The Human Proteome Project

Anno 2009/2010
Dott.ssa Manuela Santonocito
With the completion of the Human Genome Project (HGP), the emphasis is shifting to the protein complement of the human organism.
The term “proteome” (PROTEins expressed by a genOME) refers to the total set of proteins expressed in a cell, tissue, or organism.

While a genome remains unchanged to a large extent, the proteins in any particular cell change dramatically as genes are turned on or off in response to the environment.
The Monarch butterfly life cycle, with its 5 instars, has 8 stages.
The PROTEOMICS is the study of all proteins produced by cell and organism and involves the identification of proteins in the body and the determination of their roles in physiological and pathophysiological functions.
Why is proteomics necessary?

Researchers are realizing that merely having complete sequences of genomes is not sufficient to elucidate biological function because proteins are the molecules responsible for cellular functions (e.g. signal transduction).
There is no strict linear relationship between genes and the 'proteome' of a cell. Gene sequence do not provide an accurate profile of a protein abundance, structure and activity in a cell (alternative splicing, post-translational modification etc.)

Proteins may undergo more than 200 different types of post-translational modification, including:

- phosphorylation
- glycosylation
- acetylation
- deamination
- farnesylation
- myristoylation
- palmitoylation
- proteolysis

Such a wide range of modifications cannot be predicted purely from DNA sequences.
How many proteins are in the proteome?

Since the human genome has about 22,000 genes, we might guess that the human proteome has about 22,000 proteins. But this number is both too small.

Why is it too small?

The classic dogma of genetics “one gene equals one protein” is oversimplified, in fact there are more proteins than genes.

There are many more proteins in a proteome than genes in a genome.
It is difficult to predict the actual numbers of encoded proteins based on genomic data, for a number of reasons.

The concept of one-gene to one-protein is over-simplified since an RNA can be differentially spliced and can produce various protein products. Furthermore, the protein may be affected by more than 200 different types of post-translational modifications. Finally, as a result of compartmentalization and translocation, the same protein can be found with different properties and functions in different locations.
It is believed that through genomics and proteomics, new disease markers and drug targets can be identified, which will ultimately help design products to prevent, diagnose, and treat diseases.
Why is protein analysis more complicated with respect to gene analysis?

- BIOLOGICAL PROBLEMS
- TECHNICAL PROBLEMS
Whilst humans are estimated to have 22,000 genes potentially encoding several different proteins, alternative RNA splicing and post-translational modification (PTM) may increase this number up to 500,000 proteins or protein fragments.

As a consequence the proteome is far more complex than the genome.
Proteomics can be classified into three types:

1. **Expression proteomics**
   - Proteome profiling
   - Biomarker discovery
   - Proteomic data

2. **Structural proteomics**
   - Structure of protein complexes
   - Protein-protein interactions

3. **Functional proteomics**
   - Subcellular proteomics
   - The mechanisms or functions of protein compartmentalization/translocation

**Genome project**

**Potential applications**

**Benefits**

- Prove the existence of genes
- Elucidate the mechanisms of diseases, aging and protein functions
- Facilitate gene therapy, etc.

**Integration**

- Genomic data
Major proteomics directions
The aim of all types of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional (3-D) map of protein localizations and interactions in a cell or an organism.

The completion of this map will certainly require contributions from various disciplines such as biochemistry, molecular biology, biophysics and bioinformatics.
Methods of protein measurement

The 2-D polyacrylamide gel electrophoresis method is the most commonly used technology for analyzing protein variability between samples. This method enables to distinguish up to 10,000 proteins.

Mass Spectrometry (MS) is the current method of choice for the identification of proteins, since this method offers high analytical sensitivity and the capacity for high-throughput protein identification.

Protein arrays use antibodies of known affinity and specification and they are stamped on the surface of oligonucleotides. This method allows the observation of the biochemical activities of thousands of proteins.
A) Two dimensional gel electrophoresis (2-DE): Separates proteins based on two properties: their size, and their individual unique isoelectric points.

