Review

Norman G. Anderson
Alastair Matheson
N. Leigh Anderson

Large Scale Proteomics Corporation, Rockville, MD, USA

Back to the future: The human protein index (HPI) and the agenda for post-proteomic biology

The effort to produce an index of all human proteins (the human protein index, or HPI) began twenty years ago, before the initiation of the human genome program. Because DNA sequencing technology is inherently simpler and more scalable than protein analytical technology, and because the finiteness of genomes invited a spirit of rapid conquest, the notion of genome sequencing has displaced that of protein databases in the minds of most molecular biologists for the last decade. However, now that the human genome sequence is nearing completion, a major realignment is under way that brings proteins back to the center of biological thinking. Using an influx of new and improved protein technologies – from mass spectrometry to re-engineered two-dimensional (2-D) gel systems, the original objectives of the HPI have been expanded and the time frame for its execution radically shortened. Several additional large scale technology efforts flowing from the HPI are also described.

Keywords: Proteomics / Human protein index / Two-dimensional electrophoresis / Cybernomics / Review PRO 0002

Contents

Correspondence: Dr. N. Leigh Anderson, Large Scale Proteomics Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA
E-mail: leigh@lsbc.com
Fax: +301-762-4892

Abbreviations: HPI, human protein index; **HUGO**, Human Genome Organization; **MAN**, molecular anatomy; **NCI**, National Cancer Institute

1 Introduction

As proteomics now takes center stage in molecular biology, it is appropriate to take stock of progress to date and consider the major strategic objectives that can be achieved during the next stages of its evolution. Just as the sister discipline of genomics set itself the task of sequencing the complete human genome, proteomics now aims to map and identify the entire human proteome, and to compile the human protein index (HPI) as a comprehensive, tissue-specific inventory of the proteins expressed in our species. This database will characterize the differences between the estimated 252 different cell types in man and provide a basic foundation for systematic discovery of the molecular changes underlying a variety of diseases and markers for them, discovery of what drugs actually do in the human body, and identification of new and unique targets for therapeutic agents. Further, the reference data required to characterize different types and subtypes of cancer and stages in cancer progression will be availabe. The HPI is therefore a central project in proteomics as it is presently defined.

The strategies required to approach completion in genomics and in proteomics are quite different. The genome is composed of four chemically similar components arranged in the linear coding order of DNA. Sequencing DNA involves fragmenting DNA strands and, through an automated repetetive process, determining the sequence. DNA molecules are so alike that the sequencing process applies to DNA from any source, a technological universality that explains the exponential rate of progress in genomics. However, as is well appreciated in proteomics, DNA has no function except to store information: it has the intellect of a piece of magnetic tape.

In sharp contrast, proteins are made of approximately 20 amino acids that exhibit a variety of different chemical properties. The amino acids are arranged in different ways in different proteins, and in addition, proteins fold to give complex three-dimensional structures that have thousands of different functions. In essence, while DNA stores information almost without modification for periods up to eons, proteins do everything else. Further, while the DNA in all of the estimated 252 different somatic cell types in man has basically the same sequence, the protein compositions of different cell types are different, both qualitatively and quantitatively. To further complicate matters, the three-dimensional structure, function, and final form of proteins cannot be predicted with certainty from the linear codes of genes, and many, if not most proteins are modified after they are synthesized. The world of individual proteins is thus far larger, more complex, and potentially more rewarding than the world of the genome. Proteins fit together and self-assemble to form the subcellular organelles of cells which are complex mechanical and chemical machines. A complete HPI must ultimately describe where individual proteins are in cells and what they do. To make matters more difficult (and interesting), all cellular proteins are continuously made and destroyed at rates which differ for different proteins, and which change under different physiological conditions.

Current technology is sufficient for the first stage of the project, now well under way, which is high resolution mapping of major tissues and readily obtainable cell types. The second stage will involve cell separation, and will require both existing and new technologies, and considerable progress in this direction has already been made [1]. The third stage involves precision subcellular fractionation and the fractionation of soluble protein mixtures using affinity columns and other methods. The systems for this work either exist or are now in development. The fourth stage involves the production of an antibody library which parallels the HPI. Such a library is required to confirm the cellular and intracellular location of specific proteins, to produce clinically useful tests for candidate disease and inujury markers, and to produce solid state protein chips for routine clinical use (the fifth stage). It is clear that more and different technologies and disciplines are involved in the HPI than in the parallel Human Genome Project, and that much of the technology remains to be invented, developed and refined. It is also evident that the HPI, while a more costly and longer term effort, is more directly related to advances in the medical sciences than is the Human Genome Project. In the final analysis, all human diseases involve changes in the structure, location, or abundance of proteins.

2 Proteomics as “Big Science”

It has been proposed that proteomics will soon become, in words made popular in physics, “Big Science”. While the term has different connotations in different settings, the basic idea is that such projects allow the acquisition of information and products not obtainable any other way. Equally importantly, big science projects allow individual researchers to do things they otherwise could not do. Particle physics (through accelerators) and space exploration (through space vehicles) are the usually cited examples of big science – little science cooperation, cases which immediately bring to mind the fact that big science has been almost entirely publicly supported. Big science itself arose initially from a wartime governmental decision (the Manhattan Project).

