REPORT OF THE

HUMAN PROTEIN INDEX TASK FORCE

DECEMBER 29, 1980

EXECUTIVE SUMMARY

Proteins, precisely constructed from coded information in DNA, are the functional machinery of all cells. Most human diseases, and the degenerative aspects of aging, result from changes in the function, amount, or location of proteins. The number of different proteins is quite large, and estimates of this number range from 20,000 to 50,000. Only a few percent have beeen isolated and studied, and only a fraction of one percent can now be measured clinically.

Recent technical advances make it feasible to begin to construct a **Human Protein Index** (HPI) which would be a library of information about all human proteins, and would enable us to identify among the entries those proteins causally related to each of the thousands of human diseases. The key to this effort is a new procedure, two-dimensional gel electrophoresis (2-D gel) that makes possible the simultaneous measurement of charge, mass and quantity of a thousand or more proteins from a single sample of body cells or fluids.

When combined with techniques of genetic engineering, monoclonal antibody production and automated clinical testing, this advance makes the HPI possible. Association of specific protein changes with human diseases can be deduced as well as strategies, products and procedures for the prompt diagnosis and treatment of them.

While the completion of the index will take some years, sets of proteins that behave most reproducibly can be defined at the outset for immediate use.

A program to accomplish these goals is suggested and crucial to it is a central coordinating activity that takes advantage of existing scientific research mechanisms but also requires:

* cross compatability of data between laboratories

* development and provision of standards of purity and identity

* acquisition and sharing of data by a common computer program

* collaboration on technical improvements

* adoption of common terminology

Not only should the new information be of incalculable benefit to our understanding of normal and disease states, but also could lead to a new family of commercial products and services such as enhanced diagnostic and treatment modalities and more sensitive and meaningful tests for the monitoring of environmental and toxicological effects of chemicals.

This program initiative could be accomplished with funding in the range of $150 million over an 8–10 year period and in our opinion ranks as an outstanding and affordable opportunity.

REPORT OF THE HUMAN PROTEIN INDEX TASK FORCE

EXECUTIVE SUMMARY

- FOREWORD

A. INTRODUCTION

- 1. General
- 2. The New Technology
- 3. The Human Protein Index

B. USES OF THE INDEX

- 1. Clinical Medicine
- 2. Monitoring
- 3. Basic Research

C. INTERRELATIONS WITH OTHER TECHNOLOGIES

D. IMPLEMENTATION OF THE INDEX – ITS EVOLUTION

E. TECHNICAL DEVELOPMENT REQUIREMENTS

F. TYPES OF COLLABORATION AND SUPPORT NEEDED

G. ORGANIZATIONAL REQUIREMENTS

H. FUNDING REQUIREMENTS

APPENDIX

- STRUCTURE OF THE HUMAN PROTEIN INDEX

DECEMBER 29, 1980

FOREWORD —

The Human Protein Index Task Force was formed subsequent to a review meeting on the uses of two-dimensional gel electrophoresis technology held in the office of Senator Alan Cranston, The Majority Whip, on August 25, 1980.

At that meeting, attended by representatives of government, science and industry, the potentials of an exciting new technology applied to biomedical sciences were outlined. Such were the implications for rapid advances in the diagnosis, treatment or prevention of human disease using this technique that Senator Cranston requested a report further explicating the rationale and requirements to advance progress in the field.

Dr. Norman Anderson, Argonne National Laboratory, was appointed as Chairman of the Task Force, Dr. Robert Stevenson, American Type Culture Collection, is Secretary-Treasurer. Members of the Task Force were: Vincent Abajian, Electro-Nucleonics, Inc; N. Leigh Anderson, Argonne National Laboratory; Irving Johnson, Eli Lilly Co; Edwin McConkey, University of Colorado at Boulder; Walsh McDermott, M.D., Robert Wood Johnson Foundation; James Neel, M.D., University of Michigan; Stephen Thomas, Foundation for Integrated Biomedical Research, and Edwin C. Whitehead, Technicon Corp.

Providing liason with government agencies were Jeffrey Koplan, Center for Disease Control; William Raub, National Institutes of Health and William Stroud National Aeronautics and Space Administration.

Mr. Daniel Perry, Special Assistant to Senator Cranston, participated in the discussions.

