

A Human Proteome Project

**Determining and characterizing the proteins encoded by the
human genome.**

AUGUST, 2008

Mathias Uhlen

Department of Biotechnology
AlbaNova University Center
Royal Institute of Technology (KTH) 106 91
Stockholm,
Sweden

Alexander W Bell

Department of Anatomy and Cell Biology
McGill University
3640 University Street
Montreal, Quebec H3A 2B2 Canada

Laura Beretta

Public Health Sciences Division
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, M5-A864
Seattle, WA 98109 USA

Christoph Borchers

Dept. of Biochemistry and Microbiology
University of Victoria
Victoria, BC. V8P 5C2
Canada

Tom Conrads

University of Pittsburgh School of Medicine
Co-Director, Cancer Biomarkers Facility and Mass Spectrometry Platform
Member, University of Pittsburgh Cancer Institute
B401 Magee Women's Research Institute
204 Craft Avenue
Pittsburgh, PA 15213 USA

A Human Proteome Project

Michel Desjardins

Département de pathologie et biologie cellulaire, Université de Montréal,
C.P. 6128, Succ centre ville,
Montréal, Québec, H3C3J7 Canada

Eric W Deutsch

Institute for Systems Biology
1441 N 34th Street
Seattle, WA 98103 USA

Will Dracup

Nonlinear Dynamics Ltd
Keel House, Garth Heads
Newcastle upon Tyne, NE1 2JE UK

Henry Duewel

Protein Technologies and Assays
Sigma-Aldrich
2909 Laclede Ave.
Saint Louis, MO, 63103 USA

Ola Forsstrom-Olsson, CEO

Ludesi
Engelbrektskatan 15
SE-211 33, Malmo
Sweden

Jack Greenblatt

Donnelly CCBR, Room 906
9th floor, Terrence Donnelly
Centre for Cellular and Biomolecular Research
160 College Street
University of Toronto
Toronto, Ontario M5S 3E1
Canada

Rudolf Grimm

Life Science Solutions Unit
Agilent Technologies Inc.
5301 Stevens Creek Blvd.
Santa Clara, CA 95051, USA

David Juncker

Biomedical Engineering Department
and McGill University and Genome Quebec Innovation centre
740 Penfield Avenue

A Human Proteome Project

Montreal, H3A 1A4 Canada

Bonghee Lee

Center for Genomics and Proteomics
Lee Gil Ya Cancer and Diabetes Institute Gachon University of Medicine and Science
7-45 Songdo Dong, Incheon 406-840, Korea

Lennart Martens

Computational Omics and Systems Biology Group
VIB Department of Medical protein Research- B-9000
University of Ghent
Belgium

Helmut Meyer

Medizinisches Proteom-Center
Ruhr-Universität Bochum
Raum 2.053, Bldg. ZKF II
Universitätsstraße 150
D-44801 Bochum, Germany

Tommy Nilsson

The Research Institute of the McGill University Health Centre and the Department of
Medicine, McGill University,
687 Pine Avenue West,
Montreal, Quebec H3A 1A1, Canada.

Gilbert S Omenn

Center for Computational Medicine & Bioinformatics
University of Michigan
100 Washtenaw Avenue Rm. 2017F Palmer Commons Bldg.
Ann Arbor, MI 48109-2218

Young Mok Park

Korea Basic Research Institute
804-1 Yangcheong-Ri,
Ochang-Myun,
Cheongwon-Goon,
Chungcheongbuk-Do 363-883,
Korea

Peipei Ping,

UCLA School of Medicine
Suite 1609/1619 at CVRL
675 CE Young Dr.
Los Angeles, CA 90095-1760 USA

A Human Proteome Project

Fredrik Ponten

Uppsala University
Department of Genetics and Pathology
Rudbeck Laboratory, Uppsala University
S-751 85 Uppsala, Sweden

Jan Schnitzer

PRISM Proteogenomics Research Institute for Systems Medicine
11107 Roselle St
San Diego, CA 92121. USA

William M. Skea

Protein Forest Inc.
128 Spring Street
Building B, Level 500
Lexington, MA 02421 USA

Michael Snyder

Department of Genetics, MC: 5120
Stanford University
300 Pasteur Drive., M-344
Stanford, CA 94305-5120 USA

Sudhir Srivastava

Chief, Cancer Biomarkers Research Group
Division of Cancer Prevention
National Cancer Institute
6130 Executive Boulevard, Suite 3142
Rockville, MD 20852 USA

Tadashi Yamamoto

Niigata University
Institute of Nephrology, Graduate School of Medical and Dental Sciences
Niigata University
Niigata 951-8510, JAPAN

John Bergeron

The Research Institute of the McGill University Health Centre and the Department of
Medicine, McGill University,
687 Pine Avenue West,
Montreal, Quebec H3A 1A1, Canada.

A Human Proteome Project

Table of Content

Project Summary.....	7
Ensuring Data Quality and Reproducibility.....	8
Time Lines and Deliverables for the Pilot Phase of a Human Proteome Project	11
Antibody-Based Profiling	16
Signaling Pathway Profiling	17
Proteomics Based Profiling.....	18
Networks-Based Profiling.....	19
Integrative Bioinformatics	20
Conclusion	21
Drafting of the HUPO proposal	21
Antibody-based profiling (complete program)	22
Executive summary.....	22
Introduction.....	22
Process	23
Overall deliverables (output)	24
Time frame.....	25
Subprograms	25
1. <i>Antigen (protein) production.</i>	25
2. <i>Affinity reagent generation</i>	26
3. <i>Generic antibody validation</i>	26
4. <i>Application-specific validation</i>	27
5. <i>Proteome analysis (using affinity reagents)</i>	27
The demonstration project	28
Technology development.....	29
Synergies with the other HUPO subprograms	29
Clinical relevance.....	30
Proteomic-based profiling.....	30
Executive summary.....	30
Introduction.....	30
Process	33
Ongoing HUPO and other initiatives that feed into the project.....	33
Multiple Reaction Monitoring	33
Posttranslational modifications	34
The Pilot phase.....	35
Full-Scale Phase.....	36
Technology Development.....	36
Post Translational Modifications	37
<i>Protein Phosphorylation</i>	37
Pilot Phase.....	38
Full-Scale Phase.....	38
<i>Protein Glycosylation</i>	38
Human Proteins Interaction Networks.....	39
Executive summary.....	39
Introduction.....	39

A Human Proteome Project

Protein Interactions Networks in Health and Disease.....	41
The Human Proteotheque Initiative (HuPI): Building a repertoire of comprehensive maps of the human protein interaction network	42
Clinical Implications	45
Process	45
Conclusions.....	49
Integrative Bioinformatics	50
Introduction.....	50
Mission statement	50
Process	50
Objectives	51
Implementation	52
Output	53
Synergies with the other HUPO Project components	53
Outcome.....	53
Signaling Pathway Profiling	54
Executive Summary	54
Introduction.....	54
Process	55
Reverse protein arrays.....	55
Applications	56
Technology Development.....	57
Synergies with the other HUPO subprograms	57
Clinical relevance.....	57
Relevance for the Communities.....	58
Relevance to the Biology Communities	58
Clinical relevance.....	58
Stem cell relevance	60
Contingency plans.....	61
Conclusion	61
References.....	62

A Human Proteome Project

Project Summary

The human genome sequence provides the basic blueprint for the 21 000 genes [1] present in all the 219 or so cell types that make up the body, yet what defines each cell's function are localization, interactions and modifications of its expressed proteins. Current and emerging advances in proteomic technology have created an opportunity to understand cellular function through a complete characterization of each of the major isoforms of the 21 000 proteins expressed by the human genome by proteomics.

This project (see Figure 1) will create a resource of immediate applicability to the clinical and basic science communities in a format that assures fundamental discoveries and insight into diagnostic and treatment regimens for the patient. Following the HUPO editorial in Nature [2], the proteomics community has established standards for proteomics based profiling, antibody based profiling and network based profiling. These are quantitative, reproducible methodologies and platforms to assure the completion of the human proteome within 10 years. Taken together with standards developed for biofluid, tissue, organ, cell, and organellar samples and standards for a bioinformatics based interface, the resource will be comprehensive, accurate and responsive to the needs of the biological and clinical communities. The human genome provides a basic blueprint for the genes present in every cell. Genome-wide RNA/DNA array-based studies already offer insights into the expression of each gene in various cell types and tissues thereby providing important inroads to functional and disease-related studies. We now need information about which proteins are expressed at what time and levels in any given cell type. The Human Proteome Project, by characterizing all 21 000 genes of the known genome, will generate the map of the protein based molecular architecture of the human body and become a resource to help elucidate biological and molecular function and advance diagnosis and treatment of diseases.

The effort gathers international strengths in each component of the project. The combination of data from global sources [3] is a strength of the project with standards applied to each component to ensure completeness, accuracy, and permanency.

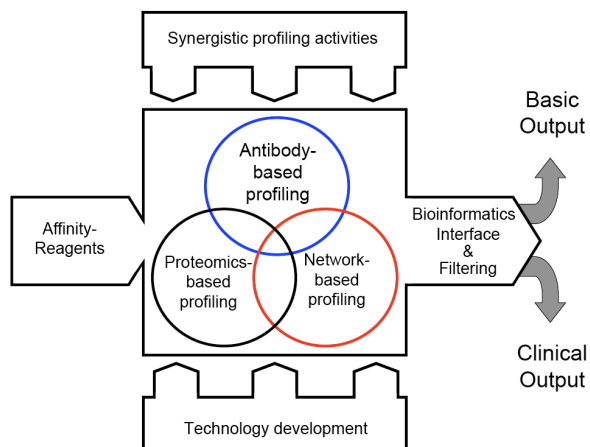


Figure 1. Antibody-based profiling, proteomics-based profiling, and network-based profiling provide complementary data-generating activities that are combined and accessed through a **single** Bioinformatics Interface and Filtering window.

A Human Proteome Project

Ensuring Data Quality and Reproducibility

Proteins must interact with other molecular components of the cell, including other proteins, DNA sequences, RNA molecules and various metabolites in order to exert their function (e.g. enzymatic, signaling, transcription factors, structural). These protein interactions define the physiological state of the cells, organs and tissues. Within the cell, proteins are segregated to specific organelles for specialized functions. Several organs, tissues, subcellular organelles will be targeted for proteomics characterization (Table 1). Preliminary proteomics LC-MS based data is available and these data will be utilized in an initial accounting of the human proteome.

Table 1: Proteome Targets

<u>Organ, Tissue, Cellular and Body Fluid Profiling</u>		
Organs (cell type specific):	Brain*	Heart muscle
	Heart*	Kidney*
	Liver*	Lung
	Pancreas	Skeletal muscle
Tissues:	Brown adipose	White adipose
Cells:	Dendritic cells	Human embryonic cell line
	Lymphoid cells	Macrophages
	Mouse embryonic cell line	Stem cells
Body fluids:	Cerebrospinal fluid* (CSF)	Plasma*
	Serum	Synovial fluid
	Urine*	
<u>Subcellular Organelle Profiling</u>		
Organelles:	Autophagosomes	Cis-Golgi network
	Clathrin coated vesicles	COPI vesicles and sub compartments
	COPII vesicles and sub compartments	Early endosomes
	Endothelial caveolae	Exosomes
	Intermediate compartment	Late endosomes
	Lysosomes	Mitochondria
	Non-clathrin coated endocytic carriers	Nuclear envelope
	Nuclear pore	Nucleoli
	Nucleus	Peroxisomes
	Phagosomes	Plasma membrane (endothelial hepatic)
	Regulated secretory vesicles	Rough endoplasmic reticulum (ER)
	Smooth endoplasmic reticulum (ER)	Stacked Golgi cisternae
	Synaptic vesicles	

*HUPO Projects

HUPO have taken note of worries that have been raised by various parties that data generated by older proteomics projects may not have proved to be as reliable as would be

A Human Proteome Project

desirable. While recent advances in high throughput techniques bring huge improvements in data reliability, HUPO has undertaken major actions to investigate the issue further with the HUPO Test Samples and Reproducibility Test projects [4].

Both of these projects show that high throughput proteomics generates reliable data, providing that there is a strong and robust data matching mechanism and harmonised data presentation.

To ensure the required level of accuracy, reproducibility and comprehensiveness, samples used throughout the project also need to be gathered, stored and distributed in a controlled and coordinated manner. For this, each participating funding agency is required to put in place a national body that oversees this aspect and which coordinates this internationally. At the time of writing, several large-scale profiling initiatives are already on their way aimed at studying extensive cohorts. As in this proteomics project, routines and standard practices are presently being put in place to ensure best possible practices. In this process, biobanks are being harmonized, standardized and quality assured. A human proteome project is in light of such efforts both timely and realistic.

The pilot phase will be used to put in place a set of standard practices that pertain to the handling of samples, the analysis and evaluation of data. Inbuilt redundancies in the overall project plan will also ensure the quality and accuracy of the obtained data. Expert laboratories are here envisaged to play an active role to ensure that standard practices can be put in place well in time for the full-scale phase. This will include application of test samples to benchmark each lab and evaluate statistically inter lab variation. These statistical methods will be employed for biological samples including statistical assessment of inter lab reproducibility for the same biological samples.

The Human Proteome Project involves a comprehensive characterization of the human proteins encoded by the genome, including:

1. Cellular and Organ Profiling – Protein Expression

A variety of antibody-based and mass spectrometry-based strategies will be used to ensure comprehensive coverage of the proteome including high sequence coverage of the identified proteins to determine isoforms and splice variants associated with each cell type, organ and body-fluid selected for the study. For low abundance proteins, the use of targeted proteomics will be implemented. Comprehensiveness will be defined by the characterization first of chromosome 21 (195 proteins), then of all 21 000 genes with chromosome to chromosome milestones.

2. Localization – Subcellular Enrichment of Cells

Sub-cellular enrichment techniques will be used to isolate cellular compartments for proteomic analysis. The need for highly enriched organelles and sub-cellular compartments will drive the development of new cell biological techniques to create highly enriched fractions as well as subtractive mass spectrometry techniques. This will be complemented by the antibody based profiling of organelles in all organs of the body.

3. Protein-Protein Interactions

A Human Proteome Project

Protein-protein interactions will be determined using available antibodies to enrich protein complexes. We anticipate this aspect of the project will drive the development of new methods to determine protein-protein interactions in addition to the creation of tagged protein resources for both pull down experiments and fluorescent localization as was done in bakers' yeast, *S. cerevisiae* [5-8].

4. Molecular profiling - Protein Modifications

The proteins will be characterized with regards to molecular weight, splice variants and proteotypic isoforms both with proteomics-based and antibody-based profiling.

Large-scale identification of modification sites will be accomplished for each cell type especially for phosphorylation and glycosylation.

5. Signaling Pathway Profiling

Signaling pathway activities will be defined at a molecular level in a quantitative manner using reverse protein arrays. The large scale, high throughput approach will include the analysis of normal activities specific to cells and organs, interpathway connectivity and kinetics of signaling.

A Human Proteome Project

Time Lines and Deliverables for the Pilot Phase of a Human Proteome Project

I. Pilot phase for current accounting of the human proteome by bioinformatics based data gathering of LC-MS proteomics profiling data.

A start is envisaged by gathering tandem mass spectra currently available from the global community to map the frequency with which proteins have been characterized on a chromosome by chromosome basis. This will be coordinated with the individual HUPO Proteome Initiative organizers to collect the raw mass spectrometer output files and all the corresponding metadata. This information will be formatted into the PRIDE submission mechanism and uploaded there. The raw data will thereby be deposited into Tranche and the Tranche hash keys made available to PeptideAtlas and GPM. These two resources will download the data via an automated process and process the data through their individual systems, including format conversion, sequence database searching and spectral library searching, statistical validation, and loading into the respective databases. The results will be merged with other existing data present in these resources to form a gene/protein centric view of the identifications that can be browsed by chromosomal coordinates. Results will be made freely available via web-portal that allows browsing on a chromosome basis. As part of this pilot phase, a robust data-matching component will be developed and implemented to ensure 100% accuracy in protein annotation. This will include a database configured in a gene-centric fashion for 21,000 proteins, i.e, a maximum of one for each protein coding gene.

Deliverables:

1. Proteotypic peptides. By visual representation through peptide heat maps of the frequency with which tandem mass spectra of high quality are assigned to tryptic peptides unique (red) to each protein coding gene. This will define experimentally proteotypic peptides mapped to each protein coding gene on a chromosome to chromosome basis. **Timeline: 1 year (interim report at 6 months).**
2. Shared peptides. All other tandem mass spectra will be represented again by heat maps, but in blue, to indicate tryptic peptides not unique to the protein. **Timeline: 1 year (interim report at 6 months).**
3. The degree of coverage of chromosome 21 and all chromosomes is visually represented. **Timeline: 1 year (interim report at 6 months).**
4. A robust data matching component and the annotation of all 21,000 protein coding genes with a descriptive unique name to enable unambiguous matching of tandem mass spectra in order to remove one of the major bottlenecks in proteomics. **Timeline: 1 year (interim report at 6 months).**

II. Pilot phase for Antibody Based Profiling – Chromosome 21

The initial phase of the Antibody Based Profiling Initiative is the gathering of all antibody and affinity reagent information of the 195 proteins of chromosome 21 to a

A Human Proteome Project

designated portal. This site, an initial deliverable will be integrated by the Integrative Bioinformatics group into the Integrative Proteomics Portal. This site will be updated continuously by this group as antibody and affinity reagents are generated for proteins of chromosome 21.

