The Human Proteome Organization (HUPO) and Environmental Health

B. Alex Merrick

Proteomics Group, National Center for Toxicogenomics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

The Human Proteome Organization, or HUPO, was formed to promote research and large-scale analysis of the human proteome. By consolidating national proteome organizations into an international body, HUPO will coordinate international initiatives, biological resources, protocols, standards and data for studying the human proteome. HUPO has identified five key areas to advance study of the human proteome, specifically in bioinformatics, new technologies, the plasma proteome, cell models, and a public antibody initiative. Consideration of three major issue areas may help develop HUPO's strategy for human proteome study. First is the need to distinguish the value of high throughput platforms from discovery platforms in proteomics. Second is the importance for international planning on integrating both transcriptome and proteome data and databases. Last is that effects of the environment from chemical, physical, and biological exposures alter the expression and structure of the proteome, which become manifest in long-term adverse health effects and disease. Environmental health research stands to greatly benefit from the shared resources, data, and vision of the HUPO organization as a valuable resource in exploiting knowledge of the human proteome toward improving public health. ***Key words:*** **environmental health, bioinformatics, human, proteome, proteomics, toxicogenomics, toxicology,** ***Environ Health Perspect*** **111:797–801 (2003). doi:10.1289/txg.5918 available via** ***http://dx.doi.org/*** **[Online 18 November 2002]**

The completion of the human genome is a major achievement of the 20th century. The 21st century challenge is to determine the function of the many newly discovered genes and how their gene products interact in pathways and systems to create the human body. An important approach in meeting this challenge in functional genomics is the use of large-scale analyses of the transcriptome and proteome. The human proteome is derived from the gene expression particular to a cell or tissue involving the dynamic and coordinated interaction of proteins in the body. "Mapping the human proteome" (Maher 2002) or describing the complex nature of protein structures, actions, and organizational hierarchies will be very unlike, and much more complex than, mapping the DNA sequence of the human genome. Multiple technologies and international cooperative strategies are being planned to meet the challenge of defining the human proteome and the subtle genetic variations reflected in protein polymorphisms that define each individual. This article summarizes proceedings of a new proteomics organization, comments on its goals and directions for the field of proteomics, and demonstrates why environmental health researchers have a vested interest in the agenda, cooperative studies, and shared resources that will emanate from this organization's activities.

The Human Proteome Organization, HUPO (2002), held an international meeting and workshop at the National Institutes of Health on 29 April 2002 to prioritize goals and standards for large-scale analysis of the human proteome. The mission of the organization is to *consolidate* national proteome organizations into the international body HUPO; to *engage* in scientific and educational activities that promote technologies pertaining to the human proteome and model organisms; and to *assist* in coordinating shared, public proteomic initiatives. The president of the organization is Samir E. Hanash at the University of Michigan. Currently, member countries are linked by three international HUPO divisions: North America, Europe, and Asia–Oceania, with countries from all three divisions participating at the workshop. The two major challenges for HUPO are to identify major opportunies first for international cooperation and second for joint initiatives between public and private sectors.

The HUPO meeting focused on developing specific agendas in five key areas (Figure 1) of human proteomics for immediate international development, chaired by recognized leaders in the field: "Bioinformatics," Rolf Apweiler (University of Heidelberg, Heidelberg, Germany; EMBL, European Bioinformatic Institute, Hinxton, Cambridge, UK); "New Technology," Richard Simpson (Ludwig Institute, Melbourne, Victoria, Australia); the "Plasma Proteome," Gilbert S. Omenn (University of Michigan, Ann Arbor, Michigan, USA); "Cell Models and Tissues," Ronald Taussig (University of Michigan), Cell Signaling Alliance and Pei Pei Ping (University of California, Los Angeles, California, USA); and the "Antibody Initiative," Mattihias Mann (Odense University, Odense, Denmark). A brief discussion of each area follows.

Bioinformatics

The bioinformatics group will define the proteomic data platforms such as 2D (two-dimensional) gels, protein arrays, mass spectrometry, and structural data into a defined infrastructure for data submission and annotation. A major bioinformatics issue is to determine a direction toward either a linked, interoperable consortium of small distributed proteomics databases or the alternative of a large, centralized database. Annotation standards need to be defined using a controlled vocabulary and data confidence measures. Because journals contain many raw and processed proteomic data for potential database incorporation, copyright and accessibility issues need to be resolved.