Large privately funded laboratories alter this equation, and will be major factors in discussing how the HPI is being done. In particular, the recognition by capital markets that biotechnology will prove to be a potent source of valuable products and revenues in the coming decades has caused an influx of quasi-governmental scale resources into corporate biotechnology laboratories. In the extreme case, this could lead to a situation paralleling the one in solid state physics where essentially all scientific work on the element silicon is carried out in industry (for application in semiconductor chips). The recent “friendly” competition between public and private human genome sequencing indicates that there is more interest in maintaining public participation in human biology than in silicon physics, but the balance of resources available for large focused projects is clearly moving towards the private sector. As C. P. Snow has noted, decisions regarding the earlier large scale science projects were essentially made in secret, in the sense that the scientists ultimately engaged had little initial say. Politicians maintained overall control precisely because of the need for public accountability for the funds employed. With the involvement of private funds and organizations and with intense public interest in biotechnology, this can no longer be true.

3 History of molecular anatomy

The HPI is not a new initiative – in fact, parts of it predate the Human Genome Project by approximately forty years.

Big science and little science were topics thoughtfully dealt with several decades ago by Alvin Weinberg, in his book, *Reflections on Big Science* [2]. A recent editorial in *Nature* asked why Weinberg never proposed a large project in biology, to which he replied that one had been proposed [3], but he did not elaborate. Since the project to which he referred is the lineal parent of both the genome and proteome projects, and, since it is almost entirely unknown, we review it briefly here before discussing concepts regarding future directions.

The Molecular Anatomy (MAN) program, which first proposed a complete analysis of human cells, was presented at Dr. Weinberg's invitation in 1960 as a position paper in a series asking how Oak Ridge National Laboratory (originally involved in the production of enriched uranium for the Manhattan Project) could diversify into other areas of science. After considerable discussions, a program was organized with the support of the National Cancer Institute (NCI), the National Institute of General Medical Sciences, and the Atomic Energy Commission. A summary of the technical accomplishments of the first five years was published as an NCI monograph [4] and was described in a general article in 1967 under the title: "*Molecular Anatomy: Next Major Science Programme*" [5]. This article begins: "Cells and the aggregates into which they arrange themselves – plants and animals – are by far the most complex systems which now engage the attention of scientists. It would be misleading to suggest that the detailed exploration of such systems will be easier than exploring space or the atomic nucleus or that it will ultimately cost less. ...budgets for space and atomic energy exceed by several orders of magnitude the budget for cell exploration at a molecular level. I believe that the reason for this is not that the exploration of space is more important for mankind, is inherently more interesting, or that the goals are more easily realized. Rather the reason is that no comparable programme for the biomedical sciences has been seriously proposed."

The strategy was to develop several different technologies in parallel. These included some of the first high pressure liquid chromatographic systems to resolve low molecular weight compounds [6] including sugars [7] and the constituents of nucleic acids [8], high resolution centrifuges to resolve cell components [9, 10], and methods to separate proteins including rapid recycling immuno-affinity chromatographic systems to simplify complex protein mixtures [11, 12]. Since it was realized that automatic fractionation methods would yield large numbers of samples to be analyzed for enzyme activities, the first computerized enzyme analyzer (the centrifugal fast analyzer) was invented, developed, and came into general use in clinical chemistry laboratories worldwide [13].

Since one of the objectives set by the NCI was to develop the means for making a human cancer vaccine, considerable attention was paid to developing physical methods for detecting, isolating and characterizing unculturable human pathogens, and to the development of large-scale vaccine centrifuges with the aid of the engineering staff of the Separations Systems Division (Gas Centrifuge Project) of the Oak Ridge Gaseous Diffusion Plant. Among the interesting discoveries was the high titer of viruses in the ocean [14], and the first electron micrographs of hepatitis B virus and the Australia antigen [15]. Two-dimensional, so-called s-ρ centrifuges were developed and shown to be able to isolate trace quantitites of viruses from tissues [16]. This collaboration resulted in the development of a series of some 56 different designs for zonal rotors [17] (originally invented by Norman G. Anderson), and the K-II centrifuge for vaccine purification [18]. The K-II centrifuges were used to produce the first purified influenza vaccine [19], and the first hepatitis vaccine. Many of these instruments are still in use for the large scale production of human immunodeficiency virus (HIV) and other viruses. Considerable attention was paid to developing contained systems in which virus-infected tissues could be processed, and large containment facilities were constructed [20]. The first biohazards committee on record oversaw the work.

This Oak Ridge MAN program was the first demonstration of what could be done with what was (at the time and from the nave viewpoint of a biologist) unlimited scientific and technical support. While the project included work on nucleotide chromatography, knowledgeable biochemists advised that DNA sequencing was chemically impossible. A separate project to isolate and sequence tRNA was set up under Dr. David Novelli, who missed publishing a first complete sequence (and getting a Nobel prize) by weeks. However, the MAN program provided invaluable experience in how extraordinarily diverse groups in engineering, mathematics, chemical engineering and physics could be focused on problems of biological interest, and also provided an experimental platform for testing out new organizational concepts. The excitement of bringing large scale resources to bear on key biological problems appealed to many with actual experience with large scale projects, but the basic idea was uniformly rejected by biologists. The failure of the several "wars on cancer" (which were premature), the conclusion that human cancer generally was not due to viral infection, and misconceptions as to both the purpose and the organization of large scale science contributed to the eventual demise of the original Oak Ridge MAN program. And the problem of efficiently resolving complex protein mixtures had not been solved.

