

Advancement of biomarker discovery and validation through the HUPO plasma proteome project

Gilbert S. Omenn

*University of Michigan, A510 MSRB I, 1150 W. Medical Center Drive, Ann Arbor, MI 48109-0656, USA
Tel.: +1 734 763 7583; Fax: +1 734 647 8148; E-mail: gomenn@umich.edu*

Abstract. The Human Proteome Organization (HUPO) Plasma Proteome Project has mounted a Pilot Phase focussed on key problems essential for standardization of specimen collection, specimen handling, choice of fractionation and analysis technologies, and search engines and databases for protein identifications. This international collaboration will lay the groundwork for many large-scale clinical and epidemiological studies of health and disease.

Keywords: Plasma proteome, human proteome organization (HUPO), reference specimens, protein databases/data repository, biomarkers

1. Introduction

Proteomics is maturing into a powerful approach for comprehensive analyses of disease mechanisms and disease markers [1,2]. Nevertheless, applications of proteomics to clinical diagnosis and drug discovery have been described as “in a state of flux” due to many factors, ranging from sample preparation and storage to varied, evolving technological platforms for analysis to non-standardized protein identification and annotation [3]. Organized efforts to accelerate biomarker discovery, quality assurance, validation, and specification of test performance (sensitivity, specificity, and positive predictive value) in screening scenarios, include the National Cancer Institute’s Early Detection Research Network [4].

The Human Proteome Organization (HUPO) was formed in 2001 to enhance development of the field of proteomics through international collaborations in research and education [5]. In a short time, HUPO has organized several major initiatives, of which the Plasma Proteome Project (PPP) is the most salient for biomarker discovery. The PPP has three long-term goals:

1. Comprehensive analysis of the protein constituents of human plasma and serum
2. Identification of physiological, pathological, and pharmacological sources of variation within individuals over time, leading to validated biomarkers
3. Determination of variation across individuals and across populations due to genetic, nutritional, environmental, lifestyle, cultural, and other factors.

Blood samples have the special advantages of being highly accessible sources of new human specimens, of being available in already-existing large specimen banks, and of capturing proteins shed or secreted by cells in every organ of the body. With the power of modern proteomics methods to identify and directly or indirectly quantify large numbers of proteins simultaneously [6–8], it is now feasible to explore and characterize protein molecular signatures diagnostic of specific diseases and treatment responses [2,9]. Hopefully, this work will overcome the recent paucity of new blood tests with high clinical value [10].

2. Planning for the HUPO plasma proteome project: Pilot phase

Highly interdisciplinary groups of experts from academe, government, and industry proposed at HUPO planning meetings in April 2002 and September 2002 and then the First World Congress on Proteomics in Versailles in November 2002 that a Pilot Phase be organized to address the following scientific and practical issues:

1. What standard operating procedures for collection, handling, storage, and thawing of specimens are best suited to proteomic analyses with various technology platforms?
2. How should one choose whether to collect serum or plasma? If plasma, which anti-coagulant (citrate, EDTA, heparin)?
3. What is the sensitivity of various techniques to detect and identify proteins over the huge dynamic range of concentrations in the circulation?
4. What are the advantages and disadvantages of specific methods of depleting or pre-fractionating the most abundant several proteins?
5. How should proteins that are visualized and identified be enumerated and categorized, with special attention to biologically significant post-translational modifications and tissue of origin?
6. Which leads to larger numbers of confidently identified proteins, separation of intact proteins or digestion first, then separation of peptides before mass spectrometry?
7. What are the advantages and limitations of gel-based versus liquid phase multi-dimensional separation methods? Or specific chemical labeling to create subproteomes?
8. How can separation methods best be linked with mass spectrometry to achieve high throughput needed for population-based epidemiological studies and clinical trials?
9. What are the comparative costs of various separation and analysis schemes?
10. How can comparability of datasets and interoperability of data mining be achieved to facilitate reliability and reproducibility of biomarker testing? The need for standardization in all aspects, including the database development, formatting of data submissions, and exchange and analysis of datasets was highlighted in our planning.

3. PPP pilot phase activities (see Table 1)

3.1. Development of reference specimens

In order to compare the attributes of various technology platforms, it is essential to have reference specimens. The options considered ranged from a potential single individual to the vast American Red Cross donor pool. A custom-tailored plan for the PPP was developed with BD Diagnostics and the Chinese Academy of Medical Sciences. A special meeting on needs for standardization in proteomics convened by the US Food and Drug Administration Division of Biologics in January 2003 identified an existing source at the UK National Institute for Biologics Standards and Control (NIBSC). Given the lack of a standardized protease inhibitor cocktail and anecdotes about interferences arising from such proteins and chemicals, we chose to omit use of protease inhibitors, leaving this matter for special analyses later.