Number of protein targets: 195 (all genes from human chromosome 21)

Deliverables:

1. A public database portal for validated antigens (including standards). **Timeline: 1 year.**
2. A public database portal for validated affinity reagents (including standards). **Timeline: 1 year.**
3. Antigens to a majority of the 195 chromosome 21 encoded protein targets using different strategies (preferably both full-length recombinant proteins and selected recombinant protein fragments). **Timeline: 2 years (assessment at 1 year for progress)**
4. Various types of affinity reagents to a majority of the 195 chromosome 21 encoded protein targets (preferably both multi-epitope and single-epitope affinity reagents). **Timeline: 2 years.**
5. Standard operating procedures for generation of renewable affinity reagents. **Timeline: 1 year.**
6. Results from technical studies aimed to show the feasibility of high-throughput generation of such renewable affinity reagents. **Timeline: 2 years (assessment at 1 year for progress)**

III. Pilot phase for the Protein Interactions network.

An initial tasks of the Protein Interactions Network will be the integration all protein-protein data from within the group with all protein-protein interaction data available globally into HuPI (<http://hupi.ircm.qc.ca/>). All protein interaction data will be mapped to the integrated site that emphasizes human proteins and incorporates protein interaction data and interrelated bio-information from all species. Chromosome 21 proteins will be highlighted to a single page. One of the deliverables of this group is the HuPI site that will be integrated by the Integrative Bioinformatics group into the Integrative Proteomics Portal. *Number of protein targets:* 195 (all genes from human chromosome 21)

Deliverables:

1. A selection of cell lines and proteins to be targeted by the project's discovery platform **Timeline: 1 year.**
2. A consortium that oversees the deployment of the discovery platform at various sites. **Timeline: 1 year.**
3. The design standard operation procedures (SOPs) for characterizing protein interactions. This includes the choice of affinity tagging, purification and mass spectrometry methods, specifically selected so that the datasets produced at different sites will be compatible and can be integrated into large interaction networks. **Timeline: 1 year.**

A Human Proteome Project

4. Generation of constructs encoding tagged proteins from chromosome 21 followed by purification of protein complexes and LC-MS/MS. **Timeline: 2 years (assessment at 1 year for progress).**
5. Selection of mice for their generation from targeted traps in hybrid ES cells and generation of protein-protein interactions from 100 different tagged protein coding genes (100 mice). **Timeline: 2 years (assessment at 1 year for progress).**

IV. Creation of the bioinformatics interface. To integrate the LC-MS based protein data, the antibody and interaction based profiling of proteins encoded by chromosome 21. Mapping of the frequency (index of abundance) of protein characterization by LC-MS proteomics profiling to their spatial localization (Human ProteinAtlas) and with integration of frequencies (tandem MS) to protein partners as well as integration with, for example, the UniProt knowledge base.

Deliverables:

- 1 A bioinformatics interface that merges the datasets from I, II, III as one into one output. **Timeline: 1 year (interim report at 6 months).**
- 2 A proteome that determines how many proteins have been assigned to human genes of all chromosomes in the various tissues, cell types, organelles and body fluids of the human that has been analyzed up until present time. This will then be compared to the Human Protein Atlas. In this way the complementary annotation via antibodies will define the degree of overlap and current coverage. Hence the merging of the mass spec based resource and antibody resource will be a starting point benchmark. Another aspect is that this assessment phase will guide us in the choice and requirements of the proteomics based profiling project (see V). **Timeline: 1 year (interim report at 6 months).**
- 3 A proteome map of all proteins encoded by chromosome 21 including abundance, spatial distribution and interacting components. **Timeline: 1 year.**

V. Pilot phase for reagents for mass spectrometry based quantitation.

Global MS comparison of HUPO test samples of proteins at equimolar abundance (already completed) and for proteins to test for their relative abundance (over three orders of magnitude). This will ensure standardization of protocols for trypsinization and LC MS/MS as well as to test for peptide complexity quantitatively.

Generation and characterization of full-length proteins for all predicted protein coding genes for chromosome 21. Elucidation of proteotypic peptides by LC-MS/MS in addition to those determined in I (above). Generation of heavy isotope labeled tryptic peptides corresponding to proteotypic peptides. Quantitation of protein abundance for all proteins in chromosome 21 by multiple reaction monitoring (MRM). This will be done following the generation of new samples previously determined from projects I and II to be enriched in these proteins. This will be compared to quantitation using SILAC for cells and animals. During the pilot phase, the quantitative annotation of proteins derived from chromosome 21 will also be completed by including, for example, spiked-in heavy-

A Human Proteome Project

isotope-labeled reference peptides. Here, proteomics-based profiling by 1DE will be performed on a selected sub-set of targeted samples, for example, the human liver and lung as well as selected organelles (plasma membranes, mitochondria) to demonstrate the feasibility of quantitative proteomics. As such, the pilot phase is a demonstration project as well as a means to take stock of where we are in terms of assigning tandem mass spectra to the human proteome.

Deliverables:

1. Standard operating protocols for collaborative MS based proteomics using test samples. **Timeline: 1 year**
2. Generation and characterization of full-length proteins for all predicted protein coding genes for chromosome 21. **Timeline: 2 years.**
3. Elucidation of proteotypic peptides by LC-MS/MS for chromosome 21 (195 proteins). **Timeline: 1 year.**
4. Quantitation of protein abundance for all proteins in chromosome 21 by multiple reaction monitoring (MRM). **Timeline: 1 year**

VI. Phosphoproteomics pilot phase.

Project 1: Phosphoproteomics Standards Initiative

As part of the HUPO standards and technology initiative, a phosphopeptide standards sample will be globally distributed as a test sample phosphopeptides spiked into lysates. Different enrichment techniques, mass spectrometric identification, database search engines, and MRM transitions and absolute quantification will be assessed for complete characterization, leading to robust and reproducible methods for phosphoproteomics. As an extension of I above, tandem mass spectra currently available from the global community will be gathered to map the frequency with which specific phosphopeptides have been characterized on a chromosome by chromosome basis. This information will be formatted and deposited into PRIDE, PeptideAtlas, GPM and other databases. Results will be made freely available via the bioinformatics interface outlined above. As part of this, a robust data-matching component will be developed and implemented to ensure 100% accuracy in site-specific phosphopeptide annotation.

Deliverables:

1. Robust proteomics standards for the analysis of protein phosphorylation. **Timeline: 1 year**
2. Phosphoproteotypic peptides. A visual representation via peptide heat maps with which tandem mass spectra of high quality are assigned to unique phosphoproteotypic peptides to each protein-coding gene. **Timeline: 1 year**

VII. Glycomics pilot phase. Targeted proteomics using protein-specific antibodies followed by MS-analysis will be used to profile all 195 proteins encoded by chromosome 21 in terms of their N- and O-linked glycosylation profile. Proteins will be immunoprecipitated (see II above) from plasma, liver, large intestine, and adipose tissue and treated either with endoglycosidases or mild base to release N- and O-linked oligosaccharides, respectively. Released oligosaccharides will then be separated through LC and analyzed through MS using standard protocols. This pilot phase will be conducted by 3 main sites,

A Human Proteome Project

each analyzing at least 2 of the targeted body-fluids/tissues. This will allow for technology development and implementation of MS analysis in respect to glycomics. The project will be phased with the international glycomics consortium aimed at elucidating glycosylation and function with relevance to human disease.

Deliverables:

1. Complete oligosaccharide profiles of each of the 195 proteins encoded by chromosome 21. **Timeline: 1 year for the 20 proteins of chromosome 21 with antibodies**

VIII. Extension to the production phase

1. Criteria and standard operating protocols for the extension of the pilot phase of the Human Proteome Project to the full scale phase as based on the productivity in each component of the pilot phase.
2. Selection on international committee for cells and organs and sample preparation for mass spec based protein characterization profiling as based on the 6 month interim report from the pilot project.
3. Selection of international committee for criteria and standards for the preparation and characterization of organellar and sub-organellar biological samples as based on 6 month interim report from the pilot phase.

Timeline for completion: 10 years (with yearly assessments of progress and milestones).

A Human Proteome Project

Antibody-Based Profiling

The vision of this part is to generate a map of the human proteome using validated affinity reagents that recognize each of the 21,000 human protein [9]. The project consists of an initial program to demonstrate the feasibility and costs of the generation of renewable affinity reagents for proteome-wide efforts. The long-term objective is to obtain two renewable (paired) antibodies to all human proteins, including also modification-specific antibodies and antibodies recognizing various splice forms or proteolytic fragments. An important mission is to create a first genome-wide draft of the human proteome with normal and disease tissue within a time-frame of 10 years. The antibodies created within this program can also be used to complement the molecular profiling, signal pathway profiling, and the analysis of plasma/serum and other body fluids (Table 1) as well as targeted proteomics of low abundance proteins. The deliverables consist of physical resources and a human proteome map with publicly available data. The physical resources include (i) validated antigens and cDNA clones representing all human genes and available through public database portals and (ii) validated affinity reagents representing a majority of the human proteins and available through public database portals. The human proteome map will contain protein profiling in both normal and disease tissues from numerous different organs, cells and body fluids. In addition, subcellular profiling in all major organelles and other cellular compartments will be presented. The antibodies will also be used to facilitate the molecular profiling of the protein targets, such as molecular weight and splice forms that will continue well beyond the completion of this human proteome project, the characterization of all major isoforms expressed by the 21,000 genes throughout our body.

This project complements and synergizes with the proteomics-based profiling effort. Indeed a bioinformatics based overlay of the proteomics and antibody based profiling efforts will ensure the delivery of the complete characterized human proteome in the shortest time feasible (≤ 10 years).

The initiative is divided into two phases; (i) a demonstration phase to explore different technologies with a focus on the generation of renewable affinity reagents and to develop necessary standards and databases (ii) a full-scale phase in which affinity reagents are generated and validated and a first draft of the global expression profiles of the human proteome is generated followed by a subsequent explorative phase in which the affinity reagents are used by the science community for in depth analysis in various disease models (see clinical relevance below). Here is a short description of the two phases:

Pilot phase

Number of protein targets: 195 (all genes from human chromosome 21)

Deliverables: standard operation procedures for generation of antigens and affinity reagents, feasibility studies using various approaches, database portals for validated antibodies and standards for submission of data, generation of affinity reagents for the majority of the selected protein targets

Full-scale phase

Number of protein targets: 21,000 (the complete human proteome)

A Human Proteome Project

Deliverables: generation of affinity reagents for the majority of the human proteome, generation of database portals with results from the validation, the creation of various databases with results from the use of the affinity reagents to explore normal and disease tissues, cells and body fluids.

The deliverables consist of physical resources and a human proteome map with publicly available data. The physical resources include:

1. Validated antigens and cDNA clones representing all human genes and available through public database portals and a public resource.
2. Validated affinity reagents representing a majority of the human proteins and available through public database portals and a public resource.

The draft of the human proteome including expression profiles in normal and disease tissues as well as human cell lines and body fluids. The information obtained by the antibody program to be included in the first draft of the proteome map is:

1. Protein profiles in normal and disease (organs, cells, body fluids, and subcellular profiling as outlined in Table 1).
2. Molecular profiling of protein interactions (networks, maps, protein complexes etc.) and physical characterization (MW, modifications, splice forms etc.)
3. Renewable antibodies.

Signaling Pathway Profiling

Cells can be considered to be complex networks of interacting molecules and the analysis of intracellular molecular signaling pathways is much less complex as the complete understanding of the products of the about 21,000 human genes. Pathways are conserved throughout the animal kingdom and the link between disease and signaling pathways has been firmly established and validated for many genetic disorders. Reverse-protein microarray approaches will analyze signaling pathways and the currently already available set of about 200 highly validated antibodies will allow a quick entry in the creation of pathway activity atlases, to be expanded over the term of the project to cover substantially the “signalomes” of disease relevant cell lines and tissues. With the ongoing generation of new, highly validated antibodies, comprehensive quantitative pathway activity atlases will be generated to define normal values for pathway activities and protein expression for a wide variety of cells and tissues.

In contrast to sandwich-type protein microarrays which carry immobilized capture molecules (e.g. antibodies) directed to the analytes of interest, cell lysate arrays (reverse arrays) carry the whole proteome of cell or tissue samples immobilized on the chip surface. Cell lysate arrays (reverse arrays) enable the investigation of proteins analyte sets in crude proteomic samples with low amounts of starting material in high throughput mode due to the parallel approach provided by array technologies

Components of pathways in their active and inactive forms are detected by specific

A Human Proteome Project

antibodies against these components. The ratio of active over inactive forms of the proteins provide a measure for the activity of this pathway node in a particular cell or tissue type.

Proteomics Based Profiling

The vision of this part is a comprehensive description of the molecular differences between different fluids, organs and their underlying cell types using a gene-centric approach. This part of the project will further the technology to make fluid, tissue and cell-mapping studies more sensitive, faster, and more comprehensive. In addition, information obtained through this project will aid in the annotation of the 21 000 human genes obtained with antibody-based profiling and network-based profiling. Reagents for such proteotypic peptides will be generated and provided to the community to facilitate the quantitation of the proteome by MS based profiling.

Pilot Phase

High quality data will be gathered and served to the community in a web portal for several HUPO related projects (Table 1) that meet the criteria and standards established for such samples and the proteomics platforms utilized. Proteins that are organ-specific and shared, will be characterized in a label free quantitative mode. These data will be integrated with the ongoing proteomic profiling of organelles derived from rodent organs (Table 1) that already the quality control criteria now rigorously established for such “divide and conquer” strategies. Within 9 months, this compendium of data will have merged all approved MS based proteomics efforts and for the first time have defined a systems wide measurement of protein expression levels.

Such profiling has been demonstrated to be comprehensive and accurate as shown in *S. cerevisia* [10] and *D. melanogaster* [11]. Exactly where we start with the annotation of all 21 000 genes will be realized through the pilot phase.

Full-Scale Phase

Protein characterization is currently best accomplished by using tandem mass spectrometry as it can provide an accurate and sensitive determination of amino acid sequences. To determine the presence of proteins, to uncover splicing variations, and to determine posttranslational modifications will require the use of the available forms of tandem mass spectrometry and the variety of database searching programs and de novo interpretation methods. These methods will provide the sensitivity and scale required to characterize every human protein.

The HUPO test sample effort, after feedback to investigators, achieved a 100% success rate in the characterization of a 20-protein test sample. This defines a standardization methodology for all mass spec based platforms [4]. Coupled with the standardized reporting protocols for mass spec experiments via the HUPO PSI Initiative and data repositories such as PRIDE, TRANCHE, Peptide Atlas and Global Proteome Machine (GPM), data acquisition in a standardized manner is also assured. The project will focus on a subset of tissues and body fluids which are immediately tractable by proteomics

A Human Proteome Project

(Table 1). These are major organs that fulfill two main criteria. They have a high relevance to human disease and represent the main tissue types in the body. The experiments will be done on the organs themselves and a call for all rigorously defined cellular subsets of these organs will be made for a further analysis of individual cell types of these organs .

The tissues and body fluids will be coordinated with a new collection of biological samples from stakeholders including the NCI of the NIH and ongoing HUPO initiatives. High quality clinical samples obtained with Institutional Review Board approval and the informed consent of each donor undergoing surgery as a normal course of therapy are available for immediate characterization. These tissue banks and their already separated major cell types with selected high quality isolated organelles are available for lung, colon, prostate, breast, bladder, and kidney, pancreas, liver, and ovaries. Visceral and sub-cutaneous fat are also available.

Samples for Proteomics Based Profiling

The rodent liver will be selected for a full organellar mapping (Table 1) using the highest standards for organelle purity and characterization. These methods are well established for rodent liver but will be extended to human liver.

The quantitative profiling, targeted profiling by MRM and SILAC, generation of prototypic peptides as reagents and use of existing proteomics repositories and computational biology to gather and integrate the data with the antibody based profiling and networks based profiling efforts will be the resource generated. A focus on antibody based generation of protein complexes specific to each complex [12] will be realized as well as a focus on glycosylation and phosphorylation.

Networks-Based Profiling

The vision of this part is to define protein interaction networks for various human and mouse cell types and tissues with the purposes of elucidating the biological roles of human proteins on a genome-wide scale and facilitating the development of diagnostic and therapeutic agents as based on the pioneering work in budding yeast [5, 6, 8].

Proteins must interact with other molecular components of the cell, including other proteins, DNA sequences, RNA molecules and various metabolites, to exert their functions. Mapping protein interaction networks in cells and tissues is expected to produce fingerprints of the physiological states of these cells and tissues; similarly, mapping changes in protein interaction networks occurring during disease progression should generate signatures for specific pathological states. Because protein interaction maps represent multi-variable, complex descriptions of physiological/pathological states, they also provide the descriptions needed to more realistically and reliably address issues such as the causes, the diagnoses and, eventually, the cures of human diseases.

A Human Proteome Project

This repertoire of human protein interaction maps, termed the Human Proteotheque, will be built via a concerted, inclusive initiative, the Human Proteotheque Initiative (HuPI); this will involve discovery platforms aimed at defining protein-protein, protein-DNA, protein-RNA and protein-metabolite interactions, and will integrate data into comprehensive protein interaction maps through bioinformatics. A collection of cell lines each expressing affinity-tagged proteins is used as a systematic, unbiased discovery platform to identify human protein interactions and characterize human protein interaction networks. Computational procedures are used to select high-confidence protein interactions and to build comprehensive maps of high-density interaction networks of the mouse counterpart of the Human Proteotheque Initiative (HuPI). For the mouse component of this effort, a collection of mice, each with a knocked-in, affinity-tagged protein is used as a systematic, unbiased discovery platform to identify mouse protein interactions and characterize mouse protein interaction networks. Computational procedures are used to select high-confidence protein interactions and to build comprehensive maps of high-density interaction networks.

The integration of these efforts using tagged proteins for network-based profiling will clearly complement the antibody and proteomics based profiling parts. Since phosphorylation of proteins in these networks will also be mapped either in the tagged strategy for network profiling or by co-immunoprecipitations in the proteomics based profiling, mechanistic insights into the proteome will be immediately available to the community.

Integrative Bioinformatics

The vision in this part is of a reference portal for the human proteome. This portal will develop an infrastructure that makes information of all the individual proteomic domains truly *accessible* to the community. Each of the profiling parts outlined above and in Figure 1 will have a distinct bioinformatics program tailored to the required data analysis and necessary computational biology to ensure the rigorous annotation of the proteins characterized by the antibody-based, proteomics-based, and network-based profiling parts. It is the integration of the data generated from these profiling parts and their transformation into one format and the construction of an interface for the basic science and clinical communities that will be achieved in this integrated bioinformatics component. It will allow all levels of users to explore the human proteome as an integrated resource at the same time as retaining the integrity of the specialized data bases from which it draws its information. Thus, the vision is of a user-friendly window that provides live information from multiple sources via a single world wide web-based interface.

The implementation of both DAS and BioMart as standard protocols at each of the contributing databases will assure the use of established protocols to access information from all contributing databases.

A Human Proteome Project

Access will be protein oriented with each protein annotated with a summary page that interfaces to all available data. The summary page will provide a descriptive name, a unique universal identifier capable of accessing all databases, a brief description of function, subcellular localization, organ and tissue distribution and a curated specific bibliography. Detailed information will be readily retrievable via drop-down menus that access related protein databases (e.g. NCBI, UniProt), databases related to genomics, transcriptomics, proteomics, protein structure, protein-protein interaction and antibodies and also informatics related tools (e.g. blastp and blat search tools, structure prediction).