New Technology

The new technology group had several objectives that included determining lead technologies for best discovering protein interactions, quantifying proteins over a wide dynamic range, fractionating cellular and subcellular compartments to acceptable levels of purity, and identifying housekeeping genes for normalization. The group plans to establish web-based HUPO protocols and make available sets of protein standards that are platform-independent. The group was interested in whether high throughput technologies could be developed to define protein states such as posttranslational modifications, protein conformations, cellular localization, splice variants, covalent modifications, proteoloysis,

This article was previously published in the inaugural Toxicogenomics Section of *EHP*.

Address correspondence to B.A. Merrick, NIEHS, D2-04, PO Box 12233, Research Triangle Park, NC 27709 USA. Telephone: (919) 541-1531 Fax: (919) 541-4704. E-mail: merrick@niehs.nih.gov

I thank M.D. Waters and E.K. Lobenhofer for a critical reading of the manuscript and recognize R.W. Tennant, J.K. Selkirk, and K.B. Tomer for creative discussion and critique. I also thank B. Mills of Image Associates, Raleigh, North Carolina, for the graphics artwork.

Received 31 July 2002; accepted 18 November 2002.

and ligands. A goal was set to identify 5,000 proteins from a specific cell or tissue type to generate an enriched dataset useful for future studies.

Plasma Proteome

The plasma proteome section discussed plans for a comprehensive proteomic analysis of human plasma constituents in the general population and identification of the major sources of variation in the plasma proteome such as age, nutrition, gender, menstrual cycle, exercise, medication, and disease state. There was much discussion about formulating a pooled, multiethnic "reference sample" of plasma or serum to be shared among participating laboratories. Because only a few hundred soluble blood proteins are known, the group plans to more completely identify and catalog plasma proteins with future plans for a plasma proteome database. An informal poll among participants showed a preference of serum over plasma for proteomic analysis. A number of issues were identified, including genetic variation of plasma proteins, liquid-phase multidimensional separation schemes compared with gel-based separation methods, parameters for high throughput links to mass spectrometry (MS), removal of high abundance proteins, and use of antibody arrays or multiplexed enzyme-linked immunosorbent assays (ELISAs) for plasma protein analysis. The question of using pooled or individual plasma samples was raised, as well as the desirability of animal models to help sort the factors accounting for variations in the plasma proteome.

Cell Models and Tissue

The cell models group plans to develop criteria for "attractive" cell and tissue systems that would be widely recognized as a reference type and to recommend specific pilot studies for construction of a cellular proteomic database. Criteria discussed for model cells/tissues included specific organs, amenability to functional assays, high interest to biology or medicine, availability of biological material and funding resources, and ease of identifying proteins. Some participants commented that standardizing protocols might be difficult in a research setting but could be promoted by an international body like HUPO. Existing DNA and protein databases for many nonhuman models were discussed as incomplete, which might make such species less desirable for model adoption. The merits of primary cells were compared with those of immortalized cell lines and stem cells. Also considered was the possibility that several cells or organs might have to serve as models because of the wide range of participant interests. Overall, there was strong sentiment that choice of model and of pilot studies should be a biologically driven decision rather than a technology-driven exercise.

