content) comprise the *glu* form. Although most cellular proteins do not display as much microheterogeneity as is seen in the plasma proteins, the characterization and expression of the detailed results of chemical modifications is nevertheless very important.

The notion of named sets makes it possible to give similar names to a set of polypeptides sharing some important characteristic. In cells, the mitochondrial proteins, phosphoproteins, or fibroblast-specific proteins can thus be enumerated together using a class:member identifier. Thus the second member of the set of mitochondrial proteins may be

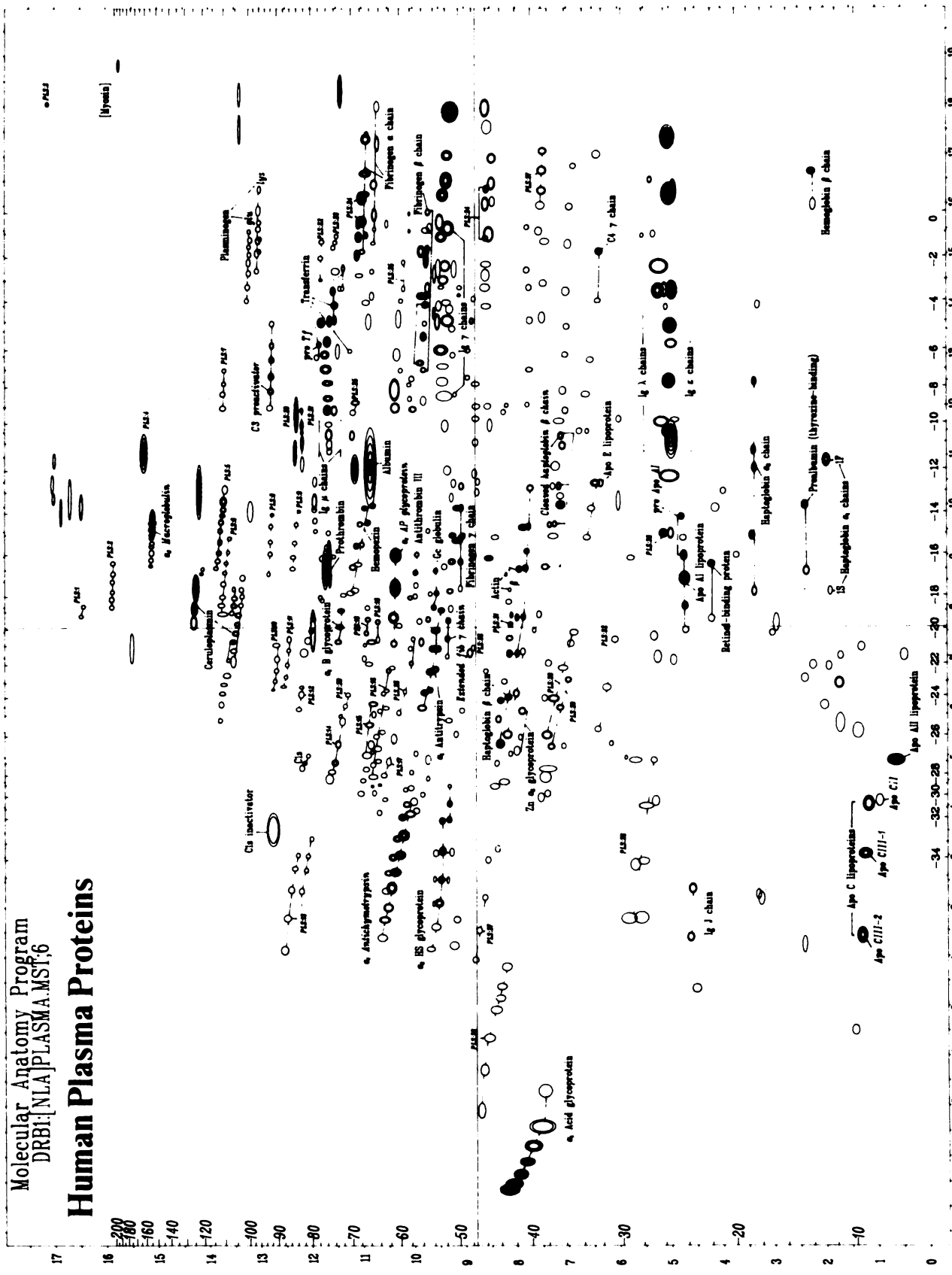

Fig. 2. Two-dimensional electrophoretic map of human plasma proteins
The main identified proteins are indicated, as well as the principal unidentified molecules (PLS:n codes). (Adapted from Anderson et al, ref. 15)