A. INTRODUCTION —

1. General. Proteins, precisely constructed from coded information in DNA, are the functional machinery of all cells. Most human diseases, and the degenerative aspects of aging, result from changes in the function, amount, or location of proteins. The number of different proteins is quite large, and estimates of this number range from 20,000 to 50,000. Only a few percent have been isolated and studied, and only a fraction of one percent can now be measured clinically.

Recent technical advances make it feasible to begin to construct a **Human Protein Index** (HPI), a library of information about all human proteins, and to identify among the entries those proteins causally related to each of the thousands of human diseases. This paper examines the advantages to be gained from creating such an Index, some of the research and development required, and the organizational and support problems to be solved.

2. The New Technology. The key technical advance, 2-dimensional gel electrophoresis, that makes possible simultaneous measurement of 1000 proteins has been developed in the last few years. This is a dramatic improvement over previous methods, which were limited to one or a few proteins at a time. The new technique is a two-step procedure which makes use of the fact that different proteins will move at different rates in response to an electric field. The separations are carried out on thin sheets of a jelly-like material called acrylamide; for these reasons, the short name "2-D gels" is applied to this technique, which will play an important role in gathering the information for the Human Protein Index. The data provided by this technique takes the form of a pattern of spots, each of which represents a protein. Figure 1 shows an example of such a pattern from human lymphocytes, a class of white blood cells. These 2-D gel spot patterns have the important advantage that they can be converted to computer manipulable form for analysis and storage in a data bank. Therefore, if the appropriate technological developments are made, this 2-D gel method can be applied to the measurement of hundreds of human proteins on a clinical scale, making available a new tool of immense power for diagnostic medicine.

3. The Human Protein Index (HPI). The 2-D gels, along with other developments in molecular biology to be mentioned below, now render feasible rapid progress in the construction of a Human Protein Index. This will be a unique catalog of human proteins and their properties. All the known chemical properties of each protein will be listed, along with the normal function of each protein, when this is known. An essential feature of the HPI will be the cataloguing of the relationships of each protein and its variants to human disease. The history of the appearance or disappearance of each protein in normal development and aging will also be determined. Many techniques will be used to acquire this information, such as the 2-D gels referred to above, but the HPI will not depend upon any single technique. Development of the HPI will require a major effort extending over several years, but even in its early stages, it will provide a library of information about human biology that is totally distinct from anything available today. The more that is learned about each protein, the greater will be our understanding of disease and our ability to develop preventive measures or design new therapies that treat the primary cause of each disease. Details concerning the structure of the HPI are to be found in the Appendix.

A major feature of the HPI will be its interaction with efforts in other areas of biotechnology. Genetic engineering now allows (in theory) any human protein to be manufactured in quantity, and hybridoma technology for the production of monoclonal antibodies makes possible the development of specific clinical tests for any single protein. The Human Protein Index will be essential in determining which of the thousands of proteins could be most promptly exploited by these technologies in order to yield medical benefit

B. USES OF THE HUMAN PROTEIN INDEX —

Immediate application of the data in the Human Protein Index to major aspects of human biology can be predicted. The three most obvious areas that will benefit from the HPI are (1) clinical medicine, (2) monitoring programs, and (3) basic research.

1. Clinical Medicine. One of the capabilities of the Human Protein Index Program will be to search for correlations between amounts or properties of a protein and a particular disease. These correlations will be sought with two-dimensional gels and with other techniques as they exist or are developed. When it is found that a change in a particular protein is diagnostic for a particular disease, a simple assay for that protein can be devised. These assays will usually involve antibodies, as in the radioimmunoassays widely used today. The HPI should lead to a large increase in the number of such assays available to the physician.