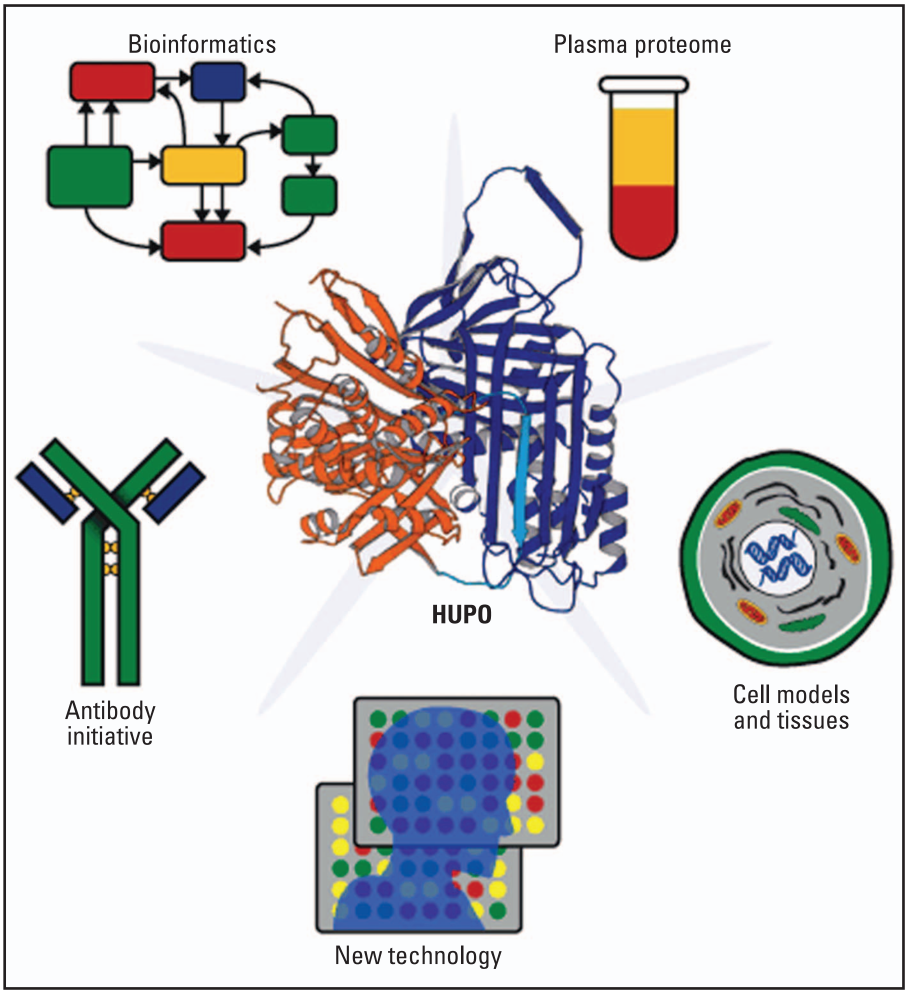

Figure 1. HUPO has developed an agenda in five key areas for international proteomic research. See text for details.

Antibody Initiative

The HUPO antibody initiative gained considerable attention at the workshop because of the pervasive use of antibodies in biological research, the duplication of effort in making antibodies at commercial and research sources, a growing need for a standardized, public antibody collection characterized by application for use, and a widespread desire for developing antibody arrays (Borrebaeck et al. 2001), multiplexed ELISA (Moody et al. 2001), and microfluidic antibody systems (Walter et al. 2002). The intent of such a public antibody collection would be to develop high-quality antibodies for every human protein, with distribution to researchers at a minimal cost. The creation of antibodies was discussed as an internationally funded effort that would likely use bioinformatically chosen peptide antigens to produce polyclonal antibodies as a relatively rapid and inexpensive means to produce a high-affinity product. Selection of the animal host required further discussion ranging from using avian species for noninvasive and high-efficiency collection of IgY antibodies (chicken immunoglobulin derived from the egg yolk) (Tini et al. 2002) to more traditional mammalian species such as rabbit.

Members of HUPO fundamentally agree that describing the human proteome will be vastly more complex than the Human Genome Project but can greatly benefit the scientific community from a cooperative international effort with shared public resources, tools, and data. Each of the five working groups was challenged to identify specific milestones and objectives to be eventually set forth by the leadership of each group. As group agendas become more refined by HUPO, three major issues were identified that should be articulated in the mapping of the human proteome.

Platforms for Discovery and High Information Density in Proteomics

A major goal of HUPO is the mapping of 5000 human proteins. Mapping the human proteome is well recognized as an imperfect analogy drawn from the human genome's assembling of DNA sequences. Yet, a description or map of changing levels of all cellular transcripts and gene products over time under a variety of conditions is an important functional approach to derive meaning from the genome. In this regard, a prime advantage of DNA microarray technology is the relatively large volume of transcript information gained from a single analysis, which could be viewed as a "high information density" technology. The starting source of biological material for microarray studies, RNA, is the same regardless of tissue. The presence of thousands of gene transcripts in isolated RNA are rapidly queried against cloned sources of thousands of sequence-verified genes or synthetic oligomers arrayed on small surfaces by nanotechnologies. The two major platforms, cDNA and oligomer chip arrays, each use easily renewable resources in constructing arrays that are greatly assisted by robotics and automation. The sheer volume of DNA microarray data, or "high information density," its storage, analysis, visual display, clustering, and relation to other microarray data sets are major factors that drive bioinformatics.