designated Mitcon:2. As becomes apparent when analyzing any classically purified protein in a 2-D system, generic protein names such as serum albumin, transferrin, actin, etc. also refer to classes of molecules varying in degree of deamidation or sulfhydryl oxidation, or even produced by similar (but different) genetic loci or alleles. A useful protein nomenclature should recognize the hierarchical nature of protein definitions explicitly, and by so doing allow use of equivalent class:member, generic protein, or exact nomenclatures wherever appropriate. The acceptability of synonyms allows the creation of a new short name whenever increasing specification makes the current name cumbersome.

3) The original description of a protein should include

substantially more information than simply its position on a 2-D map, particularly if a name more specific than a simple class:member name is contemplated. Sufficient information to allow recognition of the protein in another laboratory's 2-D system should be given. This requirement could be satisfied, for example, by a description including the information that the protein has approximate pI of X, SDS-molecular mass of Y, is phosphorylated, co-isolates with the major membrane proteins when technique ABC is used, and is dephosphorylated in cells treated with compound Q. Although this information is far short of that required to deduce the function of the protein, it is sufficiently detailed to indicate that someone is thinking seriously about the molecule and may have a need to refer to it more than once. Less-unique information, limited to the knowledge that a protein is phosphorylated, for instance, is often not sufficient for the generation of a unique name, because there are so many phosphoproteins. In such a case, a class:member name (PO4protein:1, etc.) could be more appropriate. Irreproducible characterizations, consisting for instance of the identification of a protein based on reaction with an antiserum that is not generally available, should not generally suffice for naming.

4) Although a serious attempt should be made to ensure that a protein to be named has not previously been named, it should be recognized that no currently implementable system will be leakproof in this regard. Therefore the system itself, in particular the computer systems that maintain the nomenclature, should provide for the use of numerous synonyms. A major benefit of such a liberal approach is that it correctly recognizes that many, if not most, of the names assigned will ultimately be changed to reflect functions when these become known. In addition, the capability to handle synonyms allows us to make use of class:member names before generic names are given.

5) One alternative name for a protein is its detailed chemical specification. Although the polypeptide sequence is generally too long to be of use as a name, modified versions of a particular polypeptide could be named in a way that reflects the chemical nature of the modification. Such a nomenclature could be very useful in describing regulatory and genetic relationships between proteins (spots). We suggest that a system be developed that can indicate chemical modification status explicitly by a shorthand code. Such a code could follow the "root" protein name where necessary. Code symbols might include p for added phosphate, pser for phosphate located on serine residues, ptyr for phosphate located on tyrosine, or ptyr(142) for phosphate located on tyrosine residue 142 counted from the protein's amino terminus. Likewise for carbohydrate additions, Csial could stand for added sialic acid residues, Cns for neutral sugar structures of undetermined composition, or Chm for a high-mannose sugar structure, each followed by an amino acid residue number if known. When the chemical basis of the modification is unknown, for instance when the protein is more acidic by one charge unit, then the limitation of what is known should be explicit (i.e., code chg-1). The general formula for a modification code would then be:

:MODIFICATION CODE (SITE IF KNOWN) NUMBER OF GROUPS IF NOT 1,

with successive codes separated by colons. A rather complex example, illustrated in Figure 3, describes some of the molecular forms of the haptoglobin beta chain. Consider, for example,

Haptoglobin_beta:Cns2:Csial6.