B) Difference in gel electrophoresis (DIGE): minimizes gel-to-gel variation and reduces workload as the number of gel needed to be run is reduced.
Limitations...

- it is a labor-intensive and tedious procedure.

- many large or hydrophobic proteins do not enter the gel during the first dimension. In addition, proteins with extreme pH (below 3 or above 10) are not separated, but focused as vertical lines on both sides.

- low-copy number proteins either cannot or can hardly be detected on 2-D gel. Although this problem can be overcome by increasing the sample loading, there may be a risk of overloading the system and reducing the resolution.
Proteomic techniques utilizing mass spectrometry include:
(i) protein-protein interactions;
(ii) post-translational modifications;
(iii) structural proteomics;
(iv) protein quantitation or differential modifications;
(v) protein identification.

These techniques are used by laboratories to isolate proteins and identify protein characteristics to help clinicians determine between the healthy and disease states of the patients
The main components of a mass spectrometer are:

- an ion source,
- one or several mass analyzers that measure the mass-to-charge ratio (m/z) of the ionized analytes,
- a detector that registers the number of ions at each m/z value.
Extract proteins from gel and split into fragments of 5-10 amino acids

Compare with existing sequence databases to find matches
It is well recognised that the complexity of the human proteome far exceeds that of the genome.

When variables such as alternative gene splicing events and post-translational modifications are taken into account, the number of different molecular protein species in man is likely to be at least an order of magnitude greater than the number of genes, i.e. about 500,000 proteins.

A protein-detecting microarray comprises many different affinity reagents (frequently antibodies) arrayed at high spatial density on a solid support. Each agent captures its target protein from a complex mixture (such as serum or cell lysate), and the captured proteins are subsequently detected and quantified.
a. In a **sandwich immunoassay**, capture antibodies are immobilized on the solid support, and bound proteins are detected using a second, labeled detection antibody.

b. In an **antigen capture assay**, proteins are similarly captured by immobilized antibodies, but the captured proteins are detected directly. This is usually accomplished by chemically labeling the complex mixture of proteins before applying them to the array. In the two-color version of this assay, two samples are labeled independently with distinguishable fluorophores, and the samples are mixed before applying them to the array.

c. In a **direct assay**, the complex mixture of proteins is itself immobilized on the solid support, and specific proteins in that mixture are visualized using labeled detection antibodies.
Antibody-based Protein Array

1. Incubate soluble sample on plate or membrane-immobilized antibody array

2. Add a mix of labeled soluble antibody against the same set of antigens

3. Incubate with developing system and quantitate signal
Suspension or bead-based arrays use different fluorescent beads. Each bead is coated with a different antibody, and all beads are spectrally resolvable from each other. The beads are incubated with a sample to allow protein binding to the capture antibodies, and the mixture is incubated with a cocktail of detection antibodies, each corresponding to one of the capture antibodies.

The detection antibodies are tagged to allow fluorescent detection by a flow cytometer system.
A major objective of this project would be to characterize all proteins encoded for in the human genome and determine their range of variation in health and in disease.
Aim of the Human Proteome Project

- Characterization of the entire proteome (an “atlas approach”) of a cell, tissue or organism by systematic analysis.

- Spatial and temporal characterization of protein expression in a cell/tissue by systematic analysis of individual cell fractions such as nucleus, plasmic membrane, cytoplasmic, etc. and/or individual cell population in a tissue.


- Quantitative/qualitative study of global changes in proteins expression between treated and non-treated and/or normal and disease, to look for toxic effects/responses or disease markers respectively.
The Human Proteome Organisation (HUPO) is an international scientific organization representing and promoting proteomics through international cooperation and collaborations by fostering the development of new technologies, techniques and training.