By contrast, proteomics has a comparatively greater diversity of platforms that reflect the many properties of proteins to be measured in addition to its primary sequence identity. Starting biological material for proteomics studies often must be freshly isolated by biochemical methods from individual tissues. At present no single proteomics platform can deliver an information density of identified proteins at a level comparable to DNA microarrays, a fact that has hampered development of bioinformatics in proteomics. However, genomic sequence does not predict which proteins interact and how, subcellular localization, posttranslational processing and modifications, or structure and topology of the processed gene product. Further, many signaling processes and pathway cross-talk are not transcriptionally dependent. The challenge in proteomics is to take such unique properties of proteins and to analyze them on a global scale. Many established proteomic technologies such as 2D gel-MS or multidimensional liquid chromatography–MS currently function extremely well as discovery-based platforms capable of linking gene products to function (Gagnon et al. 2002) or localities (Bruno et al. 2002) within the cell (Jung et al. 2000). Recent technologic advances make possible the identification of hundreds of proteins in a single experiment at high throughput commerical facilities and through the release of new technologies like ICAT (isotope-coded affinity tags) (Gygi et al. 1999a) for tandem MS analysis that permit simultaneous detection of high abundance proteins and low copy gene products alike (Hille et al. 2001; Honore 2001).

HUPO has taken a platform neutral stand, not favoring any particular methodology or device, but would like to move forward in achieving its goal of mapping 5,000 human proteins. The exact human cell type and environmental context are still under consideration. However, the rapid development of antibody microarrays (Fung et al. 2001; Haab 2001) may represent an attainable proteomics platform for high-data density comparable to DNA microarrays by using highly parallel detection and quantitation methods for specific proteins from complex solutions. Thousands of antibodies can be arrayed to recognize the primary sequence for identifying specific gene products from tissue lysates or biological fluids. In addition, it eventually will be possible to array antibodies produced to recognize specific posttranslational modifications within a protein that are involved in cell signaling processes critical to cellular response to environmental stress and disease. High-quality antibody libraries as proposed by HUPO will be fundamental in building such arrays and may be the most realistic means for dramatically increasing the data density of proteomics studies (Kodadek 2002).

Integration of the Transcriptome and Proteome

A major challenge for HUPO is development of a strategy to integrate tissue transcript and protein expression datasets. Protein abundance is generally related to mRNA expression for various cellular processes, but initial reports that compared transcript expression and proteomic technologies have suggested that the levels of mRNA and the corresponding gene product were quite different (Gygi et al. 1999b). While some biological conditions such as rapid signaling-dependent responses are well suited to a proteomic approach (Fessler et al. 2002), the higher information density of transcript analysis and ease of validation by reverse transcriptase–polymerase chain reaction are often viewed as primary reasons for use of DNA microarrays in many applications. However, many scientists recognize the advantage of bringing more information to bear on biological problems and have taken a systems biology approach (Griffin et al. 2002) by using both DNA microarrays and proteomics for better hypothesis generation and for constructing biochemical and regulatory pathways (Ideker et al. 2001) in microorganisms (Hecker and Engelmann 2000) and mammalian cells (Hanash 2001). The differing and unique results stemming from transcript and proteomic technologies are often regarded as complementary (Griffin et al. 2002) where differences between these technologies can be resolved by further validation and experimentation.

Because mRNA expression and protein abundance data are significantly more complex and noisy than the underlying genomic sequence information, some researchers have proposed combining the expression data from various data sets of different laboratories into broad functional categories such as composition, function, structure, and localization (Greenbaum et al. 2002). For example, by merging and scaling data sets from yeast into a comprehensive reference set, a substantial agreement has been observed in structural and functional categories (Greenbaum et al. 2002). Careful consideration should be given to performing transcript and proteomic analysis on common tissue from the inception of an experimental study to integrate these data sets. The development of algorithms that analyze different proteomic and transcriptomic datasets from various investigators in the HUPO enterprise will be of great value to the human health research community. A targeted portion of HUPO research could be encouraged for comparative proteomic and transcript studies.

