

HUPO initiatives relevant to clinical proteomics

Sam Hanash

Abstract

The past few years have seen a tremendous interest in the potential of proteomics to address unmet needs in biomedicine. Such unmet needs include more effective strategies for early disease detection and monitoring and more effective therapies, in addition to developing a better understanding of disease pathogenesis. Proteomics is particularly suited for investigating biological fluids to identify disease related alterations and to develop molecular signatures for disease processes. However, much of the effort undertaken in clinical proteomics to date represents either demonstrations of principles or relatively small scale studies when compared with genomics effort and accomplishments or more pertinently when contrasted with the tremendous untapped potential of clinical proteomics. Clearly, we are in the early stages. What seems to be urgently needed is an organized effort to build a solid foundation for proteomics which includes developing a much needed infrastructure with adequate resources. The Human Proteome Organization is fostering an organized international effort in proteomics that includes initiatives around organ systems and biological fluids that have disease relevance as well as development of proteomics resources.

Introduction

Uncovering the wide range of protein alterations in disease states holds much promise for elucidating disease pathogenesis and for developing novel diagnostics and therapeutics ^[1]. To facilitate disease related discoveries, it would be highly relevant to develop a knowledge base of the normal human proteome. Given the enormous complexity of the

human proteome, it stands to reason that no individual institution, be it public or private, would have the needed resources to tackle the human proteome single handedly. The Human Proteome Organization (<http://www.HUPO.org>) came into existence to help identify needed resources and determine achievable objectives to further our understanding of the human proteome. HUPO's stated mission is to consolidate national and regional proteome organizations into a worldwide organization; to engage in scientific and educational activities to encourage the spread of proteomics technologies and to disseminate knowledge pertaining to the human proteome and that of model organisms. To help define needed resources and prioritize objectives, HUPO has organized numerous meetings in North America, Asia and Europe, with participation by government and industry representatives and academicians. As a result, a number of international initiatives have been formulated with goals to develop resources for proteomics and to initiate projects around organ systems and biological fluids that have disease relevance.

The development of resources for proteomics

There is a substantial need for informatics resources for practically every aspect of proteomics ^[2]. The current approaches to the analysis of protein data are highly informal and non-standardized. An important informatics related effort that HUPO is involved in is aimed at developing and adopting standardized approaches that facilitate analysis of proteomics data generated by different laboratories. The HUPO Proteomics Standards Initiative (PSI) was launched with the aim of defining community standards for data representation in proteomics to facilitate data comparison, exchange and validation (<http://psidev.sourceforge.net>) ^[3]. The initial focus has been on protein-protein

interaction data and mass spectrometry data. Although a number of studies have been undertaken to define protein interaction networks by different groups, data integration from different studies is hampered by the fragmentation of the data which resides in various databases and in different formats. The HUPO PSI, with the support of major protein interaction data providers, has proposed a community standard data model for the representation and exchange of protein interaction data. The development of common standards for data exchange in the field of mass spectrometry as well as for other fields of proteomics are in progress (Table 1). HUPO informatics efforts will extend to other aspects of proteomics such as quantitative protein expression analysis. They will also address the needs of proteome projects that target organ systems and biological fluids, with respect to data collection, storage and dissemination (Figure 1).

Table 1 Current agenda for the PSI

A- Mass Spectrometry

1. Controlled vocabulary for M/Z data exchange
2. Converters to convert supported vendor proprietary peaklist formats to
 - a. Standard XML format.
 - b. Standard XML format for search engine results
3. Search engines to be supported
4. XML schema
5. Tools: navigation, visualization
6. How to adapt PSI-MS standard data formats
7. PSI-MS data formats and MIAPE data model

B- Molecular Interactions

1. PSI MI XML schema
2. Controlled vocabularies
 - a. Procedure for maintenance of controlled vocabularies
 - b. Controlled vocabulary for external databases
3. Unified identifiers for interactors
4. Data exchange options

C- Integrating Proteomics Workflows

1. Sample preparation -- Coordination with MGED
2. 2D gel electrophoresis -- data capture, data storage
3. Modeling columns and other separation techniques
4. Integrating mass spectrometry methods and results
5. Supporting bioinformatics analyses

6. Ontology development

An important resource related initiative that HUPO is engaged in is the development of antibodies at the necessary scale for proteome wide investigations. Such a collection of antibodies would allow for example production of antibody microarrays that assay the proteins expressed in a tissue or contained in a biological fluid. The large collection of antibodies, which will be available to researchers, should also facilitate investigations of individual proteins, given that most proteins encoded in the genome do not have corresponding antibodies. The plans under development consist of initially making antibodies directed against the proteins identified as part of the HUPO organ system and biological fluids proteome projects described below and eventually making antibodies to all the proteins encoded for in the human genome. Two major execution sites for antibody production will be located in Europe and in China and additional facilities are currently being explored. A piloting phase for antibody production in China has already begun which has yielded to date some 1,000 antibodies.