4 Large scale protein mapping and the HPI project

In 1975, Klose [21], O'Farrell [22] and Scheele [23] almost simultaneously published methods based on isoelectric focusing in the first dimension and SDS-electrophoresis in the second, thus marking the start of high resolution two-dimensional electrophoresis (2-DE). Beginning in 1976 the MAN program (reestablished at the Argonne National Laboratory in the United States) and others began to develop technologies, built around 2-DE, that would allow human proteins to be separated, mapped, and screened on a mass scale (reviewed in the Symposia listed [24, 25]). With the development of this paradigm, it was intended that automated procedures would reduce human involvement in the generation of data and enable biologists instead to focus on data analysis. Within this framework, the Argonne research program and later Large Scale Biology Corporation (LSB; Rockville, MD, USA, now Large Scale Proteomics Corp., a subsidiary of a larger company renamed LSB) developed the ISO-DALT semiautomated system for 2-DE [26, 27], and the TYCHO [28] and KEPLER software systems for scanning and analyzing gel images. Covalent linkage of acrylamide gels to glass was explored, was found to decrease resolution, and was abandoned.

The culmination of the rapid rise of molecular anatomy at Argonne was the proposal to map the full repertoire of proteins expressed in every human cell type. An article written in 1979 [29] described this goal and discussed the integration of analytical technologies, automation and computer science that would be needed to implement it: "We have therefore systematically investigated how a comprehensive catalogue (of human proteins) might be made, and have devised methods for characterizing them en masse in the mapping process. Given automation of the analyses, direct gel or autoradiographic scanning, and the assistance of computers in quantitation, data storage, and analysis, the project, usually referred to as molecular anatomy, is now technically feasible."

Shortly after that article's publication, the projected database of human molecular anatomy became known as HPI and the project plan began to take shape in earnest. In 1980 the HPI Task Force was formed, following a review of the uses of 2-DE held in the office of Senator Alan Cranston, then the Majority Whip of the United States Senate. The HPI Task Force was chaired by Norman G. Anderson, then at Argonne National Laboratories, and had a membership that included several leading academics in protein analysis, representatives of major commercial organizations, and representatives providing liaison with government institutions including the National Institutes of Health, the Department of Energy and NASA. The report of the Task Force, published in 1980 [30], and refined in subsequent papers [31, 32] set out the objectives of the HPI and proposed strategies for its implementation. For each protein in the Index, a set of descriptors was proposed, including: the map location of its encoding gene; its gel spot location according to standardized coordinates; any literature references on the protein and its function; the amount of protein present per cell; the amino acid composition of the protein; its amino acid sequence when known; its subcellular localization; the coregulational set to which the protein belonged; genetic polymorphisms for the protein and their relationship to disease; and the protein's biophysical properties, including tertiary structural data. The report also envisaged the need for a central laboratory to maintain standardization of data generation, as well as an organizational framework including a permanent secretariat and a central, public access database where results would be curated.

Unfortunately the HPI project as proposed in 1980 was never realized but became consigned to a long dormancy that would last for 20 years. The immediate cause was the election in 1980 of the Reagan administration and the consequent shift away from large scale, federal research projects. Without the backing of a political consensus or federal funding, the momentum to drive the initiative forward was gradually lost during the first years of the 1980s, despite a succession of publications discussing the program and its merits. Thus the first major attempts to comprehensively document the molecular make-up of the human species proved to be an idea ahead of the times. In retrospect, these approaches to protein analysis, both in their conceptual bases and their technological realizations, were clearly an embryonic form of what is now known as proteomics.

From the discussions surrounding the initial HPI proposal it became quite clear that there was a deep divide between nucleic acid oriented molecular biologists and more classical protein chemists, and that, in the period around 1980 any attempt to give clear advantage to one group over the other would only raise problems. Hence a new proposal was written in 1983, was circulated through several government agencies, and finally published in 1985 [33]. It proposed both gene and protein projects, and reads in part as follows: "...if biomedical research and biotechnology are to achieve and deserve major government and private funding comparable to that provided for aerospace, nuclear energy, and nuclear physics, the objectives must be fully comprehensible to the average person, fill a deep-felt need, and be sufficiently broad to encompass most of biology. Man is the most complex entity (thus far discovered) in the physical universe. Although explorations of space and the atomic nucleus have achieved and deserve large scale funding, huge facilities, strong mandates, and continued public attention, the average citizen is more interested in human mysteries: reproduction, development, birth, disease, aging and death. This public interest and concern is reflected in the position of health care as our largest single industry. Assuredly, the exploration of man and the cells of which he is composed will ultimately require the best minds, the most sophisticated technology, and a large and superlative organization. Only two objectives appear to us to offer the possibility of long-term support on the scale required to maintain ...leadership in biomedical research and biotechnology: the complete sequencing of human DNA and the separation, cataloguing, and characterization of all human gene products. The first objective might be called the "plan for man" and the second the "parts list for man". The new knowledge to be gained from global sequencing is breathtaking indeed. The ultimate intellectual challenge and goal of the DNA-sequencing project is to deduce man from the sequence (or show definitively that this cannot be done). Large ... efforts do not generally arise by consensus of the scientists and technologist concerned. Neither space nor nuclear energy programs in their (then) present forms would have been approved by the scientists ultimately involved, if a vote had been taken before the programs were originally established. Hence the fundamental decision to establish a DNA sequencing or Human Protein Index effort is an almost purely political one."