Human Proteome and Environmental Health

A major benefit of describing the human proteome for human health will be its use in biomarker development for disease. There is great interest in discovering new gene products or protein modifications that might serve as biomarkers for cancer, heart disease, neurologic disorders, and many others. One aspect to be eventually examined by HUPO is the interaction and effects of the environment on the proteome. In particular, an understanding of xenobiotic exposures to toxicity and their contribution to human disease are major areas of environmental health research. Great strides are expected in the coupling of protein expression profiles of target tissues to specific cell signaling pathways, transcriptional regulation, structural organization, and systems biology after environmental toxicant exposure. Acquisition of affected and diseased tissues from experimental animals is relatively easy for use in biochemical and molecular studies, but human organ and tissue samples are much more difficult to obtain. Blood, or its derivatives as serum or plasma, is one of the most accessible body fluids that might contain biomarkers indicative of chemical exposure, toxicity, and disease (Kodadek 2002). Transcript analysis can be performed on whole blood from extracted RNA of circulating lymphocytes and macrophages, and may be useful for assessing pulmonary toxicant exposure, some leukemias, and inflammatory conditions. However, changes in blood transcript levels may not always reflect toxic responses for many organs and tissues after systemic chemical exposure.

The changing composition of the serum proteome is more likely to contain informative proteins directly related to toxic responses and disease in the affected organ or tissues (Kennedy 2001). Removal of abundant serum proteins by immunosubtraction methods can greatly enrich for disease-related proteins prior to separation by 2D gels and identification of proteins by MS and has led to the discovery of new serum biomarkers for gentamicin toxicity (Kennedy 2001). Another innovation in proteomics for discovering new serum biomarkers involves the use of SELDI, or surface-enhanced laser desorption ionization, technology (Issaq et al. 2002). Serum proteins are selectively bound to chemically active surfaces on biochips and rapidly scanned to obtain a spectrum of protein masses by a modified MALDI-Tof (matrix assisted laser desorption ionization–time of flight) MS instrument. SELDI produces a more accurate spectrum of protein masses than gel electrophoresis, which is the more conventional but less precise means of separating proteins by mass. Serum mass spectra from different patient groups can be normalized and compared for differences in key clusters of protein masses after SELDI analysis. By using training sets from known normal and cancer patients and then analyzing SELDI data with sophisticated clustering algorithms, discrete protein subsets have been identified from SELDI analysis of serum that are highly predictive of preclinical ovarian cancer (Petricoin et al. 2002), prostate cancer (Adam et al. 2002), and breast cancer (Li et al. 2002). Protein identification from SELDI biochips is actively being developed through use of Tof-Tof (tandem MALDI MS) and other tandem MS instruments specifically adapted to analyze SELDI biochips (Weinberger et al. 2002). Furthermore, detection of protein adducts in blood and serum may also serve as an indicator of chemical exposure from reactive chemical intermediates, toxicity from target organs, and idiosyncratic responses to therapeutics (Farmer 1999; Ju and Uetrecht 2002; Liebler 2002). Proteomic analysis of the human serum and plasma proteomes can yield information on disease processes and chemical exposure that specifically pertain to environmental health.

In summary, the development of an international agenda for research on the human proteome has taken a great step forward by the HUPO leadership. Although there are many opportunities for technologic development in proteomics, HUPO is striving to serve as an international body propelled by thoughtful biological questions in biology, human disease, and environmental health. The twin challenges are in matching the scientific interest, expertise, and funding toward accomplishing the agendas outlined in the five areas of concentration, and equally as important, in developing a vision that is well connected to the large body of knowledge from transcript expression studies and genomic technologies. The shared biological resources, protocols, standards, and data from the HUPO organization will greatly benefit environmental health researchers seeking to move from knowledge of the human proteome to the next level for improving public health.

References

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, et al. 2002. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 62:3609–3614.

Borrebaeck CA, Ekstrom S, Hager AC, Nilsson J, Laurell T, Marko-Varga G. 2001. Protein chips based on recombinant antibody fragments: a highly sensitive approach as detected by mass spectrometry. Biotechniques 30:1126–1130, 1132.