Conclusion

The description of the integrated human proteome will be completed in 10 years or less. The ensured outcome is that of a representative high quality quantitative representation of each major isoform encoded by the human genome. This gene-centric approach minimizes the predicted complexity of the human proteome with the characterization of 21,000 different proteins, each representative of their corresponding gene. This will include post translational modifications (some 200 different), site specific post translational modifications, allelic and splice variants. These will be incorporated from the start and will ensue beyond the completion of the characterization of the 21,000 genes. Pilot projects for the antibody based, mass spectrometry based, and networks based profiling will be completed within 18 months. A new bioinformatics based portal constructed to merge the data generated through the strategies of Figure 1 will have a pilot project to be completed within 6 months. This in turn enables ready completion of the pilot reagents for antibodies and MS based profiling as well as this interaction network. The resource, and reagents generated, including antibodies, proteotypic peptides and their labeled derivatives will be provided as they are developed to the community in a public, accessible format. The integration of the synergistic profiling activities of mass spectrometric, antibody, signaling pathway, and networks-based profiling as well as the posttranslational modifications for the representative protein from each of the 21 000 genes will define the completed Human Proteome project. The output will be designed such that further refinement like the characterization of PTMs or different isoforms related to a single gene locus (by a continued effort past the completion of this project or by individual investigators) will be easily integrated in the resource to provide subsequent, more advanced drafts of the proteome.

Drafting of the HUPO proposal

This proposal was put together at the 4th International Proteomics Conference at the Bellairs Research Institute, McGill University, in Barbados, between January the 4th and 11th, 2008. Following discussion and input from all participants, a first draft was completed and circulated for further refinements. Draft 2, 3 and subsequently, 4 and 5 contain all suggested amendments received and considered up until January the 30th, 2008. The third draft was then circulated to the HUPO Executive Council, the co-chairs of all HUPO Initiatives, the Industrial Advisory Board of HUPO and the attendees of the Bellairs conference. The 4th draft was delivered to the NIH and the EU for their consideration. This represents the 5th draft.

A Human Proteome Project

Antibody-based profiling (complete program)

Executive summary

The vision of the antibody-based engine of the HUPO project is to generate a map of the human proteome using validated affinity reagents representing major isoform(s) of each human protein. The project consists of an initial feasibility program to demonstrate the feasibility and costs of the generation of renewable affinity reagents for proteome-wide efforts. The long-term objective is to obtain two renewable (paired) antibodies to all human proteins, including also modification-specific antibodies and antibodies recognizing various splice forms or proteolytic fragments. A mission is to create a first genome-wide draft of the human proteome with normal and disease tissue within a time-frame of 10 years. Using monospecific polyclonal antibodies, replaced over time by renewable monoclonal antibodies as the methodologies for the latter became more robust and amenable to high throughput production. The antibodies created within this program can also be used to complement the molecular profiling and the analysis of plasma/serum and other body fluids. The deliverables consist of physical resources and a human proteome map with publicly available data. The physical resources include (i) validated antigens and cDNA clones representing all human genes and available through public database portals and (ii) validated affinity reagents representing a majority of the human proteins and available through public database portals. The human proteome map will contain protein profiling in both normal and disease tissues from numerous different organs, cells and body fluids (Table 1) as obtained by immunolocalization. In addition, subcellular profiling in all major organelles and other cellular compartments will be presented. The antibodies will also be used to facilitate the molecular profiling of the protein targets, such as molecular weight, splice forms etc.

Introduction

One of the great challenges in bioscience today is the need for well-validated probes to explore the human proteome. Such reagents can be used to analyze the corresponding proteins both *in vivo* and *in vitro* using assays, such as ELISA, protein arrays, Western blots, immunohistochemistry, immunofluorescence and immune capture [9]. The generation of affinity reagents on a whole proteome scale requires the selection and development of unit operations suitable for high-throughput production and this enforces the need to address strategic and technical issues, such as the choice of antigen, choice of affinity reagent type and the depth of validation of the generated binding reagents.

The human proteome consists of approximately 21,000 non-redundant proteins defined as a representative isoform from every gene locus [1]. The Human Antibody Initiative of HUPO has suggested that the aim for an international effort would be to generate paired affinity reagents to these proteins within a reasonable time-frame [9]. The paired binding reagents should preferably be produced toward two separate and non-overlapping epitopes of the same protein target to facilitate the validation of specificity by allowing comparisons between the paired antibodies with regard to staining pattern across various

A Human Proteome Project

analysis platforms. Paired affinity reagents also allow for coupled assays with high signal-to-noise ratios such as capture/detection (sandwich) assays and other assays with requirement for dual binding to increase specificity and sensitivity.

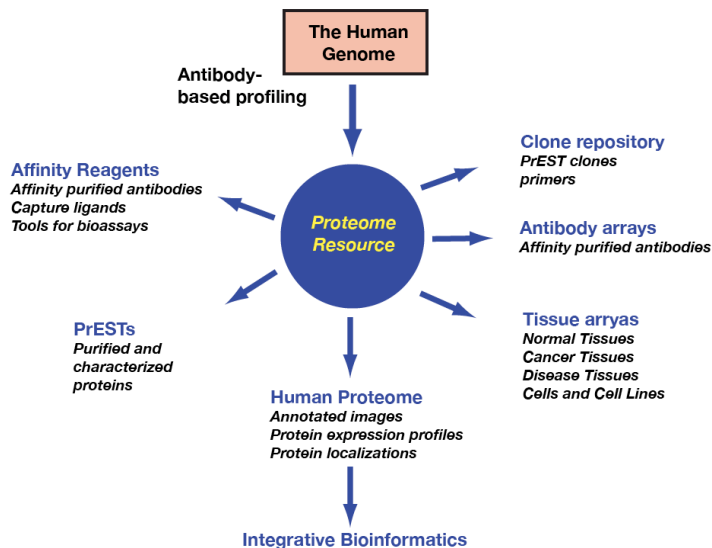


Figure 2: Antibody-based profiling as a resource in the Human Proteome Project. Mono-specific high quality antibodies are generated from the predicted proteins encoded by the human genome and used to profile tissues and body fluids as well as sub-cellular structures to yield annotated images, protein expression profiles and localization. Generated antibodies can also be used to evaluate tissue arrays or alternatively, be spotted onto arrays for direct profiling of protein samples. In addition, generated antibodies can be applied in targeted proteomics to immuno-capture low abundance proteins for proteomics profiling.

The validation of the affinity reagents is important and it has been estimated that at least half of the commercially available antibodies today have question marks around functionality, and in many cases give unreliable results across several analysis platforms. This results in an enormous amount of wasted resources, due to the purchase and use of these non-functional affinity reagents. It is not inconceivable that many hundreds of millions of dollars of research funding in academia and pharmaceutical companies are spent unnecessarily each year. More importantly, huge amounts of time and effort are wasted due to the time and effort on generated data of poor quality. There is therefore a need to initiate an international effort, preferably as a partnership among industry, academia and funding agencies, to promote the generation of affinity reagents on a proteome-wide level and to ensure that these, including renewable affinity reagents are made available to the communities as a public resource, and that the validation is publicly available through a standardized format [13, 14].

Process

The technology for generating antigens and antibodies are well established, although few systematic efforts in large-scale have so far been described [9]. For the generation of affinity-purified polyclonal antibodies, a scale-up strategy through the Human Protein Atlas program (www.proteinatlas.org) during a five year period (starting in 2003) has led to a present weekly output of approximately 50 new validated antibodies to new targets suggesting that whole-proteome efforts are feasible [15]. The results obtained from the first five years of this program suggest that this strategy can be scaled-up even further. For the generation of renewable affinity reagents, such as monoclonal antibodies and/or recombinant protein binders, no attempts have yet been initiated aimed to generate affinity reagents on a whole-proteome scale. The Human Proteome Project therefore

A Human Proteome Project

consists of an initial feasibility program to demonstrate the feasibility and costs of the generation of renewable affinity reagents for proteome-wide efforts. In parallel, the cost and success rates of systematic generation of monoclonal antibodies in large scale will be evaluated and has already commenced. An objective for the first draft of the human proteome map (5-10 years of full-scale project) will be to have mono-specific (affinity-purified polyclonal) antibodies to a majority of the non-redundant proteome and with renewable affinity reagents and the corresponding clones. The long-term objective is to obtain two renewable (paired) antibodies to all human proteins, including also modification-specific antibodies and antibodies recognizing various splice forms or proteolytic fragments. There are ample opportunities for a partnership between academia and industry, although all partners must make all reagents, such as protein antigens and affinity reagents, available to the research community at cost with no or limited intellectual property rights (IPR) restrictions. The Human proteome initiative could be divided into five major parts: All cDNA clones or hybridomas will be deposited and accessible to all.

1. To develop and evaluate new technologies and principles for all aspects of the generation of affinity reagents on a systematic (proteome-wide) level.
2. To generate validated antigens to all the major isoforms of human proteins
3. To generate affinity reagents to the major isoforms of all human proteins
4. To validate these affinity reagents by generic and application-specific methods and to make all validation data publicly available
5. To perform a proteome-wide analysis of human tissues, cells and body fluids from both normal and disease origin using the affinity reagents.

Overall deliverables (output)

The proposed deliverables of the project depend on the amount of funding and the number of participating groups. The focus will be to use the affinity reagents created in this program to perform tissue, cell and subcellular profiling. The antibodies can also be used to complement the molecular profiling and the analysis of plasma/serum and other body fluids. The deliverables consist of physical resources and a human proteome map with publicly available data. The physical resources include:

1. Validated antigens and cDNA clones representing all human genes and available through public database portals
2. Validated affinity reagents representing a majority of the human proteins and available through public database portals

The draft of the human proteome including expression profiles in normal and disease tissues as well as human cell lines and body fluids. The information obtained by the antibody program to be included in the first draft of the proteome map is:

1. Protein profiles in normal and disease (organs, cells, body fluids, and subcellular profiling as outlined in Table 1).
2. Molecular profiling of protein interactions (networks, maps, protein complexes etc.) and physical characterization (MW, modifications, splice forms etc.)

A Human Proteome Project

3. Renewable antibodies.

Time frame

The initiative could be divided into two phases; (i) a demonstration phase to explore different technologies with a focus on the generation of renewable affinity reagents and to develop necessary standards and databases (ii) a full-scale phase (“the HUPO project”) in which renewable affinity reagents are generated and validated and a first draft of the global expression profiles of the humane proteome is generated followed by a subsequent explorative phase in which the affinity reagents are used by the science community for in depth analysis in various disease models (see clinical relevance below). Here is a short description of the two phases:

Subprograms

The effort can be further divided into five different subprograms that range from the generation of antigens to the exploration of the produced affinity reagents in various human tissues. Each subprogram should be coordinated separately, but care should be taken to integrate the various infrastructures with each other to avoid duplicated efforts. Here is a short description of the various subprograms:

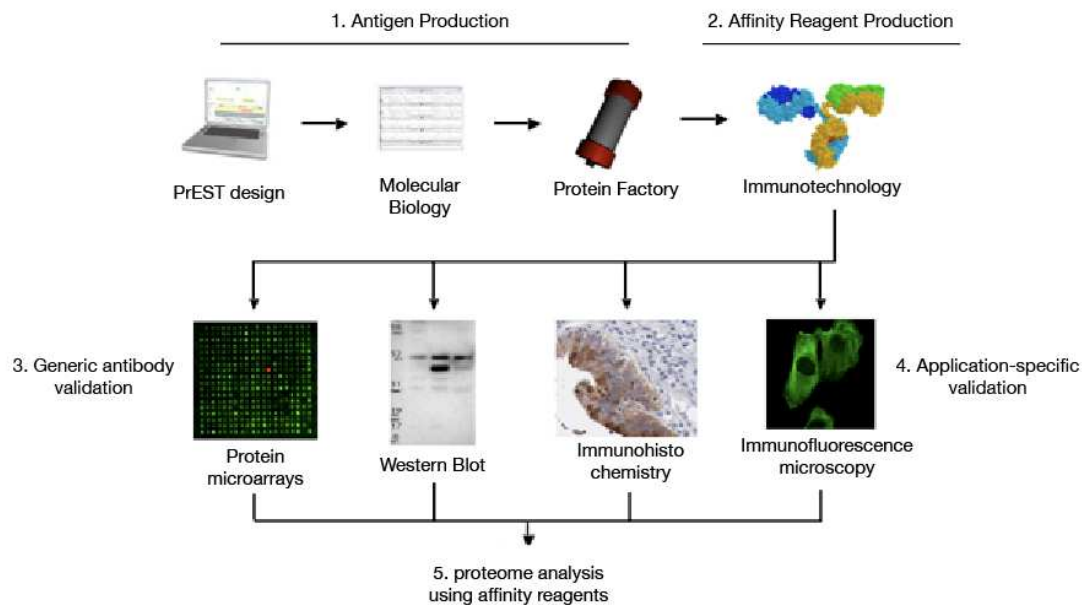


Figure 3. The pipeline for antigen production, affinity reagent production, antibody validation and application specific validation to assure a complete, accurate, and permanent characterization of representative proteins to each human protein coding gene.

1. Antigen (protein) production.

This effort aims to generate validated antigens, protein fragments and full-length proteins corresponding to all human proteins and to make these available to the scientific community. The primary objective is to produce antigens for the generation of affinity reagents (subprogram 2), but is also important to generate protein to be used as probes in the generic validation of the affinity reagents (subprogram 3). In addition, the proteins or protein fragments generated within this effort can be used to generate “proteotypic

A Human Proteome Project

peptides” .The validation of the proteins should be done according to standards, including molecular weight determination with mass spectrometry. All antigens generated within the frame-work of the Human Proteome project should be available at cost for all. Preferably, many different types of antigens can be generated to every protein target to enable the selection and validation of affinity reagents to be pursued using a wide variety of antigens, including protein fragments or peptides from different parts of the protein target. This is particular important in the generation of paired antibodies that need to be directed to distinct different regions of the protein. It is also possible to generate modification-specific affinity reagents using synthetic peptides chemically modified to mimic the modified protein target. A database portal should be developed in which all available antigens are submitted with appropriate validation data and links how to obtain the antigen. The main antigens to be generated within the frame-work of the HUPO-project should be:

1. Full-length protein (using different hosts)
2. Domain-based protein fragments (mainly structural genomics efforts)
3. Protein Epitope Signature Tags (fragments based on low homology)
4. Synthetic peptides (also modification-specific peptides)

2. Affinity reagent generation

This effort aims to use the antigens to select and produce specific affinity reagents to the corresponding protein targets. The primary objective of the pilot phase is to explore various strategies for systematic generation of renewable affinity reagents, both antibody- and recombinant-based. Strategies suitable for scale-up with regards to success rates, cost and quality, can subsequently be used for the second production phase. All antibodies generated in the production phase should be available at cost for all. These efforts will be complemented with stream-line approaches to generate affinity-purified polyclonal antibodies, which are not truly renewable but provide a reference reagent bank for a whole-proteome effort. Ideally, the final aim is to obtain at least one multi-epitope antibody (mono-specific antibody) and one single-epitope affinity reagent (i.e. monoclonal antibody or recombinant antibody fragment) for every protein target. A database portal should be developed in which all available affinity reagents are submitted with appropriate validation data (subprograms 3 and 4) and links how to obtain the protein binder. An important part of the pilot phase is to explore the use of different types of affinity reagents, such as:

1. Affinity-purified polyclonal antibodies (mono-specific antibodies)
2. Monoclonal antibodies (mouse and rabbit)
3. Recombinant antibody fragments
4. Recombinant protein scaffolds
5. Nucleic-acid-based scaffolds

3. Generic antibody validation

This effort aims to use the antigens (subprogram 1) for characterization of the generated affinity reagent (subprogram 2) using various “generic” binding assays. In addition, a comprehensive selectivity screen for example using protein arrays on which all available antigens are spotted. For such generic assays, normally the antigen is used as ligand for the analysis. The objective of the pilot phase is to define and develop standard assays to

A Human Proteome Project

be used through the project. The generation of protein arrays with a large portion of the human proteome is encouraged and these arrays should be made available (at cost) for standardized validation of the affinity reagents. For the monoclonal reagents, the determination of binding parameters is important. It might also be possible to develop systematic and standardized strategies for epitope mapping using synthetic peptides and/or the use of combinatorial libraries and surface display of randomized peptides representing the human proteome. Some examples of subprojects could be:

1. Protein arrays (as proteome-wide as possible) or other binding assays
2. Binding kinetics assays (SPR etc)
3. Epitope mapping (using synthetic peptides, surface display etc)
4. Inter-species reactivity (mouse, rat, dog, primates etc)

4. Application-specific validation

This effort aims to validate the affinity reagents in the most common application used to explore the proteome using protein binders. The protein targets have large numbers of different folding states in various applications. Proteins are often partly or fully denatured, i.e in Western blots, immunohistochemistry or confocal microscopy. Most validation assays are also semi-quantitative and context-dependent and the results will be influenced by choice of tissue and sample preparation methods. The incredible dynamic range of proteins in human body fluids and tissues with dynamic ranges covering at least a dynamic range of 10^{10} [16] makes interpretations even more difficult. This problem enforces the need to define standards for affinity reagents validation and points to the need for standard operating procedures to carry out such validation. An important objective is to use two (paired) antibodies recognizing spatially separated epitopes as a tool to validate the staining pattern in various assays. Again, the use of multi-epitope binding antibodies (mono-specific) and single-epitope affinity reagents in combination is attractive. It is important that all supporting primary data from the validation is made publicly available through database portals. The primary focus on application-specific validations within the frame-work of the Human Proteome project would be:

1. Western blots (several tissues and body fluids)
2. IHC (several tissues and cell lines)
3. Immunofluorescence using confocal microscopy (several cell lines)
4. Flow sorting (several cell lines, only for surface-displayed proteins)
5. Immunocapture (affinity capture followed by MS)
6. Frozen tissue (several tissues)
7. Protein arrays and other binding assays

5. Proteome analysis (using affinity reagents)

An important part of the effort is to use the generated affinity reagents to explore the human proteome in various tissues and cells. This could be done by many different research groups to create a knowledge-base for more in-depth studies by the whole research community. It is important that proteome analysis projects within the frame-work of the HUPO-project is committed to submit the primary data to the proteome databases. All data from should be available free of charge without IPR restrictions using one or several protein atlas portals. The Human Protein Atlas (www.proteinatlas.org) funded by Knut and Alice Wallenberg Foundation, Sweden, is an example of such a

A Human Proteome Project

proteome portal. The proteome analysis will include a whole proteome level analysis of the following:

1. Global tissue profiling in normal and disease (IHC)
2. Subcellular localization (confocal microscopy and electron microscopy)
3. Serum/plasma analysis (in relevant patient cohorts)
4. Developmental profiling (whole body profiling)
5. Brain profiling (whole brain profiling, mouse/rodent)
6. Chromosomal binding (using ChIP analysis)
7. Molecular profiling (MW and analysis of protein isoforms)

The demonstration project

The pilot phase is important to explore various approaches to generate antigens and affinity reagents. We propose that all the genes encoded by chromosome 21 are selected (195 protein targets) and a demonstration project is initiated to try to generate as many different types of antigens from these proteins and subsequently as many alternative affinity reagents based on these antigens. The analysis of an entire human chromosome ensures a relative un-biased selection of proteins with a mixture of enzymes, transcription factors, membrane proteins, signaling molecules etc. The affinity reagents are validated and used to analyze the proteome using various applications. The mono-specific reagents generated with the Swedish Human Proteome Resource (see HUPO HAI initiative, www.hupo.org/hai) can be used as a reference source and other renewable affinity reagents can thus be bench-marked and validated against those reagents. The results should be made publicly available and the data integrated using one or several database portals. The effort is evaluated annually through one or several workshops and issues relevant for scale-up for production phase (cost, throughput, success rates, quality etc) are analyzed. Due to the short time frame, the availability early in the process of different antigens from the selected proteins is crucial. All groups willing to participate to generate affinity reagents in the pilot phase should be given sufficient amounts of purified antigen to enable selection and screening of affinity reagents.