HUPO initiatives around organ systems and biological fluids

Much effort has been devoted to identifying a limited number of organ systems and biological fluids that would be the focus of projects to characterize in depth their protein constituents^[4]. The choice is based on scientific and medical relevance as well as interest on the part of governments, industry and funding organizations to commit the necessary funds for execution of the projects. The effort involved in these projects represents an ambitious undertaking that is intended to push the limits of proteomics technologies, through a cooperative international effort. These targeted proteome projects will be

executed in phases commensurate with the availability of funding and the capabilities of available technologies.

HUPO currently has one biological fluid project namely the Plasma Proteome Project (PPP), and two organ system projects namely the Liver Proteome Project (LPP) and the Brain Proteome Project (BPP). Each of these projects has been the focus of numerous activities.

The HUPO PPP

The first proteome project to be implemented is the PPP which is currently in its piloting phase. The scientific objectives of the PPP include:

1. Comprehensive analysis of plasma protein constituents in normal humans in large cohorts of subjects.
2. Determination of the extent of variation in plasma proteins:
 - Within populations in various countries
 - Across various populations from around the world
3. Identification of biological sources of variation within individuals over time and assessment of the effect of:
 - Age, sex, diet and lifestyle
 - Common medications and common diseases

Before the execution of the project would begin on the envisioned scale, a pilot phase is being undertaken. The characterization of plasma proteins is challenging because of the vast dynamic range of protein abundance. The objectives of the pilot phase include a comparison of a broad range of technology platforms for the characterization of proteins in human plasma and serum, in particular low abundance proteins, thus laying the groundwork for future studies of circulating proteins in health and disease. The pilot phase will allow HUPO to objectively assess the resolution, sensitivity, time involved, costs and volumes of sample required with various technologies. Another objective is to

clarify the influence of technical variables in specimen collection, handling, and storage. There is also a need to assess the merits and feasibility of depleting the most abundant plasma proteins and the need for additives for protein stability. To this effect standardized samples have been distributed to some 50 laboratories throughout the world. A breakdown of the various technologies and strategies implemented by participating laboratories is shown in table 2. Another objective is to begin to establish international collaborations for later-phase characterization of the normal human plasma proteome in various ethnic groups.

Table 2 Strategies for serum/plasma protein analysis being tested in the pilot phase

A- Sample preparation strategies

- Enrichment for low abundance proteins (fractionation, affinity capture...)
- Protein tagging

B- Profiling technologies

- Separation based
 - 2-D/MALDI
 - LC/MS
- Non-separation based
 - Protein microarrays
 - Direct analysis by mass spectrometry

C- Protein detection strategies

- Staining, fluorescence...
- Immune based assays

Funds have been raised by HUPO, with equal contributions from industry and from the National Institutes of Health in the United States, for the coordination of the piloting phase and for the award of modest funds for specific activities, deemed highly relevant to the pilot phase, to some of the participating groups. The pilot phase is expected to end in mid-2004. HUPO is planning to organize a meeting at the completion of the pilot phase

to review the findings resulting from the work undertaken by participating laboratories, which is expected to result in a database of identified proteins in serum and plasma and in publications in scientific journals. The technologies that would be proven to have the greatest merit, will be utilized using standardized protocols, to achieve the scientific objectives of the project. Although currently the biological fluids envisioned are limited to serum and plasma, there is also interest to consider other related biological fluids such as saliva and urine.