Bruno ME, Borchers CH, Dial MJ, Walker N, Hartis JE, Wetmore BA, et al. 2002. Effects of TCDD upon IkB and IKK subunits localized in microsomes by proteomics. Arch Biochem Biophys 406:153–164.

Farmer PB. 1999. Studies using specific biomarkers for human exposure assessment to exogenous and endogenous chemical agents. Mutat Res 428:69–81.

Fessler MB, Malcolm KC, Duncan MW, Worthen GS. 2002. A genomic and proteomic analysis of activation of the human neutrophil by lipopolysaccharide and its mediation by p38 mitogen-activated protein kinase. J Biol Chem 277:31291–302.

Fung ET, Thulasiraman V, Weinberger SR, Dalmasso EA. 2001. Protein biochips for differential profiling. Curr Opin Biotechnol 12:65–69.

Gagnon E, Duclos S, Rondeau C, Chevet E, Cameron PH, Steele-Mortimer O, et al. 2002. Endoplasmic reticulum-mediated phagocytosis is a mechanism of entry into macrophages. Cell 110:119–131.

Greenbaum D, Jansen R, Gerstein M. 2002. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics 18:585–596.

Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, et al. 2002. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. Mol Cell Proteomics 1:323–333.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. 1999a. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999.

Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999b. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730.

Haab BB. 2001. Advances in protein microarray technology for protein expression and interaction profiling. Curr Opin Drug Discov Devel 4:116-23.

Hanash SM. 2001. Global profiling of gene expression in cancer using genomics and proteomics. Curr Opin Mol Ther 3:538–545.

Hecker M, Engelmann S. 2000. Proteomics, DNA arrays and the analysis of still unknown regulons and unknown proteins of *Bacillus subtilis* and pathogenic gram-positive bacteria. Int J Med Microbiol 290:123–134.

Hille JM, Freed AL, Watzig H. 2001. Possibilities to improve automation, speed and precision of proteome analysis: a comparison of two-dimensional electrophoresis and alternatives. Electrophoresis 22:4035–4052.

Honore B. 2001. Genome- and proteome-based technologies: status and applications in the postgenomic era. Expert Rev Mol Diagn 1:265–274.

HUPO (Human Proteome Organization). 2002. Available: *http://www.hupo.org/* [Accessed 2 December 2002]

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934.

Issaq HJ, Veenstra TD, Conrads TP, Felschow D. 2002. The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem Biophys Res Commun 292:587–592.

Ju C, Uetrecht JP. 2002. Mechanism of idiosyncratic drug reactions: reactive metabolite formation, protein binding and the regulation of the immune system. Curr Drug Metab 3:367–377.

Jung E, Heller M, Sanchez JC, Hochstrasser DF. 2000. Proteomics meets cell biology: the establishment of subcellular proteomes. Electrophoresis 21:3369–3377.

Kennedy S. 2001. Proteomic profiling from human samples: the body fluid alternative. Toxicol Lett 120:379–384.

Kodadek T. 2002. Development of protein-detecting microarrays and related devices. Trends Biochem Sci 27:295–300.

Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. 2002. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 48:1296–1304.

Liebler DC. 2002. Proteomic approaches to characterize protein modifications: new tools to study the effects of environmental exposures. Environ Health Perspect 110 (suppl 1):3–9.

Maher BA. 2002. Towards a global proteome. Scientist 16:29–22.

Moody MD, Van Arsdell SW, Murphy KP, Orencole SF, Burns C. 2001. Array-based ELISAs for high-throughput analysis of human cytokines. Biotechniques 31:186–190, 192–194.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. 2002. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359:572–577.

Tini M, Jewell UR, Camenisch G, Chilov D, Gassmann M. 2002. Generation and application of chicken egg-yolk antibodies. Comp Biochem Physiol A Mol Integr Physiol 131:569–574.

Walter G, Bussow K, Lueking A, Glokler J. 2002. High-throughput protein arrays: prospects for molecular diagnostics. Trends Mol Med 8:250–253.

Weinberger SR, Viner RI, Ho P. 2002. Tagless extraction-retentate chromatography: a new global protein digestion strategy for monitoring differential protein expression. Electrophoresis 23:3182–3192.