An important part of the demonstration project is to coordinate all groups interesting in participating in the subsequent production phase. A focus will be to create “virtual resources” of antigens and affinity reagents using a community-based database portals in which data on validation is submitted in a standard format. A pilot version of such a database portal is being developed within the EU ProteomeBinder (www.antibodypedia.org) network project. The feasibility of different strategies to generate validated affinity reagents can thus be evaluated in the pilot phase to facilitate funding choices for the production phase. The deliverables from the demonstration program will be:

- A public database portal for validated antigens (including standards).
- A public database portal for validated affinity reagents (including standards).
- Antigens to a majority of the 195 chromosome 21 encoded protein targets using different strategies (preferably both full-length recombinant proteins and selected recombinant protein fragments).

A Human Proteome Project

- Various types of affinity reagents to a majority of the 195 chromosome 21 encoded protein targets (preferably both multi-epitope and single-epitope affinity reagents).
- Standard operating procedures for generation of renewable affinity reagents.
- Results from technical studies aimed to show the feasibility of high-throughput generation of such renewable affinity reagents.

Technology development

A substantial part of the funding (10-20%) of the program should be devoted to technology development to improve the various steps in the process and to enable new technologies to be implemented. During the initial period, it is important to fund efforts to create recombinant affinity reagents in large scale, as well as stream-lined efforts to create monoclonal antibodies using more classical means in a cost-efficient manner. New technologies for mapping and characterization of the epitopes of the affinity reagents are also needed, as well as various improvements for validating both the antigen and the affinity reagent.

A second important effort is the development of high throughput technologies for proteomics analysis. The exponential improvement of genomic technologies, i.e. gene expression profiling by DNA microarrays and high throughput sequencing, help sustain the continuous interest in genomics because for the same amount of funding, exponentially more challenging goals can be set, and reached, every year. Whereas protein profiling in the form of microarrays is now well established, the number of proteins that can be analyzed in parallel is limited because of (i) lack of affinity reagents and (ii) because the current technologies are not scalable and therefore not applicable to whole-proteome profiling. Following the availability of novel affinity reagents as described above, a targeted effort on developing scalable protein analysis technologies under the HUPO umbrella can greatly accelerate the emergence of high throughput proteomics technologies, such as reverse protein arrays.

Synergies with the other HUPO subprograms

There are numerous synergies between the antibody-based profiling and the other two HUPO subprograms. Most notably, the mapping of the protein profiles in cells and tissues benefits from the added validation based on the corresponding profiling using the proteomics-based effort. In addition, the generated protein fragments can be used to map the proteolytic peptides of the corresponding protein targets as part of the proteomics-based profiling engine. The mapping of subcellular distribution and tissue specificity can be used to further validate the networks created within the interaction-based effort. The antibodies can also be used to validate potential biomarkers discovered through the other two efforts.

A Human Proteome Project

Clinical relevance

The antibody-based profiling of normal and disease tissue on a whole proteome level will have incalculable benefits for biomedical research. All researchers with an interest in protein targets for therapy, imaging and diagnostics will be able to use the resource to gain more insights. Potential drug targets and biomarkers can be validated using the antibodies and insights relevant for personalized medicine can be gathered such as larger clinical studies with carefully selected patient cohorts. The studies can be expanded to virtually all diseases and further used to analyze relevant human disease models in model animals. The antibodies can also be used for in depth functional studies of the corresponding protein target to explore disease models and networks.

Proteomic-based profiling

Executive summary

The proteomics-based profiling part of the human proteome project will deliver a complete characterization of the human proteome by quantitative mass spectrometry (MS) in representative cells from each major human organ and major body fluids. The strategy will define all proteins including integral membrane proteins of the representative cells, their localization, their association with other proteins and their posttranslational modifications, especially glycosylation and phosphorylation. This part will also use antibodies delivered through the antibody-based profiling part to immunoprecipitate protein complexes to define interacting components by MS. As such, this part overlaps with, and extends, the network-based profiling part as well as providing it with a firm molecular link to the cells, organs and body fluids characterized in this and the antibody-based profiling part. The initial accumulation of all available data in cells and organs via Peptide Atlas (defined as the pilot phase) will determine the degree of coverage of the 21,000 genes of the human genome already mapped by the community. This pilot phase builds on envisaged bioinformatics interfaces that effectively will clear up all redundancies and inconsistencies in the different databases using a controlled vocabulary for each protein. Such controlled vocabulary will further ensure ready and understandable access by clinicians and basic scientists. As for the antibody-based profiling part, special attention is paid to chromosome 21. With less than 200 proteins predicted as being encoded in this chromosome, its completion and implementation of technology to characterize its most difficult encoded proteins will be a milestone reached early in the project. Indeed, the provision of labeled proteotypic peptides for quantitation by Multiple Reaction Monitoring will be made available to the community for all proteins of chromosome 21 to enable accurate abundance measurements of these proteins in this pilot phase. Technology will also be developed to assure abundance quantitation for all proteins of the human genome.

Introduction

Quantitative proteomics is an important discovery component in a wide range of clinical

A Human Proteome Project

and biological research projects. From the pioneering efforts of using two-dimensional separation by polyacrylamide gel electrophoresis (2DE) to the 1DE approach used in combination with liquid chromatography (LC) MS/MS for peptide characterization, proteomics-based profiling is today a reliable and reproducible technology (for an overview of currently used technologies, see Figure 4). In addition, samples can be analyzed through gel-free strategies where proteins are digested to yield peptides that are then separated by two dimensions of liquid chromatography followed by MS/MS analysis. Such strategies are robust and will be considered as well.

As means of quantitation, it is possible to introduce distinct isotopic labels by metabolic or chemical tagging in one reference sample. The differentially labeled samples are then combined and concurrently analyzed. In one implementation of this method, the combined, labeled sample is digested and the resulting peptide mixture is analyzed by multidimensional chromatography and tandem mass spectrometry (MS/MS). Here, peptide abundance and identification are concurrently determined in a single, automated operation.

For protein differences between samples as well as for the abundance of any single protein in different samples, labeling methods such as SILAC (Stable Isotope Labeling by Amino acids in Cell culture) have been prominent. SILAC has proved to be straightforward for special cell lines and to quantify posttranslational modifications as well as to quantify tissue samples.

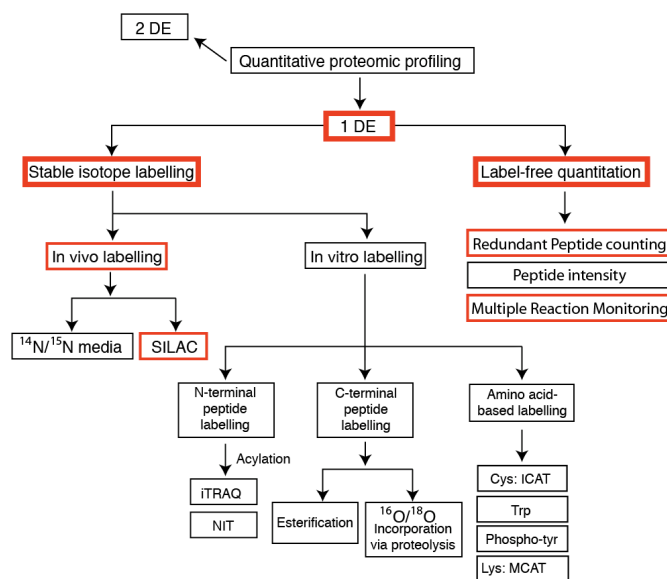


Figure 4: Quantitative Proteomics Profiling: 2DE, the intensity of staining of the resolved proteins is quantified by gel-imaging technologies. 1DE, peptides are observed by MS and quantified either by stable isotope labeling or through label-free quantitation. In vivo labeling strategies rely on the incorporation of stable-isotopes into cells cultured in media enriched in $^{14}\text{N}/^{15}\text{N}$ or with stable isotope (^{15}N) labeled amino acids followed by mixing of the proteins for quantitation via intensity of the labeled and unlabelled peptide ions while employing fragmentation of the unlabelled peptide for identification of the cognate protein (SILAC). In vitro labeling targets the N-termini (acetylation, iTRAQ and NIT) or C-termini (esterification and $^{16}\text{O}/^{18}\text{O}$) of the tryptic peptides or specific amino acids (cys by ICAT, trp, phospho-tyr or lys by MCAT). For absolute quantification, stable-isotope

labelled synthetic tryptic peptides are added to the sample and intensities are determined. In label-free quantitation, redundant peptide counting determines the sum of the number of tandem MS assigned to the cognate protein at 95% confidence. MS peptide intensity quantifies based on the intensity of assigned peptide ions. MRM employs the intensity of a signature fragment ion of the corresponding peptide ion. The figure is adopted from [17].

Several studies have adopted a "divide and conquer" strategy to comprehensively analyze specific subsets of the proteome through purification strategies. Such studies include the

A Human Proteome Project

analysis of functional multi protein machines such as the ribosome, spliceosome, and nuclear pore complex, or organelles (Table 1), and subcompartments such as defined chromatin complexes and co immune precipitates [18, 19]. Another approach has been the analysis of proteins with common distinguishing structural features, such as phosphate ester groups, cysteine residues or the ability to specifically bind to certain compounds [20, 21]. These strategies have in common that they focus on the in-depth analysis of sub-proteomes of rich biological context.

To achieve a complete and accurate description of the tissue-based cellular or subcellular proteomes in a quantitative manner, a number of challenging technical issues have recently been resolved or will be resolved in during the pilot phase. These include:

- Strategies that distinguish true and false positive protein identifications in a preparation while minimizing the number of false negatives. For example, with increasing stringency of isolation aimed at minimizing false positives, the fraction of true positive proteins that get eliminated from the sample also increases (false negatives). This holds true in tissue isolation as well as cellular and sub-cellular isolation. For example, peripheral membrane proteins are easily lost from an organelle during its isolation. In this case, these represent false negatives. With more stringent purification protocols (e.g. saltwash, repeated centrifugation, detergent extraction, use of immuno-based isolation of sub structures), proteins and organellar substructures may simply be lost. By adopting a less stringent protocol combined with a quantitative and comparative protein profiling of different sub-cellular fractions, the degree of contamination can be easily gauged, aiding in the development of optimal purification protocols for each organellar structure [22].
- The conclusive and consistent assignment of peptide and protein sequences to the tandem mass spectra collected in LC-MS/MS experiments remains challenging. The implementation of recently developed resources that use statistical principles to assign tandem mass spectra to peptide and protein sequences at consistent and known true and false positive error rates are already available. Here, probability-based empirical data accumulated from millions of spectra provides a numerical value for each predicted tryptic peptide in terms of its probability of flight. In other words, by searching for signature proteotypic peptides (as deduced by MS spectra) that have a high probability of flight, the presence of a given protein in a sample can be readily assessed. It is then possible to say with high confidence if a given protein is present in a sample or not by searching for its high probability spectra(s). As more and more MS spectra are accumulated in the community, the signature proteotypic peptides for each and every protein of the human proteome will be ascribed a numerical probability-based number with an ever increasing precision (for in depth characterization of targeted analysis, please see the Multiple Reaction Monitoring section below).

A Human Proteome Project

Process

To achieve detailed understanding of biological systems, the proteomics-based profiling part has to be comprehensive as well as responsive in regards to a multitude of external (environmental, pharmacological) and internal (genetic) variations. As shown above, there are multiple approaches that can be deployed in quantitative proteomic profiling with label-free methods gaining prominence. This will be considered during the pilot phase of the project that takes stock of the current coverage of the human proteome and quantitatively annotates all proteins of chromosome 21. Along with the development and implementation of a robust data matching component, the best strategy will then be used to carry out the remaining parts of the project.

Ongoing HUPO and other initiatives that feed into the project

HUPO has a portfolio of high quality projects that have already generated large data sets (Table 1). Peptide atlas for the plasma proteome has already accumulated over 2 million peptides as well as the data accumulation methodology that spans all available tandem mass spectrometry platforms. The brain proteome project representing over 850,000 tandem mass spectra has also defined a straightforward methodology to merge all the data from the 12 participating groups of the brain proteome pilot project. Furthermore, the PICR (protein identifier cross referencing) effort of the EBI has defined a method to update the assignment of tandem mass spectra. Hence, past HUPO initiatives and the experience of the HUPO test samples in which all current platforms work flawlessly can now be merged and updated readily.

The HUPO test sample effort [4] defines a standardization methodology for all mass spec based platforms. Coupled with the standardized reporting protocols for mass spec experiments via the PSI HUPO Initiative and data repositories such as PRIDE, TRANCHE, Peptide Atlas and GPM then data acquisition in a standardized manner is also assured.

The MCP guidelines now implemented by a majority of the community will be used and built on to continue this important start.

Multiple Reaction Monitoring

Whereas MS/MS proteomics provides an incomplete survey of all proteins present in a sample, somewhat analogous to EST sequencing in the transcriptomics field, targeted proteomics can provide quantitative measurements for all the desired proteins relevant to a study, analogous to micro arrays in transcriptomics. A technique called multiple reaction monitoring (MRM) is performed by first compiling a list of targets (proteins and peptides) to be measured, along with their expected fragmentation patterns. Then the mass spectrometer is provided with a list of pairs of precursor ion m/z values and product (fragment) ion m/z values (the combination of which is called a transition or reaction) uniquely identifying these targets to continuously monitor instead of obtaining full MS/MS spectra on the most abundant precursors. For each transition, the instrument measures intensity over time, and multiple transitions per peptide and multiple peptides per protein yields replicate measurements for each protein of interest. Targeted

A Human Proteome Project

proteomics also has its own informatics challenges, although primarily in optimal target selection.

The benefit of MRM is that high abundance proteins in a sample will not distract the instrument from the targeted proteins. When combined with spiked-in heavy-isotope-labeled reference peptides, one is guaranteed an abundance measurement or at worst, an upper limit for every targeted protein. This allows the creation of a complete protein abundance map for every processed sample.

Posttranslational modifications

Posttranslational modifications (PTMs) are important and indispensable for an understanding of the proteome. PTMs often are prerequisites or modulators of protein functions, as exemplified by phosphorylation and glycosylation. Major accomplishments by several HUPO members in the mapping of phosphorylation sites and characterization of kinases will now allow studies of this PTM on a genome-wide scale.

The most abundant post-translational modification is glycosylation, estimated to occur on 50-80% of all eukaryotic proteins. Glycans produce analytical challenges because they are usually branched, are found N-linked to Asn and O-linked to Thr/Ser, and are heterogenous at a given glycosylation site owing to the non-template driven biosynthesis. Recent advances have revolutionized the analysis of glycans using mass spectrometry to a level now practical for proteomics initiatives.

The sensitivity of mass spectrometry for glycans and glycopeptides allows for the analysis from material obtained from high throughput protocols using specific antibodies to immunoprecipitate the protein from tissues and fluids. Here, a profiling of the glycans of each glycoprotein can be achieved providing important information about the protein in question as well as how a given protein might differ between cell types and indeed, to samples obtained from patients exhibiting a particular disease or disease state. There are challenges: The complex heterogeneity of mucin-type O-glycans demands good separation methods for the high number of isomers, as do the glycosaminoglycans that require a partial oligosaccharide cleavage and special care to prevent loss of sulfate groups. Secondly, the sites of glycosylation and site-specific patterns are best determined by tandem MS analysis of the glycopeptides obtained by protease digestion. For the N-glycans, the consensus sequence for attachment is known and the glycan profiles at one glycosylation site can be more easily identified. Such direct approaches are of course ideal and can sometimes be used as the first and only analytical step. However, additional developments, both instrumental and computationally, are expected to further such direct approaches considerably.

The HUPO Glycomics initiative has undertaken a process to define the most reliable protocols for glycoprotein characterization at high throughput. These are already available for N-glycans [23], whereas for others (O-glycans), an ongoing evaluation of methods in different laboratories suggests methodological consensus. A major project within the HUPO Glycomics initiative has already focused on a 5-year program to discover, develop and clinically validate cancer biomarkers by targeting the glycan part

A Human Proteome Project

of the protein (or lipid). Current development of sensitive LC-MS systems and novel types of post-source ion fragmentation methods holds promise to facilitate glycan analyses. A major challenge, as outlined in a NIH white paper [24], is the urgent need for open access glycan databases and better tools for rapid analyses of mass spectrometry information. A concerted effort to perform high-throughput analysis of the glycoproteome would go a long way to addressing the challenge of posttranslational modifications of the proteome.

The Pilot phase

Collection of high quality data attached to a current common database for human liver, brain, kidney, heart as well as lung. These data will use the proof of principle of Peptide Atlas for the plasma proteome as well as the data accumulation and reprocessing in a standardized procedure for the brain proteome project. This will utilize protein identifier cross referencing (PICR) to assure annotation to the most recent database. The proteins which are organ-specific and common in organs will be characterized in a label free quantitative mode. This requires the accommodation of more than 100 million tandem mass spectra that will be submitted immediately as the call is given and will be completed within 9 months.