The HUPO LPP

The LPP has been the subject of meetings in Beijing, Bethesda and Montreal to formalize the objectives of the project, to assess the availability of needed resources and to identify participating countries and laboratories ^[5]. Although the plasma presents challenges related to the vast dynamic range of protein abundance, it is far less complex than an organ such as the liver, which includes numerous cell types and thus presents much greater challenges with respect to sampling, storage and distribution. An important consideration in the planning for the LPP is the availability of tissue samples for distribution as well as the feasibility of preparing purified liver cell populations for analysis. **The scientific objectives of the LPP are outlined in table 3.** The pilot phase for the LPP consists of analysis of a limited number of liver tissue samples and cultured cells to determine feasibility, as in the case of the PPP.. The importance of the LPP is demonstrated by the decision of the government of China to become a major participant in the project with the allocation of some \$30 M to the pilot phase of the LPP and a projected contribution of some \$200 M for the execution phase. Other countries with important contributions to the project include France, Canada and the United States.

Table 3 Scientific objectives of the LPP

1. Collection and banking of human liver tissue specimens
2. Characterization of the protein expression profile of human liver
3. Elucidation of post-translational modifications of liver proteins
4. Construction of protein interaction maps of human liver
5. Localization of human liver proteins in cellular compartments
6. Development of an antibody bank for human liver proteins
7. Development of an ORF bank for human liver proteins
8. Studies of liver disease with a focus on hepatitis, liver cancer and related pathologies

The HUPO BPP

The BPP has been the subject of several workshops to define scientific objectives and to develop aims for a pilot phase, along similar lines as the LPP [6]. A planning committee has been constituted to formally organize the project, with Germany and the German Proteome Organization playing a major role at the present time, with active participation by the National Institutes of Health in the United States and several other organizations. For example, a Brain Proteomics workshop was organized jointly with the National Institute of Neurological Diseases and Stroke in Washington, DC in December, 2002. The goals were to engage scientists from various disciplines of proteomics and neuroscience to address issues specific to the brain proteome and to identify the opportunities and requirements for the effective use of proteomics in understanding function and diseases of the nervous system. The merits of characterizing all brain proteins as opposed to emphasis on specific compartments such as the synaptosome or on functionally defined sets of proteins that may be involved in signaling or G protein coupled receptors etc are being considered. There is also interest in the comparative analyses of model organisms and human. As for the HLP, the development of resources including cDNA clones, monoclonal antibodies and tissue repositories are an important

component of the project. Several workshops are being planned in 2004 to finalize the organization of the project.

Informatics support for HUPO projects

Plans for informatics support for the HUPO projects are currently being developed with active participation of the European Bioinformatics Institute and bioinformaticians. A proposed infrastructure for collecting data for the various projects, once they reach the execution phase, which is under consideration is shown in figure 1 and table 4. A three-tiered structure consisting of (1) labs generating data, (2) data collection centers (DCCs) performing quality control and file validation, and (3) central repository of the validated data for public access is being considered.

Table 4 A summary of the respective roles of the DCCs and the EBI

Scientific Exchange

In addition to meetings and workshops dedicated to specific initiatives, HUPO has launched an annual proteomics world congress which is currently in its third year. HUPO's first world congress took place in Versailles, France, in November, 2002 and attracted some 1,000 participants. The second congress took place in October, 2003, in Montreal, Canada and attracted some 2,500 participants. This year's congress will take place in Beijing, China, October 25-28, 2004 with numerous sessions related to medical proteomics. The congress will be preceded by an educational program and a series of workshops dedicated to the HUPO initiatives. Although HUPO has been in existence for some three years only, substantial progress has been made in forging partnerships between the public and the private sectors and in uniting efforts in different countries to help build a solid foundation for proteomics that help address unmet needs in biomedicine.