This molecule might be cleaved in vivo to yield a circulating

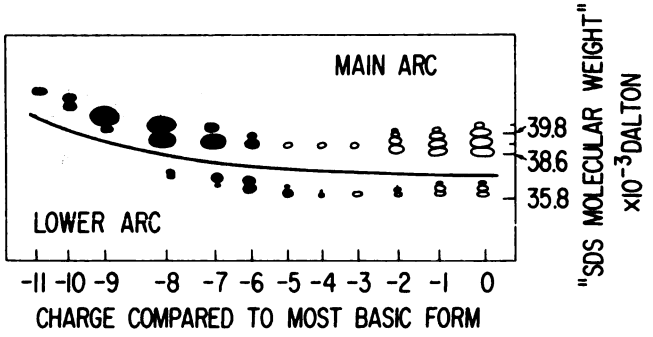

Fig. 3. Map of haptoglobin β chain, illustrating post-translational modifications

Heterogeneity within the main arc is due to glycosylation; the lower arc can be derived from the main arc by a proteolytic cleavage event removing an approximately 3000-Da peptide bearing one sugar attachment site. (From Anderson and Anderson, ref. *16*, with permission)

polypeptide form approximately 3000 Da smaller:

Haptoglobin_beta:Cns1:Csial4:Pep-5k.

Obviously, a unique and recognizable code must be devised for each type of protein modification. For ease of use, it should be possible to place the codes in any order without loss of meaning. As can be seen in the examples given, there will be a hierarchy of levels of specification, in which a conventional polypeptide name such as "haptoglobin beta chain" will occupy a middle position. From 2-D mapping, it is known that this name actually applies to a set of molecules including

Haptoglobin_beta:Cns{1to4}:Csial{4to10}Pep{0,-5K}.

Haptoglobin beta chain is itself part of a larger set:

Haptoglobin:Subunit{alpha1s,alpha1f,alpha2}2:
Subunit{beta}2,

or in other words the haptoglobin tetramer composed of two alpha and two beta chains.

6) Names should in general be devised and applied by those working for the first time with the protein. However, the relaxed use of synonyms leaves the selection of "preferred" names to the scientific public at large.

Usefulness of this and other naming schemes. As indicated, the type of computer-based nomenclature proposed here allows considerable flexibility (based on the computer's facility in handling synonyms) and affords, as one possible basis for names, some scheme for indicating chemical relationships between protein spots. There are of course other kinds of relationships: the relationship of successive enzymes in the same metabolic pathway, the relationship between interacting proteins forming part of the same cytoskeletal structure, or the relationship between a DNA-binding control protein and the protein whose expression it controls. The long-term goal of a serious effort at generating a protein nomenclature would be a system incorporating all these types of relationships. A user would be able, beginning with any molecule (or set) of interest, to explore all other proteins connected to it by any of a variety of relationships.

Graphical presentation. Names are essential for publication and for record keeping. The essential information regarding a protein may, in many instances, be displayed rapidly and conveniently by use of currently available graphics systems. Thus the substructure of a protein that is synthesized, clipped twice, glycosylated, and then sialylated may be conveniently displayed as shown diagramatically in Figure 4.

Artificial Intelligence

It has often been noted that a major difference between

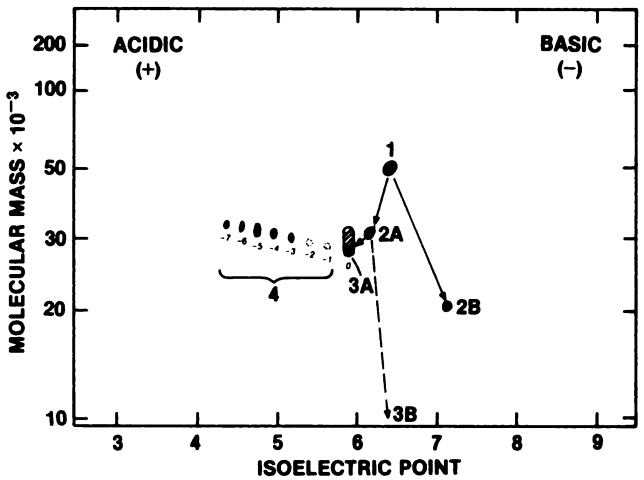

Fig. 4. Diagrammatic presentation of post-translational modification of a protein