Ongoing proteomic profiling of organelles derived from rodent organs (Table 1). These already meet quality control criteria now rigorously established for such “divide and conquer” strategies. This will be completed in nine months and again, approximately 100 million tandem mass spectra will be collected.

Ongoing proteomic profiling of cells (stem cells, macrophages) to define the current status of their proteomes will also be incorporated.

The above data will determine how many other proteins have been assigned to human genes of all chromosomes in the various tissues, cell types, organelles and body fluids (Table 1) characterized in the overall project. This will then be compared to the Human Protein Atlas. In this way the complementary annotation via antibodies will define the degree of overlap and current coverage. Hence the merging of the mass spec based resource and antibody resource will be a starting point benchmark. Another aspect is that this assessment phase will guide us in the choice and requirements of the HUPO proteomics based profiling project.

During the pilot phase, the quantitative annotation of proteins derived from chromosome 21 will also be completed by including, for example, spiked-in heavy-isotope-labeled reference peptides. Here, proteomics-based profiling by 1DE will be performed on a selected sub-set of targeted samples, for example, the human liver and lung as well as selected organelles to demonstrate the feasibility of quantitative proteomics. As such, the pilot phase is a demonstration project as well as a means to take stock of where we are in terms of assigning tandem mass spectra to the human proteome.

A Human Proteome Project

Full-Scale Phase

The project will focus on a subset of organ-specific cell types, organelles and body fluids (Table 1) that are immediately tractable by proteomics and have a high relevance to human disease as well as representing the components of the body.

The organs are: Lung, liver, kidney, brain, pancreas, skeletal muscle, heart muscle and adipose tissue (white and brown) (Table 1). The experiments will be done on the organs themselves and a call for all rigorously defined cellular subsets of these organs will be made for a further analysis of individual cell types of these organs. The cells are: Lymphocytes, dendritic cells, mouse and human embryonic and adult stem cells. The body fluids are: Plasma, CSF and urine. The tissues and body fluids will be coordinated with a new collection of biological samples from stakeholders including the NCI and ongoing HUPO initiatives. The organelles are as describe in Table 1 and include: The endothelial plasma membrane, the hepatic plasma membrane, the hepatic endocytic clathrin coated vesicles, the hepatic non clathrin coated endocytic carriers, early and late endosomes, lysosomes, autophagosomes, peroxisomes, mitochondria, nuclei, chromatin sub-compartments, nucleoli, the nuclear envelope, the nuclear pore, the rough endoplasmic reticulum, the smooth endoplasmic reticulum, COPII vesicles, the intermediate compartment, the *cis*-Golgi network, the stacked Golgi cisternae, COPI vesicles and their sub-compartments, Golgi derived clathrin coated vesicles, and secretory carrier vesicles. The first 3 years will be devoted to the full organellar mapping of the rodent (comparative profiling) and human liver using the highest standards for organelle purity and characterization. Organelles derived from other tissue-specific tissues will then follow. A HUPO organelle profiling committee will define the criteria to select the groups that can best match the requirements in this effort. An important consideration is that the isolation strategy and proteomics profiling will have to be reproduced by at least two labs.

A HUPO Proteomics profiling committee will ensure standardization of the MS-based characterization to enable the merging of all the mass spectra data with assignment of the peptides and proteins to the cognate organ-specific cell types and their organelles as visualized by heat maps. This would be matched to all proteins visualized to the cognate subcellular structures by the antibody-based profiling effort. The merging of these two resources would give the most complete annotation of cellular substructures ever known focusing initially on the liver. Protein interactomes deduced from the network-based profiling part provides functional profiling to all organelles and subcellular compartments characterized in this effort.

Technology Development

The sensitivity and reproducibility aspect of MS-based sample characterization is continuously being improved. Today, it is already feasible to fully characterize a given sample quantitatively using one instrument in less than a year providing that heavy-isotope-labeled peptides are available. The amount of sample required to meet the expected dynamic range peptides (for example 5 orders of magnitude) is no longer prohibiting. Routine LC-MS/MS analysis already operates at the femtomole level

A Human Proteome Project

requiring only microgram quantities of tissue or body fluid material for characterization. It is expected that technology development will decrease this by one or two orders of magnitude and that this will occur shortly. Already, custom-built nano-LC and nano-spray interfaces can be mounted on a FT-ICR instrument or an Orbitrap and pushes the sensitivity at least a 100-fold while retaining robustness in terms of reproducibility. Therefore, a significant portion of the funding for the proteomics-based profiling part should be put aside to allow full implementation of improved technology across participating proteomics centers. Another factor concerns quality. Standards and quality assurance is an inherent property of the envisaged human proteome. As such, methods and assays that can routinely monitor the quality must be developed and incorporated.

Post Translational Modifications

Many mass spectrometry based targeted approaches have been introduced to study protein post-translational processing and modifications, for instance by enriching for and selective analyses of protein N-termini and or protein phosphorylation, the latter enabling the temporal charting of thousands of phosphorylation events in a human proteome [25]. As is expected that at a single moment a proteome may contain more than 100,000 phosphorylation events [26, 27], methods need to become more sensitive to be comprehensive. Other modifications, in particular protein glycosylation [28] and ubiquitination [29], require even more powerful approaches to understand their function at the proteome level. Improvements in separation, enrichment and mass spectrometry based technologies are therefore still indispensable to reveal the human proteome in its full glory.

In the Human Proteome Project technologies and strategies will be standardized and optimized to enable more comprehensive proteome coverage, whereby the aim surpasses protein identification and is focusing on functional annotation.

Protein Phosphorylation

Reversible protein phosphorylation is likely the single most important mediator of signal transduction in living organisms and more than 1/3 of all human proteins are estimated to be phosphorylated during their lifetime. Many oncogenes are phosphoproteins or protein kinases that are involved in the regulation of cell growth, differentiation and death. Therefore, the analysis of protein phosphorylation is of essence for clinical applications. There is a tremendous need for novel analytical methods for the systematic and routine characterization and screening of phosphorylated proteins. Essential are 1) the specific enrichment of phosphopeptides and -proteins, 2) tools for relative quantification of phosphorylation in dynamic complexes, and 3) increase of sensitivity in the MS analysis. The use of affinity-based methods for phosphoprotein and phosphopeptide enrichment have allowed HUPO researchers to chart thousands of phosphorylation events in cells and tissue.

A Human Proteome Project

Pilot Phase

In the pilot phase program the temporal dynamics in protein phosphorylation during cell differentiation or cell activation will be studied. A model system will be the differentiation of embryonic stem cells, in conjunction with the HUPO Proteome Biology of Stem cells Initiative. Tools will be developed to systematically study micro-heterogeneity, diversity and dynamics of the phosphorylated proteins and their (changing) role in interaction and activity of protein complexes during cellular differentiation.

Full-Scale Phase

A major limitation in current phosphoproteomics are the expected ones of sampling and database limitations. The generated phosphoprotein data will be gathered in global databases and used in signaling network analyses. Moreover, the data will be used to generate new phospho-specific antibodies and possibly phosphopeptide micro-arrays for validation experiments and clinical screening of dynamics changes in protein phosphorylation.

Cells:

Human embryonic stem cells will be used.

Protein Glycosylation

The most abundant post-translational modification is glycosylation, estimated to occur on 50-80% of all eukaryotic proteins. Glycans produce analytical challenges because they are usually branched, are found N-linked to Asn and O-linked to Thr/Ser, and are heterogeneous at a given glycosylation site owing to the non-template driven biosynthesis. Recent advances have revolutionized the analysis of glycans using mass spectrometry to a level now practical for proteomics initiatives.

The sensitivity of mass spectrometry for glycans and glycopeptides allows for the analysis from material obtained from high throughput protocols using specific antibodies to immunoprecipitate the protein from tissues and fluids. Here, a profiling of the glycans of each glycoprotein can be achieved providing important information about the protein in question as well as how a given protein might differ between cell types and indeed, to samples obtained from patients exhibiting a particular disease or disease state. There are challenges: The complex heterogeneity of mucin-type O-glycans demands good separation methods for the high number of isomers, as do the glycosaminoglycans that require a partial oligosaccharide cleavage and special care to prevent loss of sulfate groups. Secondly, the sites of glycosylation and site-specific patterns are best determined by tandem MS analysis of the glycopeptides obtained by protease digestion. For the N-glycans, the consensus sequence for attachment is known and the glycan profiles at one glycosylation site can be more easily identified. Such direct approaches are of course ideal and can sometimes be used as the first and only analytical step. However, additional developments, both instrumental and computationally, are expected to further such direct approaches considerably.

A Human Proteome Project

The HUPO Glycomics initiative has undertaken a process to define the most reliable protocols for glycoprotein characterization at high throughput. These are already available for N-glycans, whereas for others (O-glycans), an ongoing evaluation of methods in different laboratories suggests methodological consensus. A major project within the HUPO Glycomics initiative has already focused on a 5-year program to discover, develop and clinically validate cancer biomarkers by targeting the glycan part of the protein (or lipid). Current development of sensitive LC-MS systems and novel types of post-source ion fragmentation methods holds promise to facilitate glycan analyses. A major challenge, as outlined in a NIH white paper [24], is the urgent need for open access glycan databases and better tools for rapid analyses of mass spectrometry information. A concerted effort to perform high-throughput analysis of the glycoproteome would go a long way to addressing the challenge of posttranslational modifications of the proteome.

Human Proteins Interaction Networks

Executive summary

The objective of this component of the HUPO Project is to define protein interaction networks for various human and mouse cell types and tissues with the purposes of helping define the biological roles of human proteins on a genome-wide scale and facilitating the development of diagnostic and therapeutic agents

Proteins must interact with other molecular components of the cell, including other proteins, DNA sequences, RNA molecules and various metabolites, to exert their functions. Mapping protein interaction networks in cells and tissues will produce fingerprints of the physiological states of these cells and tissues; similarly, mapping changes in protein interaction networks occurring during disease progression will generate signatures for specific pathological states. Because protein interaction maps represent multi-variable, complex descriptions of physiological/pathological states, they also provide the descriptions needed to more realistically and reliably address issues such as the causes, the diagnoses and, eventually, the cures of human diseases.

This repertoire of human protein interaction maps, termed the Human Proteotheque, will be built via a concerted, inclusive initiative, the Human Proteotheque Initiative (HuPI), will involve discovery platforms aimed at defining protein-protein, protein-DNA, protein-RNA and protein-metabolite interactions, integrating data into comprehensive protein interaction maps through bioinformatics.

Introduction

Citizens of developed countries expect biomedical research funded by their governments to deliver cures and diagnostic tools for diseases that threaten their health, and meeting this expectation will enjoy tremendous popular support, especially if, in addition, it stimulates the economy of their country. Genomic and proteomic research has made significant progress in this direction by identifying most of the molecular components

A Human Proteome Project

that make up a human being, including catalogues of human genes and their RNA and protein products expressed in various cells and tissues of healthy or diseased origin. Alone, this knowledge is not sufficient to explain the causes of the most common human illnesses, including cancer, diabetes, asthma and heart and neurodegenerative diseases. We now need to learn how the various molecular components of the cell interact and cooperate to sustain function, as well as how aberrant interactions can lead to disease. In other words, biomedical research now needs to produce comprehensive maps of the molecular interactions that underlie human cell function in health and disease. This knowledge is the next essential step in the race for cures and diagnostics for important diseases. It will also serve indirectly as a motor for the creation of a new generation of biotech companies that will create jobs in developed countries.

Proteins are the central functional components of human cells. They are involved in almost all cellular processes, and protein aberrations have been shown to have a causative role in many diseases. Most importantly, proteins must interact with other molecular components of the cell, including other proteins, DNA sequences, RNA molecules and various metabolites, to exert their functions. Mapping protein interaction networks in cells and tissues will produce fingerprints of the physiological states of these cells and tissues; similarly, mapping changes in protein interaction networks occurring during disease progression will generate signatures for specific pathological states. Because protein interaction maps represent multi-variable, complex descriptions of physiological/pathological states, they also provide the descriptions needed to more realistically and reliably address issues such as the causes, the diagnoses and, eventually, the cures of human diseases. Mapping protein interaction networks in health and disease is a tremendous scientific and technological challenge that will require huge efforts from many groups of scientists and huge resources from funding agencies.

Here, we propose an international research project that would unite scientists from many disciplines and various health research areas in an initiative aimed at building a repertoire of comprehensive maps of protein interaction networks in health and disease. This effort will build on the success of several large-scale projects conducted in various research centers around the world. This repertoire of human protein interaction maps, that we propose to name the Human Proteotheque, will be built via a concerted, international initiative that will involve a multi-site discovery platform aimed at defining protein-protein, protein-DNA, protein-RNA and protein-metabolite interactions, and integrating the data into comprehensive interaction maps through bioinformatics. This effort is at the heart of an emerging discipline, Integrative Systems Biology (ISB), which is aimed at developing tools and concepts to generate complex descriptions of biological systems considered globally.

The Human Proteotheque Initiative (HuPI) will have a tremendous structuring effect on biomedical research world-wide by fostering the integration of efforts in genomics, proteomics and bioinformatics and creating a wealth of new knowledge related to human cell function. It will also stimulate the application of this massive new resource by new biotech and large pharmaceutical companies, creating jobs for citizens of the participating countries and developing cures and diagnostic tools for important diseases.

A Human Proteome Project

Protein Interactions Networks in Health and Disease

Human cells function through the action of thousands of proteins that control their growth and differentiation. Most human proteins rarely work alone, but rather assemble with other proteins into complexes to exert their function. In addition, functionally related protein complexes forming specific cellular machineries cooperate during various biological processes, including gene transcription, DNA replication and repair, mRNA translation, signal transduction, etc. This situation is complicated by the fact that any given polypeptide could assemble into more than one protein complex, making the network of protein interactions sustaining cell growth and differentiation not only very complex in its organisation, but also in its dynamics of assembly. Yet another complication is that any given protein might change its interaction partners from one type of cell or tissue to another and during development from the fertilised egg to the adult organism. This unique situation highlights the pivotal role of protein interaction networks in cell function. Consequently, mapping their topology is a key issue in biomedical research, and the development of efficient technologies for doing so is an important challenge to modern research in proteomics and systems biology.

If we postulate that any given interaction made by a particular protein within a cell has evolved for a specific function, one can envision two main reasons for attempting to build complete and accurate protein interaction maps and for making them available to the scientific community. First, protein interaction networks, considered here as the set of interactions a protein makes with other proteins, DNA and RNA molecules and metabolites, are a fingerprint of the physiological status of a cell and their modulation is predicted to be a signature of specific disease conditions, including those observed in cancers and viral infections. Publicly-available protein interaction maps will undoubtedly accelerate the discovery process in biomedical research, as they will reveal new, more global (*i.e.* systemic) molecular descriptions of specific cellular conditions such as those encountered in disease; these maps may well represent the new generation of biomarkers, being more accurate and specific as they are based on multiple parameters. These maps will also reveal new targets for drug discovery. Secondly, a large fraction of the proteins encoded by the human genome remains uncharacterized and their precise functions unknown. By identifying protein interaction partners, it becomes possible to infer putative functions for many previously uncharacterized proteins. This method for defining protein function based on “guilt by association” has proved powerful in many systems. Consequently, defining protein interaction networks is invaluable for deciphering protein function in health and disease.

The literature contains a myriad of papers reporting on protein interactions. Efforts to curate and integrate these interactions into public databases have emerged in various projects around the world. These projects are important and valuable. However, because of the heterogeneity of the data and the experimental procedures used to derive the various datasets, integration efforts become extremely difficult and their results raise important questions in terms of their completeness and accuracy. This notion of accuracy and completeness is a highly relevant issue that needs to be considered seriously when developing protein interaction maps that are highly valuable and, as far as possible, not

A Human Proteome Project

misleading. Clearly, building complete and accurate protein interaction maps would be of immense value but would also represent a major challenge.

The Human Proteotheque Initiative (HuPI): Building a repertoire of comprehensive maps of the human protein interaction network

This project is aimed at generating comprehensive maps of the protein interaction networks that underlie human cell functions. The core of this project is represented schematically in Figure 5.

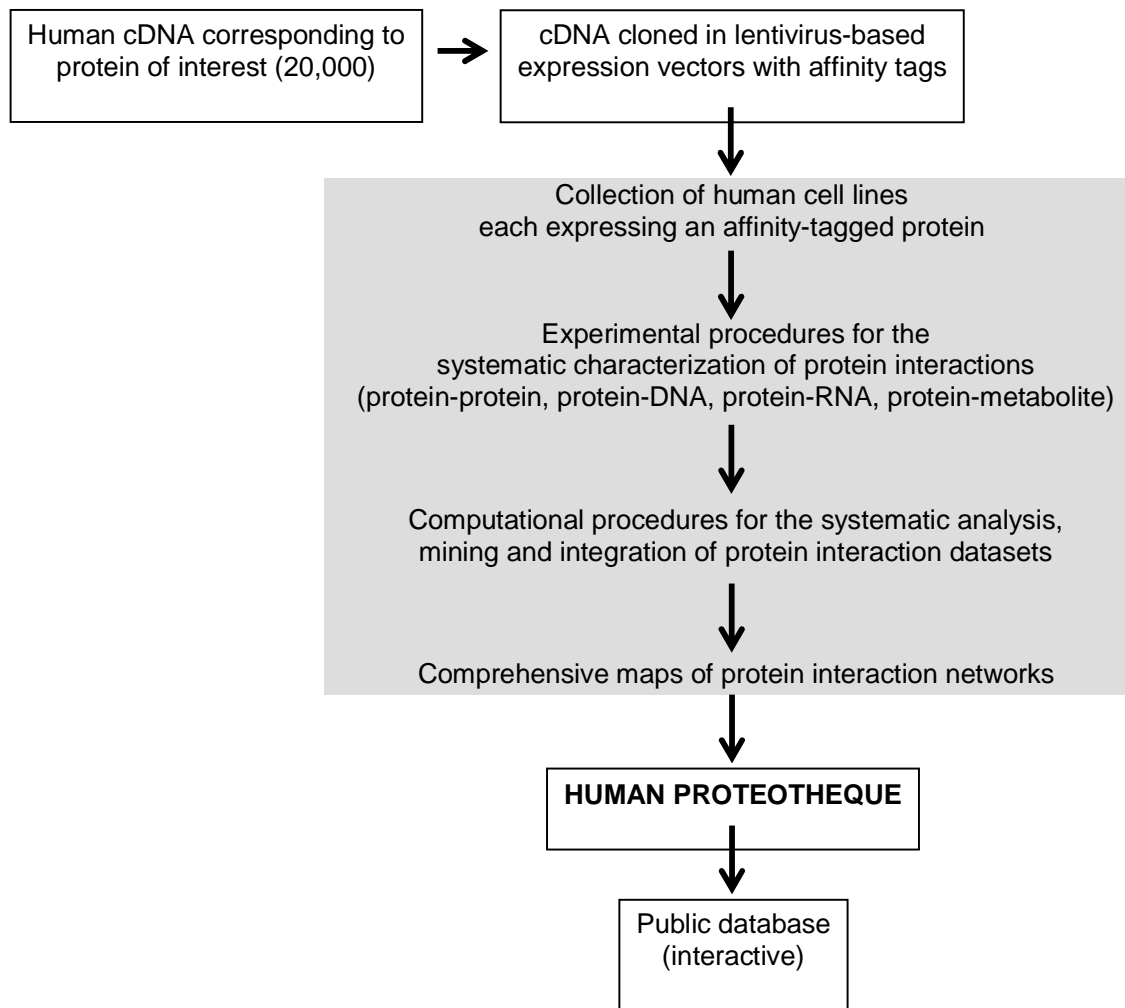


Figure 5. Overview of the core of the Human Proteotheque Initiative (HuPI). A collection of cell lines each expressing affinity-tagged proteins is used as a systematic, unbiased discovery platform to identify human protein interactions and characterize human protein interaction networks. Computational procedures are used to select high-confidence protein interactions and to build comprehensive maps, the HuPI-Maps, of high-density interaction networks.