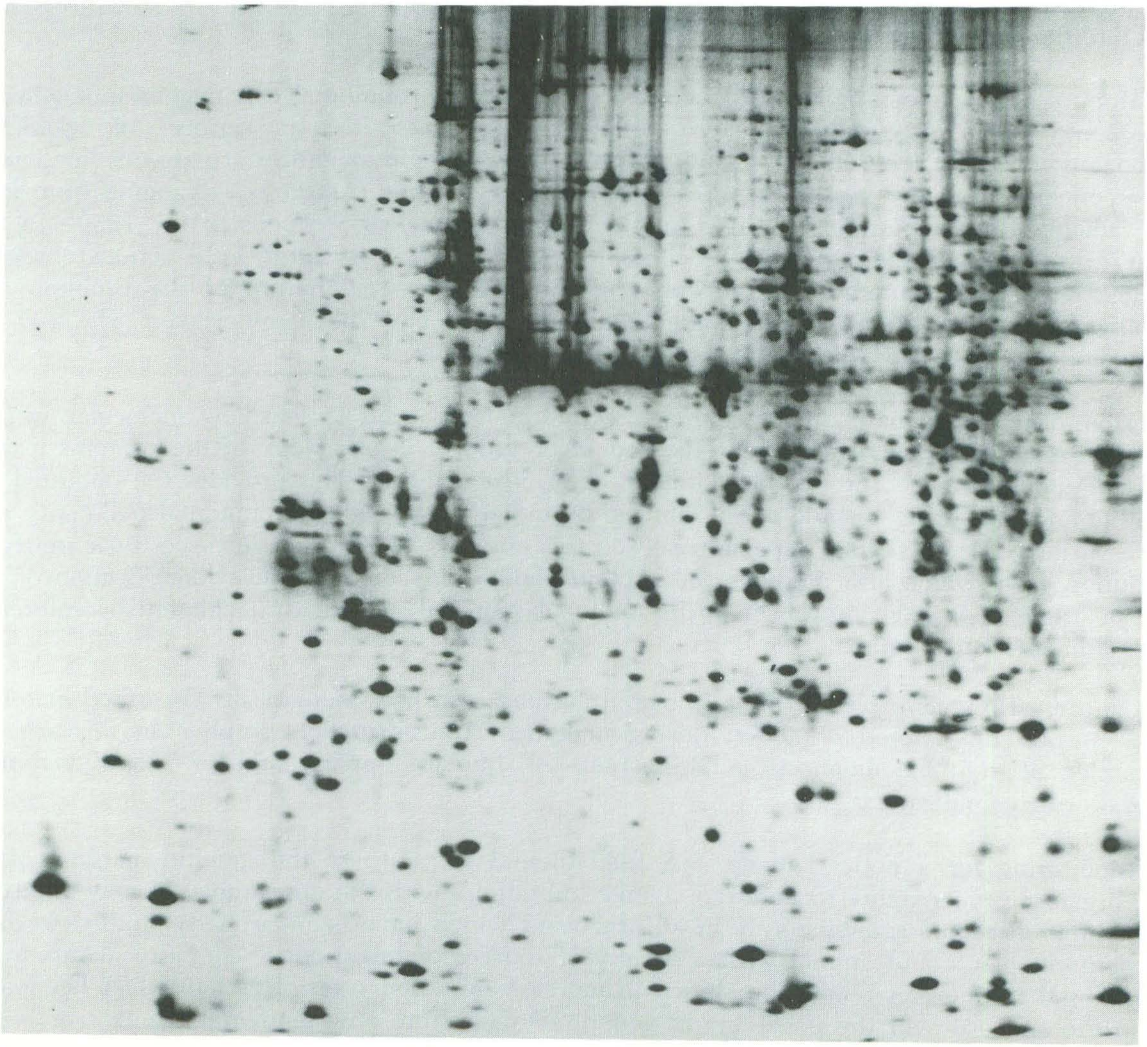


Figure 1. Two-dimensional electrophoretic pattern of the proteins produced by normal human white blood cells. The TYCHO computer analysis system has detected and quantitated 2000 protein spots in this particular pattern.

Additionally, a new type of assay may be anticipated. Modern clinical laboratories are highly automated, compared to their predecessors of a few years ago, and it is not uncommon for 10 – 20 different assays to be performed on one sample of blood. The two-dimensional gel separations that will become routinely available as a result of the Human Protein Index Program will extend this trend dramatically, making possible simultaneous measurement of several hundred (or more) proteins. This, in turn, creates the possibility of finding correlations among proteins which will be diagnostic of disease in its early stages, or the potentiality for disease, permitting physicians to detect a disorder much sooner than can be done with present techniques. If this development occurs, it could result in a schedule of early therapeutic intervention which would significantly reduce the incidence of human clinical illness and the cost of medical care.