It was not until 1984–5, however, that the possibility of a large scale sequencing program of the human genome began to be discussed on a widerspread basis by the scientific community, and not until 1988 that the Human Genome Organization (HUGO) was conceived by Sidney Brenner and others. Only in 1990 was the Human Genome Project officially launched in the United States with major federal backing and both public and private interest. It is a striking irony that many of the specific organizational proposals of the HPI finally did become a reality, but at the genetic rather than the protein level, with the formal establishment of the Human Genome Project a decade after the HPI was proposed. The concept of a "plan for man" proved more publically attractive than a "parts list for man". It is clear now that the HPI would have achieved relatively limited success had it been implemented in full during the 1980s. There are several reasons for this, not least that the complexity of the problem was underestimated. Initial estimates of the total number of human proteins were as low as 20 000, whereas we now know that over 100 000 probably exist and that the subtle modifications of such as reversible phosphorylations and glycosylations, which are critical to the functional status of proteins, greatly increase the total number.

There were also important technical shortcomings in molecular anatomy as it was practiced in the 1980s, for instance in the lack of standardization of 2-DE, a problem which would have made the pooling of information by different laboratories hard to achieve. This problem was addressed by the invention of immobilized pH gradient (IPG) gels, which nevertheless took more than 10 years to enter widespread use. Furthermore, the full identification, sequencing and characterization of proteins on an industrial scale, recognized at the time as a difficult long-term objective, only became possible in the late 1990s with the availability of genomic data and the introduction of mass spectrometry (MS) into protein analysis. From the viewpoint of the journal it is of interest to note that electrophoresis is central to all of the techniques involved since both DNA sequencing and 2-D protein analyses involve electrophoresis in gels, while MS may be said to consist of electrophoresis in a vacuum.

It was, of course, the confluence of these developments in 2-DE, MS and genomics that catalyzed the phase transition from earlier protein biology into true proteomics in the modern sense of the word. Only with modern proteomics have the goals which the HPI set for itself become tractable. It is interesting as a final historical note, however, that the original 1980 plan for the HPI suggested "the possibility of computer programs that will match up the DNA sequences of isolated genes with those implied by protein structure, and so relate specific proteins to specific genes for which, up to that time, no specific function was known." The introduction of MS into protein analysis and the development of techniques such as peptide mass fingerprinting and tagging have realized this proposal and are central, almost definitive proteomic techniques today, permitting both protein identification and the proteomic annotation of genomic sequence data. Thus not only the goal of characterizing the human proteome, but also a specific intimation of the means by which this might be acheived through the integration of genomic and proteomic data, originated more than 20 years ago and a full decade before mass molecular screening of any kind became a reality in biology.

5 Proteomics: the rebellious child of 2-DE

While genomic efforts gradually coalesced around a few large centers (funded by governments, foundations, and commercial organizations), protein analysis continued to advance in comparatively small organizations. Among these were the laboratories of Denis Hochstrasser in

Geneva (who pursued systematic improvements in 2-D technology, organized international collaborations, and oversaw the development of new software systems for analyzing 2-D gel data), Joachim Klose (who improved 2-D resolution to unprecedented levels in the course of investigating mouse genetics), Julio Celis (who developed extensive protein databases in the area of human cancer), Jim Garrels (who improved gels, software and database concepts, and gave rise to the first 2-D company: Protein Databases Inc, subsequently folded into Bio-Rad), and that of Norman G. Anderson and N. Leigh Anderson at Argonne (which gave rise to the present Large Scale Proteomics Corporation in 1985). Amos Bairoch organized SWISS-PROT, an invaluable and continuing source of protein data. The laboratories of Merill, Hanash, Neidhardt, and many others contributed to a growing array of databases, many of them dedicated to one organism, one organ, or one organelle. It is impossible to do justice here to the many individuals and laboratories who have made important contributions to this field, and who have published to date over 6000 scientific papers.