The direct identification by purification and mass spectrometry of protein interactions in human cell lines, as shown in Figure 5, has major advantages but also two key disadvantages. The first relates to the availability of human cell lines. Cell lines are not available for many types of human cells and those that are available are usually

A Human Proteome Project

transformed and far from normal. Because we expect many protein interactions to be different in different types of cells, many normally occurring interactions will be missed if this is our sole method of identifying protein interactions. This problem can be solved in large part by turning to the laboratory mouse as a model for protein interactions, as shown in Figure 6.

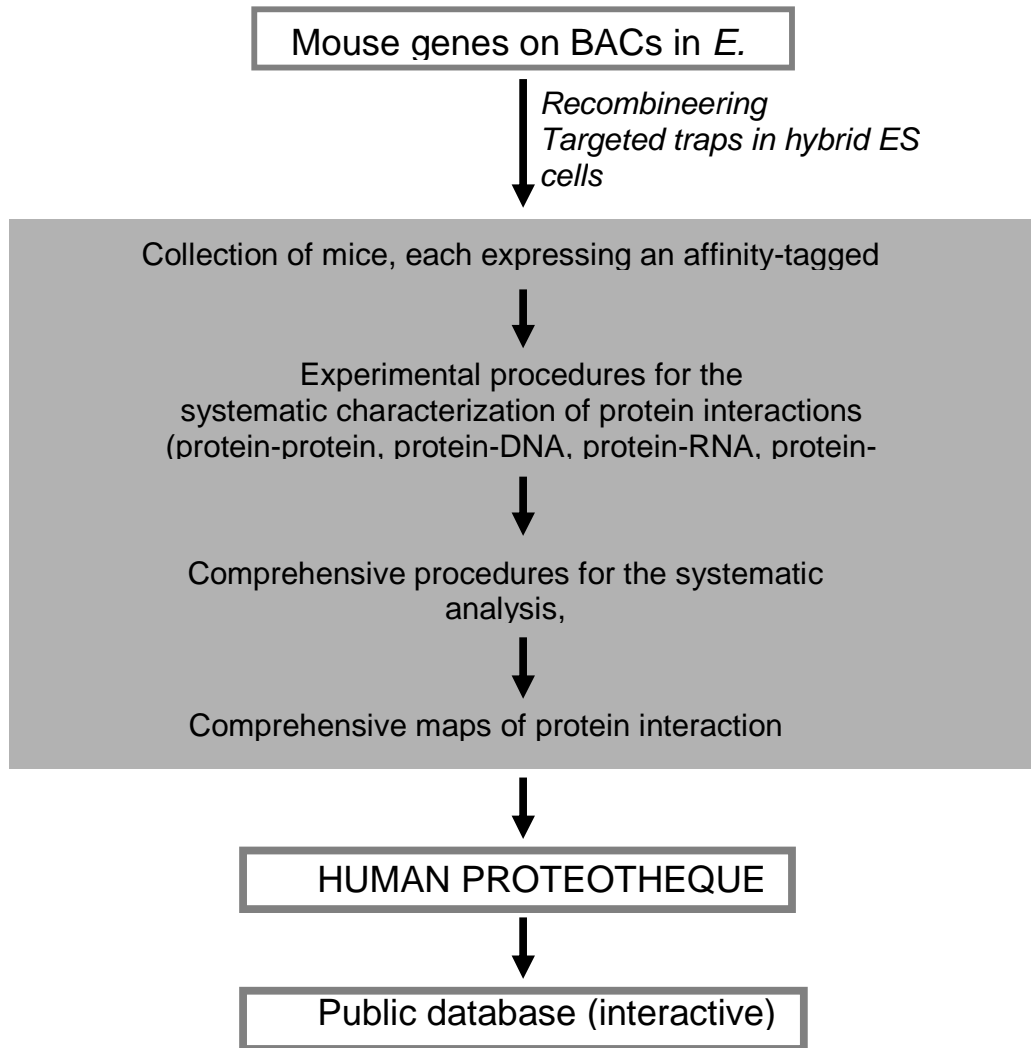


Figure 6. Overview of the core of the mouse component of the Human Proteome Initiative (HuPI). A collection of mice, each with a knocked-in, affinity-tagged protein is used as a systematic, unbiased discovery platform to identify mouse protein interactions and characterize mouse protein interaction networks. Computational procedures are used to select high-confidence protein interactions and to build comprehensive maps, the HuPI-Maps, of high-density interaction networks.

The construction of mice containing tagged proteins has several major advantages. One is that the protein would almost always be expressed at its normal level so that spurious interactions would tend to be avoided. Another is that the tag will be used in immunohistochemistry experiments to identify the mouse cells and tissues in which the protein is expressed. This will serve as a check on protein expression as determined

A Human Proteome Project

using affinity reagents (see above) for cases in which affinity reagents do not exist or give misleading results. The use of affinity reagents will serve as a control for those cases in which the presence of a tag leads to protein misexpression in the mouse or in which protein expression differs in mouse and human. Thirdly, the tagged mice will be used in immunoassays to quantify the proteins in various types of mouse cells and tissues as a check on parallel direct quantification by mass spectrometry (see below). Fourthly, the tag will be used in immunofluorescence experiments for the intracellular localisation of the protein in various types of mouse cells and tissues, again as a check on parallel methods of intracellular localisation by either affinity reagents or cellular fractionation and mass spectrometry. Finally, the tagged mice or ES cells can be provided as a resource to the international community so that mouse crosses can be done to analyse the effects of disease-causing mutations or processes on protein interactions. The relevance of interactions identified in the mouse for human cells will be tested by directed experiments with affinity-tagged human cDNAs in appropriate human cells.

The identification of protein interactions entirely by purification and mass spectrometry, while quite accurate, leaves open the issue of confirmation which, if done by an independent method, would make the interaction maps substantially more reliable. Confirmation could be achieved by constructing sets of two sets of human cDNAs with two different kinds of tags, then carrying out co-transfection and co-precipitation experiments with appropriate human cell lines.

These protein interaction network maps will be deposited in a repertoire, termed the Human Proteotheque, which will be made publicly available as an atlas describing fundamental human and mouse molecular networks. To generate valuable tools for scientists interested in basic biological and biomedical research, this project needs to be developed in conformity with three main criteria. First, the data must be both accurate and complete. In other words, both the specificity and the sensitivity of the interaction datasets must be set to the maximum achievable while developing and applying the overall experimental procedure. To attain this goal, we propose to develop an experimental platform in which data acquisition and analysis is performed in a highly systematic manner, favouring automation when possible to prevent bias that is sometimes introduced as a consequence of human decisions and error in experimental setups. Secondly, the data must be analysed, mined and integrated in such a way that (i) all the relevant information is extracted and stored and (ii) this information is transferred into interaction maps that are comprehensive. As mentioned above, systematic procedures must be developed to ensure that the data is analysed in an unbiased manner. For this reason, computational approaches are critical for the downstream part of the pipeline (Figs. 3, 4). Thirdly, both the protein interaction datasets and the comprehensive maps of their network connections must be made available to the scientific community through the internet. A web-interfaced database will be developed to help the users find all the information relevant to their research. Graphical tools for the visualisation and navigation of the protein interaction maps will play a central role in the development of the project's database and web site.

A Human Proteome Project

Clinical Implications

As mentioned above, defining maps of protein interaction networks in cells and tissues will produce fingerprints of the physiological states of these cells and tissues. Similarly, mapping changes in particular protein interactions occurring during disease will generate signatures for specific pathological states. Because protein interaction maps represent multi-variable descriptions of pathological states, they also provide the descriptions needed to more realistically and reliably address issues such as the causes, the diagnoses and, eventually, the cures of human diseases. In particular, the protein interaction networks we identify for normal cells will be combined with information about changes in protein expression observed as a consequence of disease (see below) to predict alterations in the protein interaction networks caused by disease. As well, defining the interaction partners of proteins known to participate in the establishment and/or evolution of disease (onco-proteins, tumour suppressors, etc.) is one of the most efficient methods of identifying new proteins involved in the same disease. The information we generate will serve to complement and enhance SNP-based mapping of disease genes and thereby lead to the identification of a new generation of potential biomarkers and disease targets.

Process

Building a repertoire of comprehensive maps of protein interaction networks that will systematically enhance our understanding of both normal and disease conditions and, eventually, lead to the development of diagnosis tools and cures for important diseases, requires the concerted participation of large groups of scientists with relevant expertise. Fortunately, a substantial number of groups in various countries have developed such expertise, and this project proposes to unite these groups in an international collaboration that may well revolutionize biomedical research for years to come.

The project can be divided into 6 distinct parts:

1) **Formation of the HuPI consortium** (Time line: 2008). This first step of the project is aimed at forming a representative international consortium of scientists bringing to the table two types of expertise. First, the consortium will include researchers with technical and conceptual expertise in the experimental characterization of protein interactions and in the creation of comprehensive maps of interaction networks using computational biology. Secondly, the consortium will include researchers with expertise in relevant biological and disease systems, including mouse genetics. The role of the consortium will be to (1) select the cell lines and proteins to be targeted by the project's discovery platform, (2) plan the deployment of the discovery platform at various sites according to the cells and proteins selected in (1), and (3) design standard operation procedures (SOPs) for characterizing protein interactions. This includes the choice of affinity tagging, purification and mass spectrometry methods, specifically selected so that the datasets produced at different sites will be compatible and can be integrated into large interaction networks.

Deliverable: A detailed plan, including a list of participating centres and laboratories, a list of proteins to be targeted by the discovery platform, a list of recipient human cell lines, and protocols for SOPs, will be produced (December 2008).

A Human Proteome Project

2) **Development and deployment of SOPs for the characterization of protein-protein interactions** (Time line: 2008-2009). This second step is aimed at developing the multi-site, integrated discovery platform for defining protein-protein interactions. The ability of the various sites to efficiently adhere to the SOPs will be evaluated by conducting a pilot project aimed at analysing a set of samples at the various locations; additionally, **quality control experiments will be conducted at different stages of the productive phase to ensure that equivalent datasets are produced at the various locations (see below)**. While deploying the SOPs, special care will be devoted to ensure both maximal efficiency (*e.g.* sensitivity and throughput) of mass spectrometry methods and maximal sensitivity and specificity in computational methods for establishing high-confidence datasets.

Maximizing the efficiency of mass spectrometry. As MS is central for identifying protein-protein interactions in an unbiased way, we will implement novel MS methods relying on the high sensitivity and high mass precision of the new generation of mass spectrometers, as well as the high throughput of the various procedures. Improved methods for sample fractionation that do not require gel analysis and automation of both the MS procedure and the MS data analysis pipeline will be implemented by the discovery platform.

Adapting computational methods to maximize the sensitivity and specificity of our high-confidence interaction datasets. Computational methods capable of minimizing the rates of both false-positives and false-negatives will be implemented.

Deliverables: Data from pilot experiments on 100-200 proteins for evaluating the effectiveness of the multi-site discovery platform in running the SOPs for the systematic characterization of protein-protein interactions (December 2009). The data produced in the various locations must be compatible so as to allow integration into interaction networks.

3) **Development and deployment of SOPs for the characterization of protein-DNA, protein-RNA and protein-metabolite interactions** (Time line: 2009-2012). This step is aimed at integrating procedures for defining protein-DNA, protein-RNA and protein-metabolite interactions into the multi-site, integrated discovery platform. This part of the project will require the optimization and development of a number of technologies.

As mentioned above, most proteins interact with other molecular components of the cell. To integrate protein-DNA, protein-RNA and protein-metabolite interactions into the interaction maps, many laboratories are currently involved in various technology development projects. For example, our previous work has shown that TAP-tagged proteins can be used in chromatin immunoprecipitation (ChIP) experiments aimed at defining their locations along the genome [30, 31] when coupled with the identification of the immunoprecipitated DNA fragments using systematic ChIP-on-chip experiments [32, 33] or ChIP combined with newly developed ultra-fast DNA sequencing methods, this method promises to reveal protein-DNA interactions that could then be integrated

A Human Proteome Project

with protein-protein interactions. These are also goals of the ENCODE consortium and the International Regulome Consortium (IRC), and the human cell lines and mouse ES cells that we create with tagged transcription factors will be provided to those consortia for DNA interaction analysis. For yeast proteins involved in RNA metabolism, we have previously identified the particular non-coding RNA molecules that are associated with various yeast protein complexes, and these methods could be adapted for the identification of human and mouse protein-RNA interactions. These efforts also require the development of computational methods for the integration of the various datasets.

Deliverables: Robust methods for the systematic characterization of protein-RNA and protein-metabolite (PTMs) interactions. Data from pilot experiments.

4) **Production phase for the characterization of protein-protein interactions** (Time line: 2008-2015). This part of the project will be conducted according to the plan developed in (1) by the consortium members. We expect that up to 20,000 proteins will be affinity tagged, expressed in human cells (a fraction of this number in multiple cell lines) and mice, purified and submitted to MS analysis. High-confidence interactions will be selected computationally and re-tested by co-transfection and co-precipitation. Prioritization to establish the order in which the various proteins of the human and mouse proteomes will be subjected to interaction analysis will be based in large part on input from the scientific community at large.

Deliverables: Many large datasets of high-confidence protein-protein interactions in various cell lines, as well as mouse ES cells and various tissues. Human cell lines and mouse ES cells that express various tagged proteins.

5) **Production phase for the characterization of protein-DNA, protein-RNA and protein-metabolite interactions** (Time line: 2008-2015). This part of the project will be conducted according to the plan developed in (1) by the consortium members.

Deliverables: Some large datasets of protein-DNA, protein-RNA and protein-metabolite interactions in various cell lines and selected mouse cell types and tissues. The protein-DNA interactions will be determined, in large part, by the ENCODE and IRC consortia.

6) **Integration of the various datasets and establishment of comprehensive interaction maps; development of the publicly available, interactive web site** (Time line: 2009-2014). The development of computational and theoretical tools enabling the construction of relevant, valuable maps will be carried out (for an example of an early-generation protein interaction map, see Fig. 7). The complexity of the networks that are obtained will be such that they can easily become overwhelming, thus requiring automated, goal-oriented network layout procedures that facilitate the extraction of meaningful biological information.

Deliverables: Many comprehensive, integrated and navigable maps of protein interaction networks deposited in a repertoire, the Human Proteotheque, which can be consulted through an interactive web site also to be developed during this phase of the project.

A Human Proteome Project

A Human Proteome Project

Conclusions

Building comprehensive, meaningful maps of human protein interaction networks and making them available to the scientific community through the internet would be the main goal of a HuPI component of the HUPO project. Recent progress that resulted in the publication of key papers on protein interaction networks in yeast, bacteria and human cells indicates that this objective is now achievable. In addition to assigning putative functions to previously-uncharacterized human proteins and increasing our understanding of the proteome and its regulation, we will identify the cells and tissues that express any particular protein, quantify the various proteins in those tissues, and establish the intracellular localization of the various proteins. The human cell lines and mouse ES cells that we create will be provided as a resource to the community at large. HuPI maps will provide a genome-wide, systems-based description of the relations between proteins and other molecules (proteins, DNA, RNA, metabolites) in human cells. Such a complex molecular description of the physiological status of a cell or tissue should be invaluable for the development of a new generation of effective therapeutics and biomarkers that are both sensitive and specific because they rely on more accurate, multi-parameter indicators.

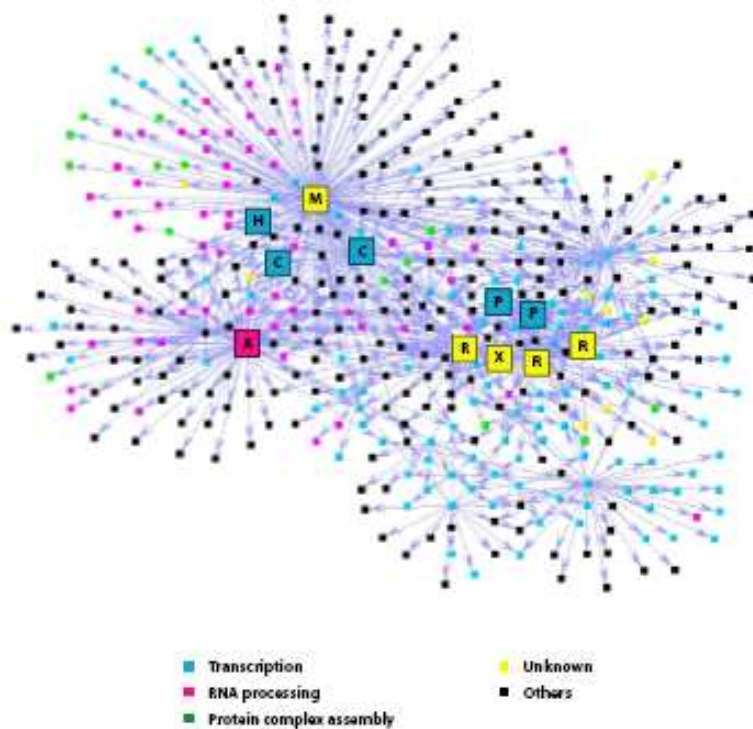


Figure 7. Map of a high-density network of high-confidence protein-protein interactions involving transcription and RNA processing factors. Affinity-tagged proteins (baits) and their co-purified interaction partners (preys) are represented as coloured squares (nodes) and are connected using arrows. The color code is defined. The previously uncharacterized proteins MEPCE/BCDIN3 (M), RPAPs (R) and XABI (X) were assigned putative functions based on their associations with the well-characterized proteins HEXIM1 (H), P-TEFb subunits (CDK9 and CCNT1/Cyclin T1) (C), hnRNPA1 (A); and, the RNA polymerase II subunits (Rpb2 and Rpb11) (P). Map derived from the data in Jeronimo et al. [8].