We also anticipate an important interface with genetic engineering programs for large scale production of human proteins. As the anomalous proteins associated with various genetic diseases are identified by the program, it is likely that private enterprise will become interested in producing these proteins in quantities large enough for clinical replacement therapy to be attempted. Use of two-dimensional gels to monitor the progress of recombinant DNA projects will accelerate the rate at which success is achieved.

2. Monitoring. a. Screening for the potential effects of environmental carcinogens/mutagens — In recent years, occupational and ambient exposures to a variety of potential carcinogens/mutagens have increased dramatically. Most of these exposures are at very low levels. Public concerns over the possible ill-effects of a variety of these exposures — ranging from low-level radiation resulting from fall-out from the testing of nuclear weapons in the 1950's to the recent Love Canal episode — have reached the point where very large legal actions charging damages have been initiated. Current uncertainty as to the real nature of the problem will probably continue until appropriate epidemiological studies have been undertaken. The 2-D gel technology can play a key role in these epidemiological studies.

b. Monitoring for somatic (tissue) damage – Injured cells tend to leak low molecular weight constituents and protein, regardless of the source or type of insult. Many of the proteins leaked are small enough to pass through the kidney and appear in urine. This provides the potential for detecting and measuring specific organ damage by radiation, burns, environmental pollutants, and a wide variety of toxic agents. Superimposed on this effect are the results of direct toxic injury to the kidney itself, which also modifies the protein composition of urine. The identification of the tissue of origin of leakage protein in urine (or in plasma) will be greatly facilitated by reference to the Human Protein Index.

c. Cancer – How cancer cells differ from normal ones can be systematically described in terms of lists of proteins deleted or added during malignant transformation. Implicit in this are the possibilities of refining diagnostic classifications of human tumors, and of identifying new candidate cancer indicators.

d. Monitoring for genetic damage – A fundamental attribute of the genetic material of all living organisms is its ability to undergo change (mutation), with the appearance of new inherited traits. Mutation occurs spontaneously in all organisms which have been properly studied to date; treatments of plants and animal with a variety of sources of radiation or chemical agents in an experimental setting can readily be shown to increase mutation rates. This latter fact has led to widespread concern over the possibility of genetic damage from environmental agents.

The reason for concern over the possibility of an increase in mutation rates stems from the fact that, in general, a fresh mutation has deleterious effects on its possessor. Although our knowledge is still very incomplete, it is estimated that at least 2% of the population are to some extent handicapped because of genetic disorders. As the infectious and contagious diseases are brought under control, genetic diseases occupy an increasing proportion of medical time, out of proportion to their numerical representation because of their chronicity.

Mutation is a rare event. Present imperfect knowledge suggests that between 1 in 100,000 and 1 in 500,000 human genes of any particular type will spontaneously change in character each generation. The detection of increased mutation rates thus requires rather extensive studies. Two-dimensional gel technology permits the visualization in a single appropriate preparation of as many as 1000 proteins. It appears feasible to develop computer programs suitable for scanning such gels for variant proteins, after which studies of the mother and father of the subject would be undertaken to determine whether the variant of interest was inherited or the result of a new mutation. Preliminary studies suggest that 2-D gels coupled with proper computer programs have the potentiality of an order of magnitude improvement in the efficiency of the screening process, compared to the techniques now in use.

The potential import of this technology on evaluating the possibility of altered mutation rates is not just due to its greater efficiency. In general, the number of children born to persons exposed to a potential mutagen will by experimental standards, fortunately, be small. It is thus important to learn as much as possible from each child. The 2-D gel technology is the most promising development in sight for doing just that, since now it may become possible to derive some hundreds (or thousands) of items of genetic information from each child.