However, 2-DE alone was not a sufficient foundation for proteomics. Stunning advances in MS finally solved what had been the most pressing problem, which was that of rapid large scale identification of proteins resolved on 2-D gels. And MS brought with it a classic case of “physics envy”: clean, expensive, metal machines that do not like liquids. As a result it is quite difficult to find self-respecting scientists who will admit to enjoying something so messy and wet as 2-DE, and the search is on for a clean “more physical” replacement. Writing obituaries for 2-DE technology has become a popular pastime. Since such a large fraction of graduate students have had experience with the technology and found it tedious and often unreliable on a small scale, it has been thought to follow that some radically different technology is required for protein separation. No one can disagree with this hope. The model sought is that provided by genomics, where off-the-shelf equipment is available, is commercially installed, repaired, and supplied with reagents, all run by readily available semiskilled workers, supervised by experienced sequencers. Proteomics, in contrast, has followed one of the rubrics of big science – if you need it to operate efficiently on even a modest scale, you have to build your own equipment, or have it custom-made on subcontract (as is true of large accelerators and of advanced space vehicles). To organizations contemplating entry into proteomics there is, therefore, a dual problem. If new surpassing technology is in the offing, one is advised to wait. If the objective is data now, then 2-DE is the only present choice. And, for serious studies, a considerable commitment must be made at the outset.

Two points deserve comment. The assumption that old technologies may be quickly superseded may require revision. It is instructive to consider that the use of silver and gelatin occurred early in the history of photography, and that despite large investments aimed at finding something better, nothing superior was found in over a century of effort – and there is nothing better now. The second point concerns 2-D separations generally [34]. The fundamental idea is that the two dimensions chosen be based on separate unrelated parameters. The supply of unrelated parameters applicable to protein separation is limited, and nearly all combinations of them have been explored in the past. One of the key elements in 2-DE was the mating of the first and second dimensions without loss of resolution. If a first-dimensional separation involves collecting fractions, not only must there be a large number of these to retain resolution, but each fraction must then be analyzed separately in the second dimension, leading to a very large number of discrete operations. This issue of mating first and second dimensions without resolution loss explains why 2-D chromatography (everyone’s first thought as a 2-D gel replacement) has so far not worked well. Only MS has, in theory but not in current practice, sufficient resolution to resolve a set of 100 000 proteins linearly, and if such resolutions were achieved, it would be even more difficult to make all the measurements quantitative. Regardless, there is every reason to try.

6 An updated view of the HPI project

Our current plan for the HPI is based on the development of new technologies, and is divided into five parts, all currently underway at Large Scale Proteomics. The first part includes mapping tissues, and regions of tissues to the limits of present dissection methods. The second includes the use of cell separation technologies to allow maps of different cell types to be prepared. The third involves the use of precision 2-D centrifugation to both allow the subcellular location of each protein to be determined, and also to increase by orders of magnitude the number of proteins that can be resolved from one cell type or tissue. The fourth area includes the systematic prepartion of antibodies against each human protein – one of the objectives of the original MAN program of the 1960s. The antibodies will be used to confirm the cell type location of each protein and its intracellular location. And the fifth area includes the development of solid state protein chips so that clinical tests for marker proteins discovered can be prepared. The HPI project as outlined assumes the development and existence of large scale automated 2-DE systems, the details of which are beyond the scope of this discussion.

The sheer throughput capacity required to complete the HPI is enormous, and is vastly increased as precision cell fractionation is included. Furthermore, the HPI is designed to document the proteome of every cell type, not only in the adult human but throughout the developmental cycle, significantly augmenting an already Herculean task. We estimate that in order to realize the HPI as a completed resource, it will be necessary to prepare, run, and analyze on the order of 100 000 gels, each of which will contain upwards of 2000 individual spots requiring quantitation and detailed, post-translational characterization of the proteins represented. Although there have been encouraging developments towards high throughput integrated proteomics systems in several laboratories, "academic proteomics" does not at present have analytical facilities with a capacity approaching that required to implement a project on the HPI's scale. Following the recent model of genome sequencing efforts where companies such as Celera, Incyte and Human Genome Sciences were responsible for obtaining large tranches of sequence data very rapidly through a combination of powerful technologies and industrial scale plant, it is therefore inevitable that the private sector will be a driving force in the HPI. Large Scale Proteomics has during this year completed the development of systems and infrastructure on a scale commensurate with the task on hand, and is now generating data with projected completion of a draft human proteome in the near future.

The involvement of private sector companies in the generation of biological data is not a perfect solution for the protein research community, since almost by definition the data generated must for a time at least be proprietary to enable the company to recoup its expenditure. However, in the current economic and political climate it is undoubtedly the fastest and most effective way to mobilize the major resources necessary to approach such a huge task. Indeed, much of the proteomic data generated to date throughout the world remains privately held by companies that are conducting projects in fields of specific clinical interest to themselves. Clearly, however, there is a need for a worldwide, public-access databank of human proteomics data to which laboratories throughout the world can contribute and subscribe. One proposal is to expand the SWISS-PROT database to include annotated sequence data on all human proteins. This project has also, ironically, been named the HPI, an acronym standing in this case for Human Proteome Initiative, although its originators were not aware of the history of the HPI when they selected this name (A. Bairoch, personal communication). Whether SWISS-PROT, an alternative existing database, or a whole new database is chosen, the need for a global site for human proteome information, and of an executive and secretariat analogous to HUGO, is now manifest.