The Human Proteome Organisation (HUPO) is a nonprofit organization promoting proteomic research and analysis of human tissues. It was launched on February 9, 2001 and has a mandate of fostering international proteomic initiatives to better understand human disease.

http:www.HUPO.org
It was founded to regroup scientists in the public and private sectors engaged worldwide in various aspects of proteomics. Over the past year, HUPO has strived to develop a mission for itself and to begin the process of organizing its activities.

The mission of HUPO as it is currently defined is to:

I. Consolidate national and regional proteome organizations into a worldwide organization;

II. Engage in scientific and educational activities to encourage the spread of proteomics technologies and disseminate knowledge pertaining to the human proteome and that of model organisms;

III. Assist in the coordination of public proteome initiatives.
HUPO consists of three major geographic spheres that include Europe, Asia/ Oceania, and the Americas.
Presently, there are eleven HUPO-sponsored Scientific Initiatives:

- Human Liver Proteome Project (HLPP)
- Human Brain Proteome Project (HBPP)
- Proteomic Standards Initiative (PSI)
- Human Antibody Initiative (HAI)
- Plasma Proteome Project (PPP)
- Human Disease Glycomics/Proteome Initiative (HGPI)
- HUPO Cardiovascular Initiative (HUPO CVI)
- Proteome Biology of Stem Cells Initiative
- Disease Biomarkers Initiatives (DBI)
- Mouse Models of Human Disease (MMHD)
- Kidney and Urine Initiative (HKUPP)

Each of these global initiatives includes subprojects that involve international research laboratories. Funding for these projects has come from National Granting Agencies and industry.
The Plasma Proteome Project (PPP)

The PPP is the first proteome project to be implemented, which is currently in its piloting phase.

The **scientific objectives:**

✓ comprehensive analysis of plasma protein constituents in normal humans in large cohorts of subjects;

✓ determination of the extent of variation in plasma proteins within populations in various countries and across various populations from around the world;

✓ identification of biological sources of variation within individuals over time and assessment of the effect of age, sex, diet, and lifestyle, as well as common medications and common diseases.
Aims of the Piloting Phase of the PPP

1. Compare a broad range of technology platforms for the characterization of proteins in human plasma and serum. Assess resolution, sensitivity, time, cost and volumes of samples required.

2. Clarify the influence of various technical variables in specimen collection, handling, and storage.

3. Assess the need and feasibility of depleting the most abundant plasma proteins and the need for additives for protein stability.

4. Develop a database structure and repository for HUPO PPP results.

5. Lay the groundwork through evaluation of technology platforms and specimen handling for future studies of circulating proteins (biomarkers) in health and disease.
The pilot phase for the PPP has begun using standardized samples distributed to some 50 laboratories throughout the world.

Another objective is to begin to establish international collaborations for later-phase characterization of the normal human plasma proteome in various ethnic groups.
Although the plasma presents challenges related to the vast dynamic range of protein abundance, it is far less complex than an organ such as the liver, which includes numerous cell types and thus presents much greater challenges with respect to sampling, storage, and distribution.
The liver is the largest organ in the body, is probably second only to the brain in complexity, and has the main digestive function for the metabolism of most substances. It has a myriad of additional functions beyond digestion, such as:

- Production of red blood cells during embryonic development
- Production of various plasma proteins
- The detoxification of xenobiotics

It is also the most effective site for phagocytosis of solid material and the guardian interposed between the digestive tract and the rest of the body. The liver has a central role in activation, catabolism, and excretion of retinols, which play various critical roles in vision, growth, reproduction, cell proliferation, differentiation, and the integrity of the immune system. In addition, the liver plays a major role in determining the pharmacokinetics of a drug because it is the major organ of drug elimination through its metabolic capacity and biliary excretion; it also influences the distribution of drugs via the synthesis of their binding proteins.