3. Basic Research. Many of the most interesting and fundamental problems in biology can be studied in terms of proteins, their properties, and variability within a species and among species. The Human Protein Index Program will have far-reaching effects on basic research, not only in humans, but on many types of organisms. These effects will accrue because of a) improvements in data processing, b) standarization of techniques, and c) easier identification of proteins.

a) Improvements in processing of data from two-dimensional gels will directly affect researchers studying human biology, who will either be able to use the facilities of the central laboratory to analyze their data, or who will obtain the computer "software" developed by the HPI and apply it in their own laboratories. Investigators studying non-human biology by the same technique, whether for academic or industrial purposes, will also have access to the HPI data processing programs.

b) Standardization of reagents and development of apparatus for automatic or semi-automatic execution of complex protein separation techniques, such as two-dimensional gel electrophoresis, will benefit all investigators using the technique, not only because the results of their own experiments will be less variable, but also because their results will be interpretable by other investigators everywhere in the world. Currently, a large amount of unnecessary duplication of effort occurs because of variations in technique, which cause data from one laboratory to be ambiguous elsewhere.

c) Isolation and characterization of hundreds (and eventually thousands) of human proteins is an essential element of the Human Protein Index Program; it is also an effort unlikely to be attempted elsewhere because of the time and money required. These purified proteins, or the antibodies to them (which will also be generated by the HPI program) will be invaluable to other investigators. In particular, they will be used to create a map, based on two-dimensional gels, with hundreds of known entries. This means that scientists studying hormone action, development, evolution or a dozen other subjects will be able to identify the proteins that vary significantly in each case. This will greatly increase the power of their analyses permitting them to make functional interpretations of complex changes.

By establishing banks of serially collected specimens on individuals it should be possible to study protein changes as a function of age. This type of data has heretofore been unavailable but should be very valuable in studying the aging process.

C. INTERRELATIONS WITH OTHER TECHNOLOGIES —

The analysis of complex protein mixtures by 2-D gels is one of the most recent in the series of remarkable advances which have characterized molecular biology during the recent past. Furthermore, the full realization of the prospects engendered by this new technology will draw extensively on all the other important technologies of molecular biology. Thus, the biochemical characterization of most of the protein moieties visualized with 2-D gels requires isolation of at least microgram quantities. This can be accomplished through the use of the new hybridoma technique for generating monoclonal antibodies. These antibodies, acting on cell lysates prepared in large volumes, will result in specific reversible protein-antibody complexes, from which sufficient quantities of the protein for characterization can be recovered. The actual characterization of the protein then utilizes such techniques as protein finger-printing and high pressure liquid chromatography. The former technique yields the material for amino acid sequence analysis. The latter technique allows rapid characterization of a number of important properties of the protein.

Currently, techniques are being developed to permit characterization of small amounts of protein *in situ* on the 2-D gels that will further short cut this process. Once an amino acid sequence is established, one can recognize the nucleotide sequence of the DNA coding for a protein. There is currently great interest in establishing "libraries" of human genes, the vast majority still of unknown function, for many of which detailed DNA sequences will become known in due time. From these developments emerge the possibility of computer programs which 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 function was known. The ultimate end product of this possibility is a genetic map of the chromosomal location of each of the portions of the total DNA which is translated into a protein equivalent. Otherwise stated, given the DNA-sequencing activities already underway, realization of the Human Protein Index is almost automatically synonymous with stunning advances in our knowledge of the organization of the human genetic material.

D. IMPLEMENTATION OF THE HUMAN PROTEIN INDEX —

1. Evolution and Accessibility of the Human Protein Index Database. The information in the database must be easily accessible to users in the research community and medicine. This requirement can be met in two ways: (1) by following the example provided by the on-line computerized literature search services (such as Lockheed Dialog or MEDLARS) so that individuals can interrogate the current database using telephone-type data links, and (2) by periodically publishing the protein maps and spot annotations in book form. Both methods would be useful (the first being more current, the second cheaper) and both are easily within current technology.

Since a large proportion of the new information concerning specific index entries will come from the research and medical communities at large, it must be possible for members of these communities to add their observations to the database. This raises several problems: 1) it must be possible to record both the origin of specific findings and the experimental means by which they were obtained, so that other users can properly attribute and critically evaluate the data. 2) There must be a mechanism for checking added information to ensure that the contents of the database are self-consistent and accurate. Perhaps the best way to maintain the integrity of such a database is to build it in a hierachical fashion: there would be (1) a core of generally accepted information maintained and updated by a team at the central facility, (2) a catalog of data and observations which has been reported by individuals or groups but not yet confirmed by others and hence considered tentative, and (3) private catalogs maintained by researchers, for their own working use, (perhaps in their own computers but in a format consistent with the main database) and ultimately to be added to the central database. Such a system would be supplementary to the normal scientific literature in the sense that its goals would be narrower (distribution of data concerning specific proteins in the index) and it could be much faster.