A Human Proteome Project

Integrative Bioinformatics

Introduction

A level I bioinformatics platform will process data generated for each of the technologies represented in Figure 1. The level II bioinformatics platform will provide an integration layer and incorporate all related data as seen in Figure 8.

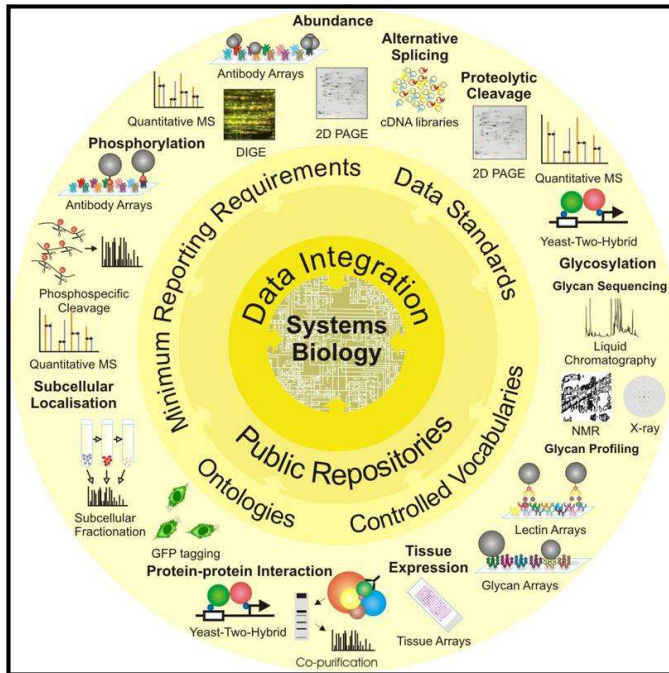


Figure 8. Integrated Bioinformatics. See Mueller et al. [34].

Mission statement

To build the reference portal for the human proteome. The construction of this portal will require the development of an infrastructure that makes information of all the individual proteomic domains truly *accessible* to the community.

Process

Each of the three major data components in this proposal, namely antibody profiling, proteomics profiling, and interactions profiling, will perform their own level I bioinformatics including quality assurance analysis and make the results available in their own resource.

The UniProt KnowledgeBase [35] provides an outstanding high-level view of manually and automatically annotated protein information with alternate identifiers, functional annotations, sequence annotation at the protein level, and cross-references to many other resources that the user can follow. The Human Protein Reference Database (HPRD) [36]

A Human Proteome Project

also displays the results of literature curation efforts, along with community contributions made via Proteinpedia [37]. The Ensembl database [38] provides a highly advanced interpretation of the human genome at the DNA, transcript and protein isoform level. The International Molecular Exchange (IMEx) Consortium groups together molecular interactions databases like DIP [39], IntAct [40] and MINT [41]. PeptideAtlas [42] and the Proteomics Identifications Database (PRIDE) [43], both partners in the ProteomExchange data exchange collaboration [44], capture mass spectrometry proteomics data gathered from a large number of laboratories from around the world. Three dimensional protein structures are contained in dedicated databases such as the Protein Data Bank (PDB) [45], and the Molecular Structure Database (MSD) [46], which collaborate in the wwPDB consortium [47].

These infrastructure components essentially fulfill two roles. First, the annotation of positional features (e.g. posttranslational modifications, protein domains, antibody epitopes and structural elements such as beta sheets or helical structures). Second, they allow complex queries spanning multiple resources to take place (e.g. ‘find all the UniProt proteins that have known three-dimensional structures, that are located on chromosome 21, with at least 5 distinct peptides observed in mass spectrometry databases, and that resulted in a positive immunostaining in liver’).

Positional features are resolved by the Distributed Annotation Server (DAS) protocol [48], while cross-resource queries are enabled by implementing BioMarts [49] in the relevant resources.

The DAS protocol is designed around a lightweight XML format that allows a resource to respond to a simple, identifier-based query with a collection of the positional and non-positional features it knows for that identifier. By aggregating the information retrieved from a variety of such resources, a client program can construct a very broad view on all the known positional features obtained across all relevant technology platforms.

A BioMart is at its core a specialized database concept that is optimized for query speed. A wide array of automatically generated programmatic interfaces is provided on top of this concept, which presents a choice of access mechanisms to prospective data consumers. Furthermore, a standardized web interface is automatically constructed that actively presents the data contents of the resource to the user for filtering and retrieval. Finally, the automatically provided interfaces are sufficiently uniform across different BioMart implementations to easily allow integrative queries across multiple BioMarts.

Objectives

The integrative bioinformatics component will provide a reference interface to all the data generated in a human proteome project and already present in existing community resources. The interface will enable access to this wealth of information to the wider biomedical community, and will provide a concise and meaningful presentation.

A Human Proteome Project

This portal will explicitly rely on existing resources rather than compete with them. Note that this strategy preserves the fine level of detail that is inherent in these repositories so that users who wish to can always retrieve the primary data.

In order to effectively use the existing resources, adaptations to make them versant in existing protocols will be needed. An important immediate benefit of having these standardized conduits implemented in these resources is that it makes the information in them readily *accessible* to the entire community as well.

Implementation

The project requires support at the individual databases for the implementation of the two standard protocols, as well as a core team of developers to provide the central reference interface.

It is critical that the design of the reference portal starts simultaneously with the implementation of the protocols at each of the contributing databases. Synchronizing these efforts will enable an iterative development process in which the core developers can continuously interface with the bioinformaticians at the contributing resources. This is important to ensure that the implementation by the data providers at the individual databases provides the necessary information required by the presentation at the reference portal.

The building of the infrastructure at the individual databases requires 1 FTE for two years at each of the databases. Each developer will become a member of the individual resource team, implement the necessary protocols at the resource site, and make the information there available via these protocols. Further, the developer will work with the portal team to ensure that the resource will provide the necessary information with the necessary speed.

The reference portal team will require 5 FTEs for the first two years, which will be expanded to a team of 15 FTEs for the next five years. The first two years will focus on requirements gathering, functional analysis, and prototyping, while the next five years will carry out the actual implementation of the reference portal interface(s).

When the infrastructure is completed, the project will move into its production phase. This phase aims to improve the consistency of the information accessed via the portal, and to expand the number of resources it covers. The integrated information for each protein will be reviewed via targeted curation and annotation of centrally selected proteins of interest, and potential discrepancies will be examined in detail for resolution. The number of resources covered by the portal will be expanded to include additional resources that were either not available or not selected for the initial phase. Furthermore, more advanced visualization tools will be built on top of the portal to allow users to make sense of the ever-growing volume and diversity of the integrated information. The core

A Human Proteome Project

development team requires 5 FTEs for five years for this second phase, while each contributing resource will require 1 FTE for these five years.

Once the seven-year project span has completed, continued funding for at least 5 FTEs will be necessary in order to maintain the portal interface and to further serve the community. These 5 FTEs will essentially continue the work of the second phase, and will therefore follow up on discrepancies, provide updates to the interface and tools, the addition of new resources developed as part of other projects. Similar to previous large scale efforts at cataloguing and disseminating large-scale data, these FTEs could be integrated in existing life sciences data management organisations, for instance the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI).

All software developed within this component will be released under a permissive open-source license (e.g., Apache2 or Creative Commons Attribution) which carries no limitations to its adoption and re-distribution by any third party, so long as original authorship is acknowledged.

Output

There are two major deliverables for this bioinformatics component.

Deliverable 1 is the reference portal for the Human Proteome. This will allow all levels of users to explore the human proteome as integrated from multiple resources by actively combining live information from multiple sources via a single interface.

Deliverable 2 is the implementation of both DAS and BioMart as standard protocols at each of the contributing databases. The reference portal will use these two protocols to gather its information and any other members in the community will be able to use exactly these already established protocols to access information from all contributing databases.

Synergies with the other HUPO Project components

The bioinformatics component effectively glues together the findings of the different components. This enhances the existing synergies between the other three components as they can readily access the information generated by the other components via the portal. Furthermore, the communications infrastructure built as part of the portal development will also enable each component to act as direct consumers of any of the contributing databases.

Outcome

We envision this reference portal to become a prominently visible aspect of the Human Proteome Project effort. It will have a major impact in the field of clinical research as it

A Human Proteome Project

will provide a powerful platform to support the development of diagnostics, prognostics and treatments.

It will also provide academic research with an unprecedented, integrated, comprehensive and continuously updated tool to fuel hypothesis-driven research.

Finally and importantly, the reference portal will also be built with an outreach area wherein it will provide the general public with a window into the ongoing efforts to understand the human proteins that are the machinery of life.

Signaling Pathway Profiling

Executive Summary

Cells can be considered to be complex networks of interacting molecules and the analysis of intracellular molecular signaling pathways is much less complex as the complete understanding of the products of the about 21.000 human genes. Pathways are conserved throughout the animal kingdom and the link between disease and signaling pathways has been firmly established and validated for many genetic disorders. Reverse-protein microarray approaches are ideally suited for analysis of signaling pathways and the currently already available set of about 200 highly validated antibodies will allow a quick entry in the creation of pathway activity atlases, to be expanded over the term of the project to cover substantially the “signamoles” of disease relevant cell lines and tissues. With the ongoing generation of new, highly validated antibodies, comprehensive quantitative pathway activity atlases will be generated to define normal values for pathway activities and protein expression for a wide variety of cells and tissues.

Introduction

Protein kinases are key actors in the cellular signal transduction networks, regulating cell growth, survival and differentiation. Aberrant activation or deactivation of such kinases by overexpression, deletions or mutations can lead to deregulation of cellular signaling cascades and has been shown to be associated with numerous diseases.

The phosphorylation status of signaling pathway components can be measured using anti-phosphoprotein antibodies that specifically recognize the phosphorylated isoforms of such kinase substrates. Thus, the activity status of multiple signaling pathways can be probed through parallel phospho-specific analysis. Besides the laborious western blot, which allows only a limited throughput, the current gold standard for this purpose is the sandwich-ELISA, which is available in many custom or commercial formats. Sandwich-ELISA's have the disadvantage that a carefully matched pair of antibodies must be developed, if one aims for site-specific analysis. Moreover, the (peptide) epitopes that have been used for the generation of the antibodies are not always accessible in the native (non-denatured) protein.

A Human Proteome Project

Recently, lysate arrays have emerged as an alternative to the sandwich (or forward) assay format. This type of array, in which a protein extract is immobilized and queried with antibodies or other reagents that bind to a specific protein in the sample, is often referred to as reverse (phase) protein micro-array (RPA). In contrast to this, forward arrays are those in which the capture reagent (e.g. the antibody) is immobilized. Among the first applications of RPA were microarrays of tissue lysates to study many proteins in micro-dissected biopsies.

RPAs have been used for several years in their most basic form, the dot blot, in which drops of cell or tissue extract are applied to a membrane or a coated glass slide. Among the different proteomics technologies that are suitable for that purpose, we propose a reverse array platform utilizing the planar wave-guide technology allowing for detection of a minimum of 1000-2000 molecules to be present in a single spot, where the total protein constant of a single spot corresponds to a single cell equivalent. Planar wave-guide, reverse phase protein arrays make it feasible to obtain reproducible and quantitative protein expression information about the dynamic aspects of cell signaling. Samples are spotted in serial dilutions on the array to obtain an on-array dose-response curve of the assay, and hence relative or absolute quantification (using an internal standard) is possible. Cells or tissue samples are subjected to a one-step extraction using denaturing conditions, under which the potentially labile protein phosphorylations are effectively 'frozen', rendering most peptide epitopes accessible and making it rather easy to translate antibody validation by western blotting into an array format. As the quality of the antibodies is key to the successful application of reverse phase arrays, a significant effort is required to their validation before the antibodies are applied.

Process

Reverse protein arrays

In a reverse protein array, a sample is immobilized and queried with labeled proteins (such as antibodies, members of protein complexes) or other reagents that bind with a specific protein in the sample. Samples are titrated (serial dilutions) (Figure 9) on the array to ensure that the assay is carried out in the linear range of the binding curve. Reverse protein arrays are highly sensitive, linear and quantitative. Cell lysate arrays (reverse arrays) enable the investigation of proteins analyte sets in crude proteomic samples with low amounts of starting material in high throughput mode due to the parallel approach provided by array technologies

A Human Proteome Project

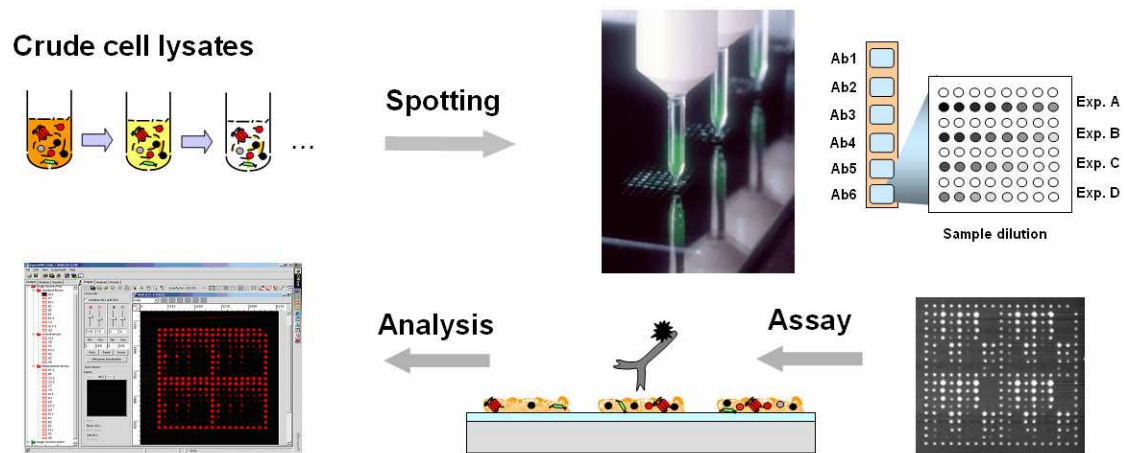


Figure 9: Crude cell lysate samples (produced with a denaturing solution) are spotted into arrays onto specially prepared glass chips by Inkjet Technology. The samples are probed by e.g. fluorescently labeled antibodies.

Applications

Due to the high sensitivity and high throughput capability of the reverse protein array approach it will be feasible to obtain protein expression profiles and signaling pathway information on a wide variety of cell lines and tissue samples. Interesting applications include i) the comparative analysis of signaling pathway(s) events in normal versus diseased tissue, ii) the comparative analysis of protein expression in various systems, iii) the elucidation of the dynamic aspects of pathway events and iv) the profiling of compounds to reveal signaling and cross-pathway effects of drug candidates. In addition analysis of healthy versus disease tissue (including animal models) will provide insights into the pathways underlying pathologies and provide a platform for Molecular Diagnostics. In a future approach, the screening of body fluids with a reverse array approach may enable to investigate a large number of individual body fluid samples for a limited set of proteins contained in them, to establish variation in protein expression levels.

Deliverables

Objectives and timelines for funding

1. Definition of canonical pathway activities in the different cell lines in a quantitative manner selected for study in the overall project including the treatment with at least 2 inhibitors for each pathway analyzed in a concentration and time dependent manner
2. Description of quantitative signaling pathway profiles in the healthy organ tissue selected for other studies in this project to provide a comprehensive atlas for normal expression and activity status of pathway components
3. Quantitative analysis of signaling pathways in diseased tissue selected for other studies in this project and selection of pathway relevant disease markers

A Human Proteome Project

Technology Development

A substantial part of the funding (10-20%) of the program should be devoted to technology development to improve the various steps in the process and to enable new technologies to be implemented. During the initial period, it is important to fund efforts to validate affinity reagents in large scale, as well as stream-lined efforts to create an accessible resource and query tools for the validated antibodies.

Synergies with the other HUPO subprograms

Substantial synergies exist between the signaling pathway profiling project and the other HUPO subprograms. Most notably, the mapping of the pathway relevant protein profiles in cells and tissues in a quantitative manner, e.g. the provision of pathway activity atlases, will benefit the other programs serving as a benchmark for the expression level of this subset of the proteome. This program will benefit from the availability of validated antibodies created in other subprograms. Cell and tissue specific quantitative pathway signaling information will assist in the validation of networks created within the subprograms.

Clinical relevance

In recent years progress has been made in ascribing pathological conditions to defects in molecular pathway components, e.g. linking dysregulation of signaling pathways to cancer and inflammatory diseases. Since kinases and phosphatases are key regulators in signaling pathways, it is not surprising that across the pharmaceutical industry a substantial percentage of drug discovery efforts are focused on targeting these enzyme classes. Especially the modulation of cellular kinase activities is one of the most rapidly growing areas in the development of novel drugs.

To understand the information flow through signaling networks and how these can best be manipulated to halt or redirect the flow of aberrant signaling is a challenging endeavor. A first step would be to describe the full complexity of signaling networks at a molecular level – their systems biology – including activities specific to a particular cell type, dynamic feedback mechanisms, pathway cross-talk, signaling kinetics and of course pathway activation states (see below) in normal and disease situations.

For a “kinase pathway” the information flow or pathway flux, mostly depends on the ratio of phosphorylated and non-phosphorylated protein species, reflecting the activation state of the biological system. If we compare cellular activity over time, at stages of disease progression or before or after drug treatment, it is likely that a correlation exists between the activation state on one hand and the biological or disease state on the other. Small molecules that modulate the activity of signaling proteins are useful tools to dissect the functional roles and connections of the individual nodes in a pathway. Using such a ‘systems approach’, one can begin to build a model that will not only provide a

A Human Proteome Project

contextual understanding of the molecular mechanisms of disease, but also has the potential to facilitate the validation of therapeutic modulation of regulatory and metabolic networks. A direct consequence of such an approach would be the early recognition of “off target” and side effects of drug candidates, as well as the identification of putative biomarkers.