The original translation product is shown at *1*. The hypothetical protein is cleaved into two unequal products, which are *2A* and *2B*. Note that the sum of their masses equals that of *1*, and that since one cleavage product has an isoelectric point more acid than *1*, the other must have a more-basic pI, as shown. Protein *2A* is clipped, and the small product, *3B*, is too small to be seen on the map. The larger product, *3A*, is glycated in a non-uniform manner to give spot *3A* plus the mass heterogeneity shown in the *hatched* region. Negative charges (sialic acid groups) are then added, giving rise to charge heterogeneity. No charge is added to spot *3A*, hence a zero is placed beneath it. Each additional charge gives rise to a separate spot, and these are numbered *−1* to *−7*, meaning that from one to seven sialic acid groups are added. The maximally glycated versions of *3A* have the most sialic acid attachment sites, the minimally glycated versions the fewest. Hence the charge train shown at *4* slopes up and to the left. Note that in this form of data presentation Cartesian coordinates are used

the biological and physical sciences is that physics is made up of a small number of theories, particles, and fields—which can be very difficult to understand. In contrast, biology deals with a very large number of entities (genes, proteins, cells, and species) whose relationships—once discovered—are individually not so difficult to understand. Biology thus deals with complex combinations of the simple, a view consistent with the evolution of biological systems through the survival of useful, simple, accidental tricks. The problem is to keep track of these tricks.

When we begin to contemplate the complexity of the systems controlling cellular gene expression, it is necessary to ask whether individual researchers will be able to absorb enough information to be able to discover the underlying patterns. The possibility that people cannot retain and organize sufficient data internally to solve some major problems leads us to consider the potential usefulness of advanced computer systems, not only to generate and store but also to help analyze the complex functions embedded in living cells. We suggest two areas in the field of artificial intelligence (AI; reference *17*) that could be of use in addressing these limitations.

1) Database structure is perhaps the most immediate area of interest. As indicated above, a protein index may contain not only the customary sorts of information about a protein (sequence, physical location, rate of synthesis, etc.), but also a wide variety of different relationships to other proteins, genes, substrates, etc. Most large database systems currently in existence (such as airline-reservation systems or inventory databases) make use of a standard scheme for all entries; an analogous scheme for a database of proteins would contain, for each protein, a set of fields covering each possible characteristic or relationship. Clearly, this is extremely inefficient for the type of protein index envisioned here, because we may have characteristics or types of relationships that are unique to each protein. It is inefficient to reserve space to describe the enzymatic characteristics of haptoglobin, for instance, because it has no enzymatic activity. What we require is a system that effectively constructs a model of the network of relationships that constitute the cell, and does this dynamically as information is added. Such a database is likely to rely heavily on methods of knowledge representation and expert system construction developed in AI research.

2) As mentioned above, the number of proteins is so great and their actions and relationships are sufficiently complex that the most useful method of communicating and understanding information at this level may be through a high-level computer-style language. Those who have recently attended seminars describing DNA technologies may conclude that a new language has already been generated in that field. Precisely because of the large number of possible types of relationships between proteins, it is unlikely that a FORTRAN-style, mathematically-oriented language could form an efficient medium for discourses about a large protein index database. Instead, a language more suited to manipulation of symbols (names) than numbers would be required. By combining aspects of "natural languge" interpretation (used to allow English-like interaction with computers) with insights derived from a knowledge of the structure of a protein index database, a reasonably "natural" biochemical language might be developed.