2. Interfacing with Current Scientific Efforts. The HPI is not a concept that supplants current scientific directions but is rather a supplement that provides a focus or integration to make these results more meaningful through extension.

The sequencing of proteins will go on as before; biochemistry and metabolic studies will be as necessary but these will be done on highly purified material not heretofore available.

Organization and retrieval of the collected information will be enhanced and within this framework a common language will evolve that will enable all disciplines to share and interpret their data.

E. TECHNICAL DEVELOPMENT REQUIREMENTS —

In order to speed up the data accumulation, it will be necessary to standardize and automate the procedures. Technically this is feasible, but it requires concerted developmental effort.

Areas for improvement include:

* Thin film technology for preparation of gels
* Protein staining
* Rapid production of monoclonal antibodies for spot recognition
* Quantitation of protein spots on gel patterns by computers
* Faster and more accurate comparison of gel patterns by computers.

None of the above are dependent upon new basic concepts. What is required is a coordinated effort to search out and support developmental efforts on these techniques as applied to the goals of the Human Protein Index.

It is expected that a 10–100 fold increase in data acquisition rates can be achieved in the next few years by improvements in the technology.

F. TYPES OF SUPPORT AND COLLABORATION REQUIRED —

The scientific community is already using 2-D gels extensively and should be encouraged to develop standard procedures and techniques so that their data will be compatible with those in the data bank.

The establishment of earmarked program funds for investigator-initiated grants in this area might be an appropriate way to stimulate research. A proviso of the grant to assure adequate standards for generated data plus the researchers ability to query the data base in a private manner should result in enthusiastic cooperation.

In line with this utilization of the conventional peer review process, it might be desirable to establish a study section(s) in appropriate department/agencies to serve as focal points for the program. As the field develops, commercial development and general availability of equipment, supplies, software and services would proceed apace. Scientists would participate in this process by developing improvements and specifications for needed instrumentation and associated products.

Since a number of diagnostic and therapeutic products should result from this activity a rather large peripheral industry should result. It is beyond the scope of this paper to explore those potentials but no new forms of stimulus are seen to be required to have them come about.

G. ORGANIZATIONAL REQUIREMENTS —

The Human Protein Index will be an evolving, growing body of information As such it will require an organizational framework that provides the following:

- * A focal point for programs and project information
- * A permanent repository for the data base accessible to all
- * A state of the art laboratory to evaluate and select technology options and to develop new ones
- * A secretariat to develop concensus and promulgate standards and terminology
- * A skilled staff to interact with government, academia, and industry in collaborative projects.

1. The Central Laboratory. The concept of the central laboratory with an explicit standardization function is key to the HPI goal. Otherwise, data will not be comparable and there will be no unifying language between investigators. Field demonstration projects (2–7) would be developed to spread the capability and work on priority projects.

2. Collaborative Arrangements. (a) Subcontracting — Prompt transfer of technological improvements to state of the art equipment would best be handled with the central laboratory contracting for such equipment.

Also, preparation of larger quantities of identified proteins, monoclonal antibodies and cell strains might be more efficiently done through subcontracting with specialized firms.

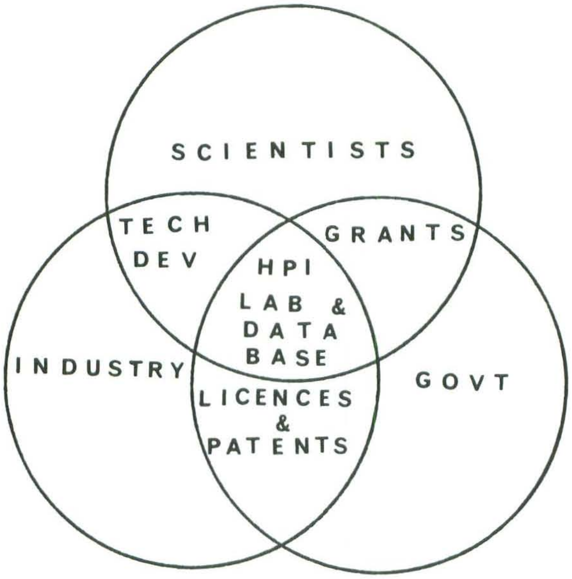
(b) Technical Assistance and Training. The central laboratory should be able to provide training in the use of the technology and be available as a resource to help individual investigators identify proteins. The training requirements and collaborative assistance efforts should be initialized and funded through regular peer review mechanisms.