Relevance for the Communities

Relevance to the Biology Communities

Progress in the fields of Cell and Developmental Biology has relied heavily on genomics-based resources to the extent that without this, many high impact discoveries would never have taken place. In light of this, an equivalent protein-based resource such as the human proteome will undoubtedly impact cell and developmental biology to the same degree for years to come. Knowing where and when and in what context a protein is expressed is at the heart of elucidating and understanding biological processes and function. This type of information, which is provided through the human proteome project along with reliable reagents, is also fundamental to the emerging field of Systems Biology. This field aims to elucidate biological function through systems analysis and modeling. So far, systems biology has mainly focused on simpler organisms such as Baker's yeast, *S. cerevisiae*, where network profiling and modeling have been deployed in an attempt to describe complex processes. Insights into such simpler systems such as yeast could potentially have relevance to mammalian biology though clearly, will fail to encompass the higher levels of complexity and sophistication for obvious reasons, i.e., yeast cells are unicellular. Nevertheless, tools used in modeling should be applicable and these include network analysis as well as stochastic modeling to explore multi-parameter space.

It is envisaged that a spatial and temporal protein map that includes sub cellular information should enable biologists to focus more on function rather than characterization. When combined with dynamic studies using, for example, fluorescent fusion proteins, or chemical fluorescent derivatives that highlight where in the cell rapid events such as phosphorylation take place, such function will be more easily attainable. Also, examples already exist where proteomics-based profiling have yielded important insights, in particular where proteomics-based profiling of organelles have been carried out [19]. We now know the composition of the nuclear pore [50, 51], much of the composition of the mitochondria, the secretory pathway, transport vesicles including synaptic vesicles, clathrin coated vesicles and COPI vesicles [52]. Through the human proteome project, such proteomes will be extended to include most of the components of the cell from different cell types yielding a formidable resource that allows biologists to explore function without spending too much of their resources and time on what will then constitute redundant large-scale efforts.

Clinical relevance

Clinical applications of proteomics involve the use of proteomics technology at the bedside with the ultimate goal to characterize the information flow through the intra and

A Human Proteome Project

extracellular molecular protein networks that interconnect organs and systems together [53-55]. The demand for new biomarkers for this field is apparent in most areas of pathology for diagnosis of diseases such as cancer, cardiovascular diseases, metabolic diseases, liver diseases, kidney diseases, brain diseases, infectious diseases, pathogenesis and pathophysiology of these diseases, which have not been clarified in spite of extensive researches by existing approaches for a long period, are now expected to be disclosed by proteomics.

The outcome from the HUPO initiative project contains MS based protein profiling, network based protein interaction as well as antibody based tissue specific protein profiles in normal and diseased human cells, tissues and organs listed above. The assured resource (the proteome) will be an indispensable tool by providing context to studied proteins, processes and linked diseases. As such, it provides a foundation to elucidate mechanistic function for each protein and diagnostics and therapeutics for human diseases.

Clinical proteomics is an interdisciplinary field and requires interaction of clinicians, statisticians, bioinformaticians and others from the beginning of the analysis.

For example, the analysis of human cancer can be used as a model for how clinical proteomics is having an impact at the bedside for early detection, rational therapeutic targeting, and patient-tailored therapy. The use of biomarker in clinical pathology for cancer diagnosis aims to determine cellular differentiation and grade of malignancy in a given cell population, i.e. over-expression or lack of expression of a given protein in a given tumor [56].

In typical cases a basic phenotype can be defined by using crude differentiation markers, e.g. cytokeratins for epithelial cells, vimentin for mesenchymal cells and the leucocyte common antigen (LCA) for cells of hematopoietic origin. The most widely used cell-type specific antibody detects the PSA protein and is an excellent marker for tumors originating from the prostate [57]. Other examples include tyrosinase related proteins for melanocytic tumors [58], thyroglobulin as a marker for thyroid carcinomas [59], and chromogranin as a marker for various endocrine tumors [60]. For morphologically undifferentiated tumors and for metastasis from unknown primary tumors, is often necessary to establish what lineage of differentiation the given tumor has. One good example has been reported as one result from HUPO initiative: A web-based tool (www.proteinatlas.org) for *in silico* biomarker discovery based on tissue specific protein profiles in normal and cancer tissues [61]. The search queries presented in this database may constitute a valuable resource to better define the proteomic landscape in tissues, support the discovery of new diagnostic and therapeutic tools and enhance opportunities for basic biological and medical research. With future discoveries in pursued project, the profiled data will be added in this database and can be used as resources for further clinical research. For another instance, chronic kidney disease (CKD) is one of the most serious diseases in the kidney, which progresses insidiously to threaten the life. As no curative or radical treatments have been developed for patients with CKD, they are finally treated with dialysis or kidney transplantation. In the consequence, the number of

A Human Proteome Project

these patients is increasing in the world and medical costs for the treatment are also elevating in many countries. To develop curative therapies for CKD, identification of disease cause and understanding of progression mechanisms at a molecular level are essential, and proteomics analysis of kidney biopsy samples is expected to identify the pathogens and to disclose molecular mechanisms of the progression. By providing catalogues of proteins in the kidney as a dataset and by comparing proteomes of normal and disease conditions, the pathogenesis and pathophysiology are expected to be clarified by proteomics.

In conclusion, the current and future protein profiles will have important translational applications for early detection, as a supplement to existing and co-evolving technological advances in diagnostic imaging, and provide a rational basis for patient tailored therapy. In the future, the clinical researchers and doctors can take advantage of this proteomic armamentarium. Diseases could be detected earlier, with greater specificity and sensitivity using mass spectrometry as the central clinical tool coupled with artificial intelligence based pattern recognition systems. Once detected, each patient's disease will be profiled through phosphoproteomic and protein network analysis combined with genomic analysis using microscopic quantities of patient tissue material and/or serum proteomic pattern analysis.

On the basis of proteomic and genomic portraits of the disease, an individualized selection of therapeutic combinations that best target the protein network will be selected and employed resulting in a paradigm shift in patient treatment and disease management.

Stem cell relevance

The human embryonic and adult stem cells are highly flexible and have the unique property to form various types of the cells in the human body by differentiation process. The control of this property *in vitro* would offer opportunities to develop treatments of diseases, especially in the area of regenerative medicine, or to design strategies for screening of drugs. Apart from these applications on the long term, molecular understanding of mechanisms controlling stem cell maintenance and differentiation would be of high value for basic understanding of the distinctive properties of stem cells.

Proteins, key players in the cell, have diverse features that are not predictable from gene sequences or from the level of transcripts. For example, posttranslational modifications (PTMs), protein-protein interactions, and subcellular locations affect the function and activity of proteins, but are not predictable using genomics or transcriptomics technology. Yet, basic as well as clinically oriented research of stem cells is confronted with many challenges and open questions. For instance, cell surface proteins and signaling cascades both for primitive stem cells as well as differentiated sub-populations are largely unknown, as are differentiation-specific proteins that can be used as biomarkers for the intermediate or terminal steps of differentiation of cells as well as discrimination of the tumorigenic cells from the pool [62]. These are all areas where proteomics can contribute significantly, and given the current state of technology in proteomics, which has matured immensely in recent years [10, 63]. Recently, the HUPO initiative 'Proteome biology of stem cells' has been established as a collaborative platform

A Human Proteome Project

bringing together stem cell biologists and researchers in proteomics [62]. The aim is effectuate the implementation of cutting edge proteomic technology in stem cell research to further our understanding of stem cell biology. This has been prompted primarily by major breakthroughs in stem cell biology and the potential of stem cells for biomedical application, and the awareness that proteomics has a place to accelerate this progress further, or to open yet unexplored areas.

Contingency plans

The draft of all proteins characterized for chromosome 21 will indicate any technical difficulties at the proteomics based profiling and antibody based profiling effort. For getting at any proteins beyond the capability of the high throughput platforms labeled peptides with MRM and paired antibodies via ELISA will be used. A microfluidics platform especially designed for this purpose will provide a contingency for these hard to detect proteins. A similar approach will be used for the remainder of the 21,000 proteins of all chromosomes. Here however the network based profiling resource will be of immense benefit sine the cognate genes of mice whose proteins have been characterized in networks will supplement the comprehensive protein profiling effort from proteomics and antibodies. Strategies that deploy so called targeted proteomics where antibodies are used to immunoprecipitate the protein before MS analysis works has been shown to be both accurate as well as quantitative thus providing additional routes to complete the proteome. It is also envisaged that much of the profiling will be transferred to array-based technologies where, for example, arrays containing pre-spotted antibodies are used to detect proteins from complex mixtures. Such arrays already exist commercially (e.g. Labvision Corporation, Thermo Fisher Scientific) and it is anticipated that the number of antibodies per array will grow to the point where the entire 21,000 proteins encoded by the human genome can be captured in a single analysis.

Conclusion

This project will define the human proteome in 10 years along with all glycan modifications and a network based profiling to deduce function. The bioinformatics interface supplemented by the computational biology of each of the engine modules used to define the proteome will create the resource which will be accessible and accessed by basic biologists and the clinical community. From this will come the fundamental discoveries in basic biology in which each assay unique to each lab or PI can focus on the relevant proteins. Similarly for clinical research, the resource of the proteome will immediately be utilized by all clinicians involved in diseases linked to any of the organs characterized. Furthermore any organs not studied here can be expedited in its proteomics characterization by the reagents and bioinformatics interface which enables the ready acquisition of any new data to select for the proteins and in common to the new organ under study. Similarly for cohorts of diseased patients this will expedite the characterization of proteins that are signatures of disease.

A Human Proteome Project

References

1. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome*. Proc Natl Acad Sci U S A, 2007. **104**(49): p. 19428-33.
2. *Proteomics' new order*. Nature, 2005. **437**: p. 169-70.
3. Deutsch, E.W., H. Lam, and R. Aebersold, *Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics*. Physiol Genomics, 2008. **33**(1): p. 18-25.
4. Bell, A.W., et al., *Collaborative MS based proteomics: equimolar test sample of 20 proteins*. 2008 submitted.
5. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae**. Nature, 2006. **440**(7084): p. 637-43.
6. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery*. Nature, 2006. **440**(7084): p. 631-6.
7. Bergeron, J.J. and R.A. Bradshaw, *What has proteomics accomplished?* Mol Cell Proteomics, 2007. **6**(10): p. 1824-6.
8. Jeronimo, C., et al., *Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme*. Mol Cell, 2007. **27**(2): p. 262-74.
9. Uhlen, M., *Mapping the human proteome using antibodies*. Mol Cell Proteomics, 2007. **6**(8): p. 1455-6.
10. Cox, J. and M. Mann, *Is proteomics the new genomics?* Cell, 2007. **130**(3): p. 395-8.
11. Brunner, E., et al., *A high-quality catalog of the *Drosophila melanogaster* proteome*. Nat Biotechnol, 2007. **25**(5): p. 576-83.
12. Cravatt, B.F., G.M. Simon, and J.R. Yates, 3rd, *The biological impact of mass-spectrometry-based proteomics*. Nature, 2007. **450**(7172): p. 991-1000.
13. Haab, B.B., et al., *A reagent resource to identify proteins and peptides of interest for the cancer community: a workshop report*. Mol Cell Proteomics, 2006. **5**(10): p. 1996-2007.
14. Taussig, M.J., et al., *ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome*. Nat Methods, 2007. **4**(1): p. 13-7.
15. Berglund, A., Odeberg, and M. Uhlen, *The Epitope Space of the Human Proteome*. Protein Sci, 2008 in press.
16. Anderson, N.L. and N.G. Anderson, *The human plasma proteome: history, character, and diagnostic prospects*. Mol Cell Proteomics, 2002. **1**(11): p. 845-67.
17. Yan, W. and S.S. Chen, *Mass spectrometry-based quantitative proteomic profiling*. Brief Funct Genomic Proteomic, 2005. **4**(1): p. 27-38.
18. Yates, J.R., 3rd, et al., *Proteomics of organelles and large cellular structures*. Nat Rev Mol Cell Biol, 2005. **6**(9): p. 702-14.
19. Au, C.E., et al., *Organellar proteomics to create the cell map*. Curr Opin Cell Biol, 2007.
20. Lu, P., et al., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation*. Nat Biotechnol, 2007. **25**(1): p. 117-24.

A Human Proteome Project

21. Olsen, J.V., et al., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks*. Cell, 2006. **127**(3): p. 635-48.
22. Gilchrist, A., et al., *Quantitative proteomics analysis of the secretory pathway*. Cell, 2006. **127**(6): p. 1265-81.
23. Wada, Y., et al., *Comparison of the methods for profiling glycoprotein glycans--HUPPO Human Disease Glycomics/Proteome Initiative multi-institutional study*. Glycobiology, 2007. **17**(4): p. 411-22.
24. Packer, N.H., et al., *Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006)*. Proteomics, 2008. **8**(1): p. 8-20.
25. Pinkse, M.W., et al., *Highly robust, automated, and sensitive online TiO₂-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster*. J Proteome Res, 2008. **7**(2): p. 687-97.
26. de Godoy, L.M., et al., *Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system*. Genome Biol, 2006. **7**(6): p. R50.
27. Zhai, B., et al., *Phosphoproteome analysis of Drosophila melanogaster embryos*. J Proteome Res, 2008. **7**(4): p. 1675-82.
28. Zhang, Z. and D.W. Chan, *Cancer proteomics: in pursuit of "true" biomarker discovery*. Cancer Epidemiol Biomarkers Prev, 2005. **14**(10): p. 2283-6.
29. Kirkpatrick, D.S., et al., *Proteomic identification of ubiquitinated proteins from human cells expressing His-tagged ubiquitin*. Proteomics, 2005. **5**(8): p. 2104-11.
30. Cojocar, M., et al., *Genomic location of the human RNA polymerase II general machinery: evidence for a role of TFIIIF and Rpb7 at both early and late stages of transcription*. Biochem J, 2008. **409**(1): p. 139-47.
31. Jeronimo, C., et al., *RPAPI, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits*. Mol Cell Biol, 2004. **24**(16): p. 7043-58.
32. Lee, Y.S. and M. Mrksich, *Protein chips: from concept to practice*. Trends Biotechnol, 2002. **20**(12 Suppl): p. S14-8.
33. Ren, B., et al., *Genome-wide location and function of DNA binding proteins*. Science, 2000. **290**(5500): p. 2306-9.
34. Mueller, M., L. Martens, and R. Apweiler, *Annotating the human proteome: beyond establishing a parts list*. Biochim Biophys Acta, 2007. **1774**(2): p. 175-91.
35. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information*. Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.
36. Mishra, G.R., et al., *Human protein reference database--2006 update*. Nucleic Acids Res, 2006. **34**(Database issue): p. D411-4.
37. Mathivanan, S., et al., *Human Proteinpedia enables sharing of human protein data*. Nat Biotechnol, 2008. **26**(2): p. 164-7.
38. Flicek, P., et al., *Ensembl 2008*. Nucleic Acids Res, 2008. **36**(Database issue): p. D707-14.
39. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update*. Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.

A Human Proteome Project

40. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data*. Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
41. Chatr-aryamontri, A., et al., *MINT: the Molecular INTERaction database*. Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
42. Desiere, F., et al., *The PeptideAtlas project*. Nucleic Acids Res, 2006. **34**(Database issue): p. D655-8.
43. Jones, P., et al., *PRIDE: new developments and new datasets*. Nucleic Acids Res, 2008. **36**(Database issue): p. D878-83.
44. Hermjakob, H. and R. Apweiler, *The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible*. Expert Rev Proteomics, 2006. **3**(1): p. 1-3.
45. Deshpande, N., et al., *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema*. Nucleic Acids Res, 2005. **33**(Database issue): p. D233-7.
46. Velankar, S., et al., *E-MSD: an integrated data resource for bioinformatics*. Nucleic Acids Res, 2005. **33**(Database issue): p. D262-5.
47. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data*. Nucleic Acids Res, 2007. **35**(Database issue): p. D301-3.
48. Dowell, R.D., et al., *The distributed annotation system*. BMC Bioinformatics, 2001. **2**: p. 7.
49. Kasprzyk, A., et al., *Ensembl: a generic system for fast and flexible access to biological data*. Genome Res, 2004. **14**(1): p. 160-9.
50. Devos, D., et al., *Simple fold composition and modular architecture of the nuclear pore complex*. Proc Natl Acad Sci U S A, 2006. **103**(7): p. 2172-7.
51. Rout, M.P., et al., *The yeast nuclear pore complex: composition, architecture, and transport mechanism*. J Cell Biol, 2000. **148**(4): p. 635-51.
52. Simpson, J.C., A. Mateos, and R. Pepperkok, *Maturation of the mammalian secretome*. Genome Biol, 2007. **8**(4): p. 211.
53. Hanash, S., *Disease proteomics*. Nature, 2003. **422**(6928): p. 226-32.
54. Hanash, S., *Integrated global profiling of cancer*. Nat Rev Cancer, 2004. **4**(8): p. 638-44.
55. Hanash, S.M., S.J. Pitteri, and V.M. Faca, *Mining the plasma proteome for cancer biomarkers*. Nature, 2008. **452**(7187): p. 571-9.
56. Petricoin, E.F., et al., *Lessons from Kitty Hawk: from feasibility to routine clinical use for the field of proteomic pattern diagnostics*. Proteomics, 2004. **4**(8): p. 2357-60.
57. Bostwick, D.G. and J. Qian, *Current and proposed biologic markers in prostate cancer: 1994*. J Cell Biochem Suppl, 1994. **19**: p. 197-201.
58. Chen, Y.T., et al., *Immunophenotyping of melanomas for tyrosinase: implications for vaccine development*. Proc Natl Acad Sci U S A, 1995. **92**(18): p. 8125-9.
59. Rosai, J., *Immunohistochemical markers of thyroid tumors: significance and diagnostic applications*. Tumori, 2003. **89**(5): p. 517-9.
60. Lloyd, R.V. and B.S. Wilson, *Specific endocrine tissue marker defined by a monoclonal antibody*. Science, 1983. **222**(4624): p. 628-30.

A Human Proteome Project

61. Denisov, I.G., et al., *Structure and chemistry of cytochrome P450*. Chem Rev, 2005. **105**(6): p. 2253-77.
62. Krijgsveld, J., et al., *Proteome biology of stem cells: a new joint HUPO and ISSCR initiative*. Mol Cell Proteomics, 2008. **7**(1): p. 204-5.
63. Maltman, D.J. and S.P. Przyborski, *Can large-scale analysis of the proteome identify effective new markers for embryonic stem cells?* Regen Med, 2007. **2**(4): p. 465-9.