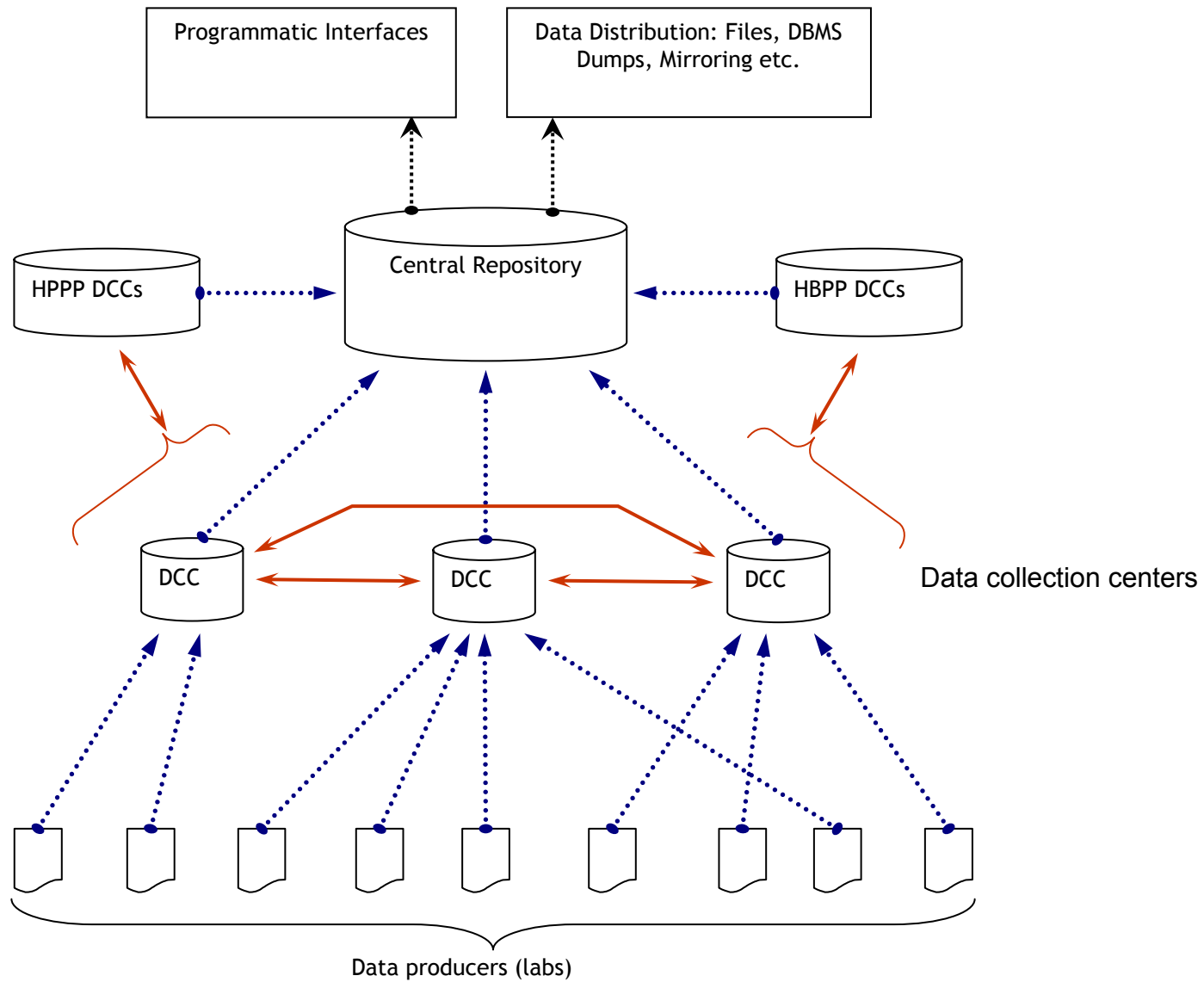
Literature cited

1. Hanash, S. *Nature*. 2003, 422, 226-232.
2. Boguski, M. S. and McIntosh, M. W. *Nature*. 2003, 422, 233-237.
3. Orchard, S., Zhu, W., Julian, R. K. J., Hermjakob, H., et al. *Proteomics*. 2003, 3, 2065-2066.
4. Service, R. F. *Science*. 2003, 302, 1316-1318.
5. Cyranoski, D. *Nature*. 2003, 425, 441.
6. Meyer, H. E., Klose, J. and Hamacher, M. *Lancet Neurol*. 2003, 2, 657-658.

DCC functions:	Central repository functions:
1. Data collection	1. Central proteomics data repository for all HUPO tissue groups
2. Conversion to an xml format, complying with the schema shared by all tissue groups (from EBI)	2. Data access interfaces, and data distribution
3. Synchronise data, in XML, between DCCs	3. Coordination of the development of PSI standards
4. Quality control	4. Coordination of the development of proteomics XML schema integrated with PSI standards
5. Mirror site of PRIDE	5. Providing tools, such as Pedro and peak list converters.
6. Services for data analyses	

Figure Legend

Figure 1. Data collection scheme proposed for HUPO projects showing relationships between the projects, and the central repository. Data producers (labs) submit processed data and results of initial analyses to Data Collection Centers (DCCs) which perform a range of meta-analyses, then pass on the results set to the central repository. The DCCs both within and between HUPO projects are expected to exchange data, methods, standards and techniques, for comparison and meta-analysis.



●.....➔ Data submission

↔ Data exchange