7 Elaborations on the HPI: expression profiles following perturbation

For the transcriptome and especially the proteome, documentation of how expression varies following perturbation is the most direct and powerful way in which analysis can currently address biological questions. Numerous proteomics and functional genomics projects are in progress to identify disease-specific patterns of expression with a view to uncovering disease mechanisms, markers and pharmaceutical targets. Combined use of expression studies with genetic techniques such as knockout and antisense can shed light on both normal and pathological cellular organization, while perturbation with drugs is central to basic cellular research, to studies of pathogenesis, and to pharmaceutical analyses including mode of action studies and toxicology.

As with the basic analysis of component molecules in the healthy cell, it would be enormously advantageous to compile comprehensive databases of drug effects and disease mechanisms in specific cell types that can be compared directly with the HPI or comparable data for normal cells. This was indeed one of the goals of the initial HPI in 1980 and is being conducted for specific diseases or toxicology programs in selected cell types by many groups. Currently we are compiling two major databases of perturbations in protein expression. The Molecular Anatomy and Pathology (MAP) database is being compiled in a systematic fashion for a range of diseases. One of its primary outputs will be the production of a comprehensive list of pathology related proteins (PRPs, pronounced "perps") for each disease. The database will therefore become a major resource for medical and pharmaceutical research. The Molecular Effects of Drugs (MED) database curates information on the action of a range of drugs on living cells. Its most important use will be as a resource for screening the mode of action and toxicity of novel drug candidates against all the major known toxicological mechanisms.

The identification of molecular profiles in perturbed cell states is achievable on a mass, comprehensive scale using existing technologies. One important proviso is that adquate care must taken to ensure that samples are obtained appropriately and studies are performed on specific cell types. Tumors, for instance, are typically composed not only of cancer cells but of various other healthy cell types, such as macrophages, endothelial cells and fibroblasts, which together form a multicellular network with characteristic properties. To approach a tumor using proteomic or functional genomic techniques, it will ideally be necessary to analyze these cell types independently. This challenge is likely to be the same for the majority of diseases. Although careful separation of cells into individual types is being carried out in studies of normal cells with great care, there has hitherto been less emphasis on cell separation with diseased cells.

The fact that protein level analysis is by definition phenotypic, and thus makes no distinction between genetic, epigentic or environmental factors, may prove to be of lasting significance both theoretically and ethically. Theoretically, we would contend that protein-level organization is the most fundamental level at which to understand diseases and their treatments, since it is at this level that disease processes are primarily manifested, that most drugs act, and that genetic, epigenetic and environmental factors are integrated during pathogenesis. Although the genetic level would appear to be the most logical plane on which to analyze the distribution of disease susceptibility in a population, the protein level is likely to be more informative and may prove the approach of choice not only for pathogenesis but also population studies. Ethically moreover, the identification of diseases and disease susceptibility using protein level analysis may prove to be more acceptable to society than genetic level testing, since there is no *a priori* implication that the positive identification of a disease-related protein change implies either heritability or inevitability. As screens of disease susceptibility move from the academic center into the general clinic and thence into wider society, protein-based methodologies may therefore offer a more palatable approach.

8 After the HPI: Where next for large scale molecular biology?

As we contemplate final implementation of the HPI two decades after its initial proposal, it is appropriate to look ahead to the next phase of large scale biology and consider what further major goals can be achieved with the mass analysis and screening technologies that are either available now or will be available for routine use within the near future. Two streams of effort are becoming clear: one a technological rebirth of proteomics in the guise of diagnostics, and a second which could be called the rebirth of theoretical biology as a legitimate child.

8.1 Proteomics: Massively parallel protein detectors

The current, 2-DE/MS-based paradigm for proteomics will remain a key methodology for analysis of new samples while there is a possibility that they contain previously unidentified proteins. However, for rapid screening and for the development of routine and clinical screening on the basis of previously identified proteins, the emergence of new techniques is inevitable. In particular, protein expression microarrays with comparable ease of use, sensitivity, analytical scope to DNA microarrays are likely to emerge within the next year or two. Such an approach is already possible with microwell plate technology and conventional fluid-handling robots.

The challenge for proteomics will be to develop these technologies not merely in an *ad hoc* style to address specific problems as probes become available for incorporation into microarrays, but to make them capable of yielding definitve data on protein expression for any chosen sample. Making such expression analysis technology truly comprehensive will require the development of a bank of specific antibodies, or other binding molecules (such as RNA aptamers), for every human protein. Techniques for rapidly raising antibodies against any selected gel spot have been in use for decades, but a proteome-wide antibody development program will nevertheless be a major undertaking on a scale comparable to the HPI, and in an intellectual sense the mirror image of that project. The benefits of completing the project, particularly when it is combined with effective microarray technology, will however be immense, particularly in the field of diagnostics. A comprehensive antibody array could also supersede 2-DE in most applications where protein solubility was not an issue, although detailed characterization of molecular variants and measurement of insoluble proteins would still require 2-DE/MS-based proteomics or an equivalent technique.