(c) Private Sector Involvement. It is expected that many companies will develop 2-D gel programs related to their efforts in genetic engineering, diagnostic reagent development, monoclonal antibody production, etc. While these activities will be privately funded, the use of the HPI data bank, technical consultation with HPI staff, callibration studies could be charged on a fee for service basis and such income used to reduce government support requirements.

It would also be expected that process patents generated by the central laboratory would be licensed to the private sector and the licensing royalties would accrue to the HPI.

Use of the Index could be basic to the development of several million dollar markets and generate new tax bases for the government.

The inter-relationships might be diagrammed as follows: (Figure 2)



H. FUNDING REQUIREMENTS —

As noted above, the establishment of a central coordinating laboratory and a home for the data base are basic to the program.

The central lab should have operating funds, capital equipment including state-of-the-art computers and an additional budget for sub-contracting technical improvements on equipment for running 2-D gels.

Provision should be made for obtaining standardized protein preparations and monoclonal antibodies to them so that they are available to collaborating laboratories in the program. Earmarked funds should also be provided to enable peer reviewed satellite demonstration labs to initiate projects of priority interest.

Details of the organizations and funding requirements should be addressed by an appropriate agency of the Government but in our opinion an effective effort could be mounted for $150 million over an 8–10 year period.'

APPENDIX

Structure of the Human Protein Index

Ultimately each of the proteins visualized on 2-D gels will be assigned a specific designation in keeping with biochemical conventions. Clearly, developing a rational system of nomenclature for 20,000 different proteins is a formidable task, not to be undertaken in an outline such as this. On the other hand, clearly from the outset there must be available some interim system for coping with the rapidly accumulating body of knowledge which will ultimately result in the HPI. We suggest that a provisional number may be attached to each protein. This would be a human protein index number, which is arbitrarily assigned, and which would tend to be in order of abundance.

The ultimate data base which is envisioned will include a computer-based storage and retrieval system which will include a number of descriptors of each protein. It will be an objective from the outset to design this base so that the contents of the Index may be retrieved in any order and searched in terms of any descriptor. The following descriptors are proposed for use in the Index.

1. *Genetic map location* – Currently cytogeneticists have developed an alphanumeric system for nomenclature which permits designating delineated regions of each of the human chromosomes. This is an expanding system of nomenclature. It is suggested that for the present, the localization of the structural genes responsible for the various proteins in the HPI be carried no further then the cytogenetic conventions for designating the position of chromosomal breaks. Obviously, in the event that nucleotide sequence maps of the human DNA become available, and the genes coding for specific proteins are assigned to specific positions on that sequence (see Section C), then a new system for designating genetic map location must come into being.

2. *Map Number* – These numbers apply to a particular preparation, and are in approximate abundance order. Map numbers are in italics. Since map numbers will be supplanted by HPI Numbers, which will ultimately be supplanted by CGO Numbers, it is evident that the index will include information on which numbers are the same as others, i.e., which map numbers from different maps have been found to be for the same protein, and which now have HPI Numbers, and eventually, which have CGO Numbers.

3. *Spot location* on "standard" two-dimensional patterns with reference to internal charge and molecular weight standards.

4. *Literature names and function* if known, including enzyme catalog (E.C.) numbers.

5. *Cell types* in which protein is found.

6. *Amount* of each protein found in each gel analyzed, later expressed as amount per cell.

7. *Subcellular localization* of each protein, and how this was determined.

8. *Composition* of the protein including amino acid sequence when known.

9. *Coregulational set* to which the protein belongs and *regulational factors* which can alter the amount or rate of synthesis of a protein.

10. *Genetic polymorphisms* of the protein listed, and their relationship to disease.

11. *Biophysical properties* including thermal denaturation inflection point in the presence and absence of cofactors and substrates, and tertiary structural data.