The following reference specimens were obtained after informed consent, prepared according to protocol, and distributed globally on dry-ice to participating laboratories:

- a) Lyophilized citrated plasma (1 ml ampoules) prepared from 25 donors by the UK NIBSC for the International Society for Thrombosis and Hemostasis Standards Committee.
- b) Three sets of four frozen reference specimens, prepared for PPP by BD Diagnostics from male/female pairs in three different ethnic groups (Caucasian-, African-, and Asian-Americans), provided as four 250 μ l aliquots of each, with serum, citrated-plasma, EDTA-plasma, and heparinized-plasma in each set.
- c) A similar set of serum and the three plasmas from Chinese in Beijing prepared by the Chinese Academy of Medical Sciences.

In order to create a quantitative calibration for a subset of proteins in these serum and plasma specimens, the PPP obtained quantitative immunoassays of 39 proteins from Frank Vitzthum and Harald Ackermann at DadeBehring, Inc. in Germany and Daniel Chan at Johns Hopkins, and additional assays of proteins and non-protein analytes from Stanley Hefta at Bristol Myers Squibb. These results cover seven orders of magnitude in concentrations. The identity of these proteins, it was agreed by participants, would be retained centrally until participating laboratories have submitted their protein IDs and abundance estimates.

Table 1
Aims for pilot phase of PPP

1.	Compare a broad range of technology platforms for the characterization of proteins in human plasma and serum. Assess resolution, sensitivity, time, cost, volumes of sample required, and practicality with reference specimens.
2.	Clarify influence of various technical variables in specimen collection, handling, and storage, especially anti-coagulation and plasma vs. serum.
3.	Determine whether the most abundant plasma proteins should be depleted, and whether anti-protease cocktails are necessary or desirable.
4.	Develop a database structure and repository for HUPO PPP results.
5.	Lay groundwork, through evaluation of technology platforms and specimen handling and through established international collaborations, for studies of plasma or serum biomarkers in health and disease across major ethnic groups.

4. Development of a collaboratory of participating laboratories

A standard HUPO PPP Questionnaire was sent to all established proteomics laboratories whose investigators had expressed interest in participating, either at the workshops and World Congress or after learning about the PPP through colleagues, the HUPO website (www.hupo.org), or press coverage. Altogether, 47 leading laboratories in 13 countries proposed to participate, utilizing their present and emerging techniques, seeking maximal identification of proteins in one or more of the reference specimens, and sharing all data collaboratively. Of these laboratories, 28 are in the United States (17 academic, 6 federal government, 5 corporate) and 19 are in other countries, including 7 in Europe, 1 in Israel, 9 in Asia, and 2 in Australia; 41 requested the NIBSC specimen, 43 the BD Caucasian-American, 18 the BD African-American and Asian-American specimens, and 18 the Chinese specimens. With regard to technology platforms, 31 indicated that they would run 2D gels, 29 liquid chromatography separations, 30 MALDI MS or MS/MS, 15 direct MS/SELDI, 18 protein labeling, and 30 various depletion or prefractionation protocols.

Combinations of technologies are certain to be required in order to move down the concentration range from albumin (40 mg/ml) to cytokines, PSA, and other proteins of interest (pg/ml).

5. Database development

A PPP database development project and repository were created by the European Bioinformatics Institute in Cambridge, UK, and the Bioinformatics Program at the University of Michigan, led by Henning Hermjakob and David States, respectively. The EBI has the lead for Protein Standards and Bioinformatics for all HUPO initiatives. The aim is to have a consolidated inventory of plasma/serum proteins, linked with organ-derived proteomes of other HUPO initiatives and

with Swissprot and other databases maintained at EBI. Specimen-tracking, descriptions of experimental protocols, study-specific database, and cross-checking of submitted data are functions of the PPP Core at the University of Michigan. Data-submission formats were prepared jointly with the Technology and Resources Committee led by Richard Simpson of the Ludwig Institute in Melbourne and the Specimen Handling Committee led by Dan Chan. Investigators were given the option of submitting results in Excel or using XML, which was highly recommended, accompanied by an offer of technical assistance. The International Protein Index (IPI) was chosen as the reference database for search engines.

6. Initial observations

In the run-up to the 2nd World Congress on Proteomics in October 2003, initial data submissions were received and analyzed; additional datasets were submitted by end of December 2003. Preliminary analyses indicate that a total of 12,830 different proteins (different gene accession numbers) have been identified thus far in one or more HUPO PPP reference specimens by multiple labs (12 labs so far have reported 200 or more proteins, and 3 have identified 2,500 proteins to 2300). Of these, 10,000 are in the IPI database (July 2003 version). Others are in RefSeq, Swissprot, or NCBI-NR databases. Only 936 (9%) have been identified and reported by a second laboratory, 274 by a third lab, etc. Multiple explanations surely apply, reflecting different fractionation methods, different analytical platforms, different search engine algorithms embedded in the mass spectrometers, and the commitment of the PPP to push the limits of detection, with risk of false-positive hits when only 2 or 1 peptide sequence is used to generate a protein match. Furthermore, assignments of protein IDs to different proteins in gene families with highly homologous genes may be inconsistent when one or several conserved peptide sequences may be identical.