8.2 Cybernomics

Cell function, largely the domain of proteins, finds no counterparts in everyday experience. The terms we use for new scientific fields should infer analogies, or contain some fragment of an explanation. But there is no system or machine in common experience, analogous to a human cell, which reproduces itself completely, manufactures all of its own parts as and when needed, constantly replaces all of its working machinery, changes its composition to be different machines (cell types) at different times and places, responds in predictable ways to chemical and physical insults, modifies itself to exhibit memory, detects heat, light, and a vast array of chemical changes in the environment. A cell can signal to its fellows and read the signals of others, and, when specific signals are received, can systematically and completely destroy itself. Further, for higher organisms, the functional units are each separately alive, and each contains a complete copy of the genetic plans. If such systems did not exist, it is doubtful that they could be imagined, or, if proposed, that the proposals would be taken seriously. The application of the term "cell", which implies an empty space, arose from early observations on cork. Surely here a new word is required.

Wiener and Rosenbluth [35] realized in 1947 the essential unity of the set of problems centering around communication, control, and statistical mechanics, whether in living tissues or machines, and further realized that progress was hampered by lack of a term for the then emerging field. They proposed cybernetics [35] from the Greek word for steersman, noting that on sailing vessels rudder position must be constantly adjusted. While the prefix cyber- has been widely used and misused today, we return to its original use and propose that a living cell, in all its aspects, be called a cybernome, with each non-DNA component undergoing continuous change and adjustment; and that the science seeking to integrate genomics, proteomics, and the rest of molecular biology, cell physiology and biochemistry into a common framework be called cybernomics. The serious side of this suggestion is this: it suggests a unique dynamic complex system about which we know little, and for which (like nuclear physics) we have no model in everyday life. The exploration of a cybernome requires the full armamentarium of genomics, proteomics, transcriptomics and what has been called metabolomics (the comprehensive characterization of low molecular weight metabolite flows). While we are only beginning to assemble these large data sets, the prospect of a complete genome, near-term availability of a draft proteome, and large amounts of transcript abundance data now make a theoretical attack on cell modeling possible. One may further speculate that, just as the advent of molecular biology followed an invasion of biology by physicists and chemists, an influx of software adepts will be required to catalyze the reformulation of molecular biology as a programing (rather than a simply descriptive) science.

9 Conclusions: genomics, proteomics, cybernomics

In this paper we have considered the prospects for large scale approaches to the challenging problems of biology and medicine in the twenty-first century. It is now 40 years since the MAN program at Oak Ridge was first proposed, and 20 years since the HPI was outlined. In that time only one major large scale project in biology, the human genome project, has been brought to near-completion. The second project, the HPI, will ultimately exceed the Human Genome Project in scale and cost. The relaunched HPI has only recently begun but will shortly produce detailed maps of all human tissues with identifications. There is a need for a human transcriptome project analogous to the HPI, since this would complete the account of information flow through from the genetic to the protein level of organization. We have also discussed some of the possibilities for further major projects comparable to the Human Genome Project and HPI. Once the three cardinal levels of molecular organization have been documented, a number of other programs, all of which lend themselves to mass approaches, are likely to be addressed. For instance, there will be a need for the HPI (and transcriptome project) to describe cell-specific expression, both of genetic polymorphisms and, more importantly, expression profiles in diseased tissues and in the presence of drugs.

The implementation of these projects will require industrial scale resources of a kind that only private companies can mobilize rapidly, which inevitably means that there will be restricted access to the data that is generated. Just as the complete human genome may be first sequenced by private companies, so too the majority of proteomic data will not, in the first instance at least, be freely available in the public domain. This is an unfortunate but perhaps inevitable consequence of the nature of large research programs which, without private sector funding, would either be completed slowly or not at all. Nevertheless, one can expect a continuous passage of data into the public domain in many cases *via* the patent literature. There is in the long term a clear need for international, public-access databases, mangaged by one or more international secretariats, to collate and curate data from around the world on the human proteome and other major projects.

Beyond the completion of databases for human genes, transcipts and proteins in normal, pathological and drug-treated conditions, we considered a number of further major projects and their suitability for mass-scale analysis. Among these, the construction of a full set of human antibodies is an ambitious but achievable objective which forms an integral part of our current HPI plan. Furthermore, the technologies now exist which will allow the full set of human protein interactions and complexes to be identified, and these will enable a map of the cell's regulatory network to be developed. There is no *a prior* reason why these challenges should not be addressed by appropriately scaled elaborations of technologies which either exist or are currently in development.

Ultimately, the compilation of data on all these levels will take us towards a theory of the cell, although major innovations will be necessary to bring empirical and theoretical approaches to regulatory architecture together. The field by which the regulatory architecture of cells is studied and manipulated (which we have called cybernomics) is likely to become the decisive platform from which genomic and proteomic data are integrated to tackle fundamental biological questions from an information-theoretic perspective. The challenge of the future is to put molecular biology back together.

Received 2 September 2000

10 References

[1] Tietz, P., de Groen, P. C., Anderson, N. N., Sims, C., Esquer-Blasco, R., Meheus, L., Raymackers, J., Dauwe, M., LaRusso, N. F., *Electrophoresis* 1998, *19*, 3207–3212.

[2] Weinberg, A. M., *Reflections on Big Science*, MIT Press, Cambridge, MA 1967, pp. 182.

[3] Weinberg, A. M., *Nature* 1999, *40*, 738.