Mutation Monitoring

Two-dimensional electrophoresis offers the possibility of detecting point mutations at an extraordinarily large number of loci, and hence over large regions of coding DNA (Figure 5). Such a capability is important for environmental studies and studies of ionizing radiation and chemical mutagenesis. Approximately 76% of base substitutions in structural genes lead to amino acid substitutions (*18*), and approximately one-third of amino acid substitutions involve a change in the charge of the protein involved. On 2-D maps, single charge changes can be detected easily over a molecular mass range of approximately 10 000 to 100 000 Da (Figure 6). Thus a protein composed of 400 amino acids (the size of actin) can serve to monitor for base substitutions in the 1200 base pairs in the coding region of the corresponding gene at an overall efficiency of about 25%. This is equivalent to detecting any single base change in 300 base pairs. If the 716 proteins on the gel analyzed in Figure 5 are weighted by molecular mass (number of amino acids) and summed, then it is apparent that more than 1.1×10^6 base pairs of DNA are represented in the pattern. Thus this gel pattern represents a system equivalent to a detector capable of revealing any change in approximately 275 000 base pairs of coding DNA. This is far more DNA than is presently monitored by any competing method. Since large proteins serve to reflect changes in longer sequences of DNA, it is important to improve resolution in the high-M_r region of the pattern.

Mutations involving changes in mass (additions or deletions) may also be observed on 2-D gels. Mass changes of about 2% are readily detected, which means that single amino acid additions or deletions generally will not be seen as changes in the second or SDS dimension.

Conclusions

The developments occurring over the past two years, and reviewed at this meeting, have gone far towards establishing the systems and generating the confidence necessary to undertake comprehensive protein-mapping projects. Given the continued development of 2-D technologies, it seems clear that major information resources will be created based on this approach. These advances will result in an ever-

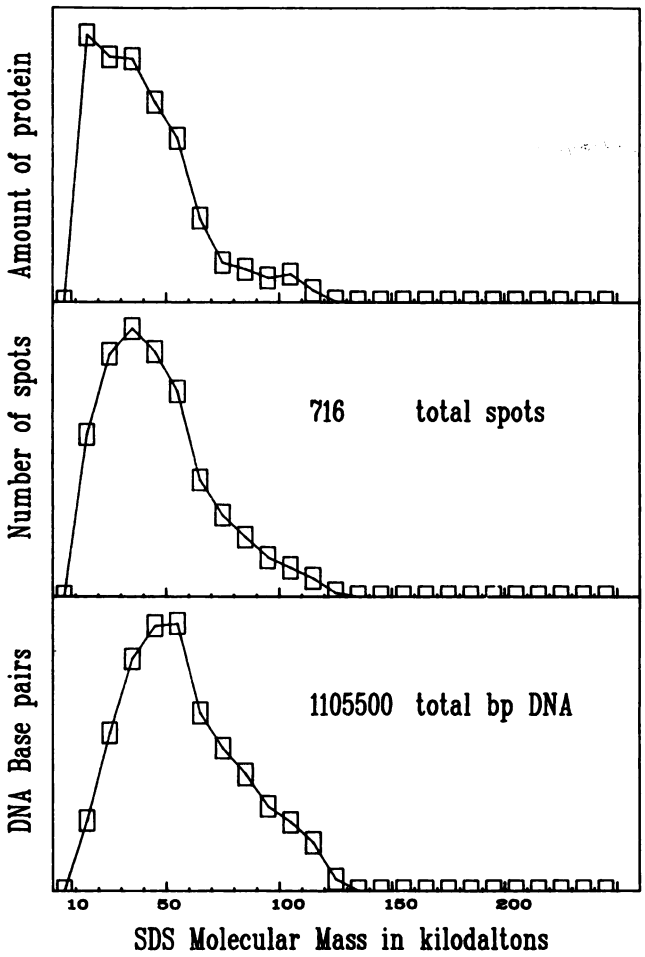

Fig. 5. Histograms relating the amount of protein (*top*), the number of spots (*center*), and the number of DNA base pairs required for coding (*bottom*) to protein molecular mass

These data are from a 2-D gel of the proteins of human lymphoblastoid cells. Note that while most of the mass of the protein is found in relatively low-*M*, proteins, the number of different proteins as a function of mass shows a different distribution and is skewed toward a higher molecular mass. As shown in the bottom section, the higher-*M*, protein, as expected, requires a proportionately larger amount of DNA to code for it. If mutations occur at a constant frequency per nucleotide, then the efficiency of mutation detection is much higher for high-*M*, proteins