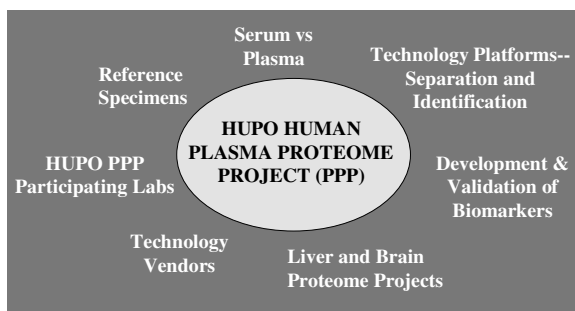


Fig. 1. Scheme showing aims and linkages of the HUPO Plasma Proteome Project.

7. Next steps

During 2004 extensive cross-specimen, cross-technology platform, and cross-laboratory analyses will be completed. The aim is to address as many of the original questions (see above) and specific aims (see Table 1) of the Pilot Phase of the PPP as feasible. Additional special projects have been launched early in 2004 to focus on multiple parameters of temperature, time, and specimen handling for stability; on enhancements of experimental protocols to delve deeper into the proteome; and on elaborate data-mining, spectral analyses, and annotations to enhance the analyses and the usefulness of the data repository. It has become clear that raw spectra from MS/MS sequencing in a variety of high-end instruments need to be compared directly, not just the sequences or m/z peak lists. A special effort is needed to identify specific proteins in SELDI proteomics patterns [11]. We are confident that the HUPO initiatives (Fig. 1) will accelerate biomarker discovery and validation for many important diseases.

We intend to wrap up the Pilot Phase in time for the 3rd World Congress on Proteomics in Beijing 23–27 October, 2003, with extensive publications and open database access shortly thereafter.

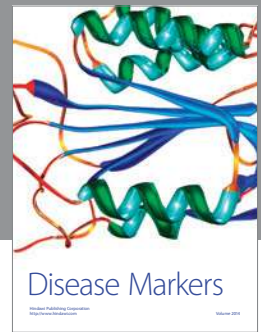
During this same period, we are beginning to plan for the long-term applications of these findings to protocols for population studies of health and disease.

Acknowledgements

The Plasma Proteome Project is supported by a consortium of NIH institutes (NCI, NIA, NIAAA, NIDDK, NIEHS, NINDS) under grant CA84982 and by the following companies: Johnson & Johnson, Pfizer, BD Diagnostics, Abbott, Amersham, Bristol Myers Squibb, Novartis, Proctor and Gamble, Invitrogen, CIPHERgen, and Dade Behring. I gratefully acknowledge the efforts of the HUPO PPP technical committees and participating laboratories and data centers (see www.hupo.org). Work at the University of Michigan is supported in part by the Michigan Life Sciences Corridor (MEDC-238).

References

- [1] R. Aebersold and M. Mann, Mass spectrometry-based proteomics, *Nature* **422** (2003), 198–207.
- [2] S. Hanash, Disease proteomics, *Nature* **422** (2003), 226–232.
- [3] M. Tyers and M. Mann, From genomics to proteomics, *Nature* **422** (2003), 193–197.
- [4] S. Srivastava and R. Gopal-Srivastava, Biomarkers in cancer screening: a public health perspective, *J Nutr* **132** (2002), 2471S–2475S.
- [5] S. Hanash and J.E. Celis, The Human Proteome Organization: a mission to advance proteome knowledge, *Mol Cell Proteomics* **1.6** (2002), 413–414.
- [6] J.N. Adkins, S.M. Varnum, K.J. Auberry, R.J. Moore, N.H. Angell et al., Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry, *Mol Cell Proteomics* **1.12**, (2002), 947–955.
- [7] R. Pieper, C.L. Gatlin, A.J. Makusky, P.S. Russo, C.R. Schatz et al., The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins, *Proteomics* **3** (2003), 1345–1364.
- [8] N.L. Anderson, M. Polanski, R. Pieper, T. Gatlin, R.S. Tirumalai et al., The human plasma proteome: A non-redundant list developed by combination of four separate sources, *Mol Cell Proteomics* **3.4** (2004) 311–326.
- [9] E. Petricoin and L. Liotta, Clinical proteomics: application at the bedside, *Contrib Nephrol* **141** (2004), 93–103.
- [10] N.L. Anderson and N.G. Anderson, The human plasma proteome: history, character, and diagnostic prospects, *Mol Cell Proteomics* **1.11** (2002), 845–867.
- [11] L. Zhang, W. Yu, T. He, J. Yu, R.E. Caffrey et al., Contribution of human alpha-depletion 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor, *Science* **298** (2002), 995–1000.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