[4] Anderson, N. G., (Ed.) *Nat. Cancer Inst. Monogr.* 1996, *21*, 525 pp.

[5] Anderson, N. G., *Science J.* 1967, *3*, 35–41.

[6] Green, J. G., Nunley, C. E., Anderson, N. G., *Nat. Cancer Inst. Monogr.* 1996, *21*, 431–440.

[7] Green, J. G., *Nat. Cancer Inst. Monogr.* 1966, *21*, 447–467.

[8] Anderson, N. G., Green, J. G., Barber, M. L., Ladd, F. C., *Anal. Biochem.* 1963, *6*, 153–169.

[9] Anderson, N. G., *Nat. Cancer Inst. Monogr.* 1966, *21*, 9–38.

[10] Anderson, N. G., *Methods Biochem. Ana.* 1967, *25*, 271–310.

[11] Anderson, N. G., Willis, D. D., Holladay, D. W., Caton, J. E., Holleman, J. W., Eveleigh, J. W., Attrill, J. E., Ball, F. L., Anderson, N. L., *Anal. Biochem.* 1975, *68*, 371–393.

[12] Anderson, N. G., Willis, D., D., Holladay, D. W., Caton, J. E., Holleman, J. W., Eveleigh, J. W., Attrill, J. E., Ball. F. L., Anderson, N. L., *Anal. Biochem.* 1975, *66*, 159–174.

[13] Anderson, N. G., *Am. J. Clin. Path.* 1970, *53*, 778–785.

[14] Anderson, N. G., Cline, G. B., Harris, W. W., Green, J. G., in: Berg, G. (Ed.), *Transmission of Viruses by the Water Route*, Interscience Publishers, New York 1967, pp. 75–88.

[15] Harris, W. W., Anderson, N. G., Bartlett, T. W., Rutenberg, E. L., McCauley, L. L., Kniseley, R. M., *Nat. Cancer Inst. Monogr.* 1966, *21*, 389–394.

[16] Anderson, N. G., Harris, W. W., Barber, A. A., Rankin Jr. C. T., Candler, E. L., *Nat. Cancer Inst. Monogr.* 1966, *21*, 353–383.

[17] Price, C. A., *Centrifugation in Density Gradients*, Academic Press, New York 1982, pp. 430.

[18] Anderson, N. G., Waters, D. A., Nunley, C. E., Gibson, R. F., Schilling, R. M., Denny, E. C., Cline, G. B., Babelay, E. F., perardi, T. E., *Anal. Biochem.* 1969, *32*, 460–494.

[19] Reimer, C. B., Baker, R. S., Van Frank, R. M., Newlin, T. E., Cline, G. B., Anderson, N. G., *J Virol.* 1967, *1*, 1207–1216.

[20] Cho, N., Barringer, H. P., Amburgey, J. W., Cline, G. E., Anderson, N. G., McCauley, L. L., Stevens, R. H., Swartout, W. M., *Nat. Cancer Inst. Monogr.* 1966, *21*, 485–502.

[21] Klose, J., *Humangenetik* 1975, *26*, 231–243.

[22] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[23] Scheele, A. G., *J. Biol. Chem.* 1975, *250*, 5375–5385.

[24] Special Issue: Two-Dimensional Gel Electrophoresis, *Clin. Chem.* 1982, *28*, 737–1092.

[25] Special Issue: Two-Dimensional Electrophoresis and Protein Mapping, *Clin. Chem.* 1984, *30*, 1897–2108.

[26] Anderson, N. L., Anderson, N. G., *Anal. Biochem.* 1978, *85*, 341–354.

[27] Anderson, N. G., Anderson, N. l., *Anal. Biochem.* 1978, *85*, 331–340.

[28] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P., Anderson, N. G., *Clin. Chem.* 1981, *27*, 1807–1820.

[29] Anderson, N. L., Edwards, J. J., Giometti, C. S., Willard, K. E., Tollaksen, S. L., Nance, S. L., Hickman, B. J., Taylor, K. E., Coulter, B., Scandora, A., Anderson, N. G., in: Radola, B. J. (Ed.), *Electrophoresis '79*, Walter de Gruyter, New York 1980, pp. 313–328.

[30] Anderson, N. G., Abajian, V., Anderson, N. L., Johnson, I., McConkey, E., McDermott, W., Neel, J. V., Thomas, S., Whitehead, E. C., *Report of the Human Protein Index Task Force*, available from Large Scale Proteomics Corp., Dec. 29, 1980, Rockville, MD.

[31] Anderson, N. G., Anderson, N. L., *Clin. Chem.* 1982, *26*, 739–748.

[32] Anderson, N. G., Anderson, N. L., *J. Autom. Chem.* 1980, *2*, 177–178.

[33] Anderson, N. G., Anderson, N. L., *Am. Biotechnol. Lab.* 1985, Sept/Oct., pp. 1–3.

[34] Wankat, P. C., *Sep. Sci. Technol.* 1984-5, *19*, 801–829.

[35] Wiener, N., *Cybernetics or Control and Communication in the Animal and the Machine*, MIT Press, John Wiley and Sons, New York 1962, p. 11.