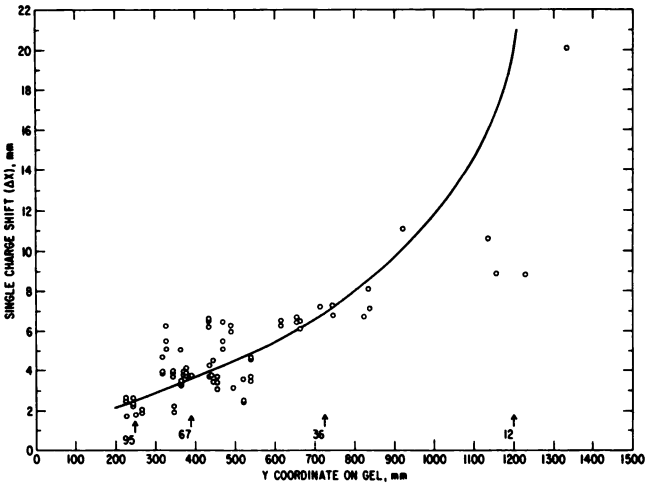

Fig. 6. The size of single charge shifts as a function of molecular mass for a variety of proteins in human lymphoblastoid cells

As expected, charge shift size is roughly proportional to the inverse of molecular mass (*solid line*), but nevertheless it varies by as much as twofold among various proteins of similar size, owing to differing amino acid compositions. Using this wide-pH range 2-D system, single charge shifts in proteins of *M*, 100 000 average 2 mm, or approximately four to six times the spot width. Single charge shifts were generated by carbamylating a whole-cell extract (ref. *18*). Numbers with *arrows* are kilodaltons

closer association between gel methods, sophisticated computing techniques, and the hard information content of biology. Because grasping complexity is the most difficult intellectual aspect of biology, such a unified approach should make possible important new advances in basic biology, agriculture, and medicine.

References

1. Anderson NG, Anderson NL. The Human Protein Index. *Clin Chem* 28, 739–748 (1982).

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

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

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

5. O'Farrell PZ, Goodman HM, O'Farrell PH. High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* 12, 1133–1142 (1977).

6. *Nucleotide Sequences 1984: A Compilation from the GenBank and EMBL Data Libraries,* special supplement to *Nucleic Acids Research,* IRL Press, in press. For further information on the NIH-funded Genetic Sequence Data Bank, contact Bolt, Beranek and Newman, Inc., 10 Moulton St., Cambridge, MA.

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

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

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

10. Righetti PG, Gianazza E, Bjellquist B. Modern aspects of isoelectric focusing: Two-dimensional maps and immobilized pH gradients. *J Biochem Biophys Methods* 8, 89–108 (1983).

11. Williams Jr SB. Protection of plant varieties and parts as intellectual property. *Science* 225, 18–23 (1984).

12. *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. (*Ed. note:* a new edition of this work is expected shortly.)

13. Dayhoff MO (Ed.). *Atlas of Protein Sequence and Structure,* vols 1–5 and supplements, National Biomedical Research Foundation, Silver Spring, MD, and further data-base development headed by W. Barker, NBRF.

14. The Hybridoma Data Bank, supported by the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions, and constructed at the American Type Culture Collection, Rockville, MD.

15. Anderson NL, Tracy RP, Anderson NG. High resolution electrophoretic mapping of human plasma proteins. In *The Plasma Proteins,* 4, 2nd ed., F Putnam, Ed., Academic Press, New York, NY, 1984. In press.

16. Anderson NL, Anderson NG. Microheterogeneity of serum transferrin, haptoglobin, and α_2HS glycoprotein examined by high resolution two-dimensional electrophoresis. *Biochem Biophys Res Commun* 83, 258–265 (1979).

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

18. Anderson NL. Comparison of organisms and cell types using two-dimensional electrophoresis. In *Uses and Standardization of Vertebrate Cell Cultures. Monograph No. 5,* RE Stevenson, Ed., Tissue Culture Association, Gaithersburg, MD, in press.

19. McKusick VA. *Mendelian Inheritance in Man,* 5th ed., The Johns Hopkins University Press, Baltimore, MD, p xiv, 1978.