Liver diseases, such as viral hepatitis, liver cancer, and alcohol-related and drug-related liver injury, are great challenges for modern medicine. Worldwide, there are over 350 million hepatitis virus carriers, and over a million deaths per year (about 10% of all deaths in the adult age range) can be attributed to viral hepatitis, and over a million deaths per year are due to hepatocellular carcinoma.
**HLPP--**

**Modern Prometheus Myth**

Prometheus  Liver

Eagles  Liver diseases

Heracles  HLPP

*Prometheus stole fire from the gods and gave it to mortals.*

*Heracles Liberates Prometheus*

*Prometheus Bound*
Scientific objectives of the HLPP

1. Collection and banking of human liver tissue specimens
2. Characterization of the protein expression profile of human liver
3. Elucidation of post-translational modifications of liver proteins
4. Construction of protein interaction maps of human liver
5. Localization of human liver proteins in cellular compartments
6. Development of an antibody bank for human liver proteins
7. Development of an ORF bank for human liver proteins
8. Studies of liver disease with a focus on hepatitis, liver cancer, and related pathologies
According to the global plan, HLPP could be divided into two phases:
1) pilot phase
2) action phase

The pilot phase for the HLPP consists of analysis of a limited number of liver tissue samples and cultured cells to determine feasibility, as in the case of the PPP.
The importance of the LPP is demonstrated by the decision of the government of China to become a major participant in the project with the allocation of some $30 million to the pilot phase of the LPP and a projected contribution of some $200 million for the execution phase.

Other countries with important contributions to the project include France, Canada, and the United States.
The Brain Proteome Project (BPP)

It is dedicated to the analysis of the brain proteome and has initiated two pilot studies in order to elaborate a standardised system for data collection and reprocessing.

It started in 2003 shortly after the 1st HUPO World Congress in Versailles and is chaired by Helmut E. Meyer, Bochum, and Joachim Klose, Berlin, both Germany.
The brain is of paramount interest in medical research and in pharmaceutical industry because of the widespread social impact of the more common neurological diseases such as Alzheimer's disease, Parkinson's disease, multiple sclerosis, prion diseases and stroke.

One aim of the HUPO BPP is the characterisation of the human and mouse brain proteomes and the usage of the data (identified proteins, mRNA profiles, protein/protein interactions, protein modifications and localisation, validated protein targets) in understanding neurodegenerative diseases and aging.

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

The prevalence of some of these diseases is increasing significantly, e.g. every 5th person over 80 years in industrial countries suffers from Alzheimer’s disease.
The goals of BPP

✓ to analyze the brain proteome of human as well as mouse models in healthy, neurodiseased and aged status with focus on Alzheimer’s and Parkinson’s Disease

✓ to perform quantitative proteomics as well as complementary gene expression profiling on disease-related brain areas and bodily fluids

✓ to advance knowledge of neurodiseases and aging in order to push new diagnostic approaches and medications

✓ to exchange knowledge and data with other HUPO projects and national=international initiatives in the neuroproteomic field

✓ to make neuroproteomic research and its results available in the scientific community and society.
In the mouse pilot study, brain tissue from normal mice of three developmental stages will be analysed by quantitative proteomics, while in the human pilot study, two human brain tissue samples from an autopsy and a biopsy, respectively, will be analysed by quantitative/qualitative proteomics techniques.
The validation of the proteins identified is the next essential step.
The PSI has three major kinds of product:

I. standard file formats with which to move data from point A to point B;

II. controlled vocabularies of well-characterized terms with which to describe data and protocols (complementary to the aforementioned formats);

III. reporting requirement documents that make explicit the minimum information to be provided in reporting experimental data and analyses thereof.
Where are proteomic data collected?

The completion of the sequencing of the human genome and the concurrent, rapid development of HT proteomic methods have resulted in an increasing need for automated approaches to archive proteomic data in a repository that enables the exchange of data among researchers and also accurate integration with genomic data.
The PeptideAtlas Project

Peptide-Atlas addresses these needs by identifying peptides by tandem mass spectrometry (MS/MS), statistically validating those identifications and then mapping identified sequences to the genomes of eukaryotic organisms.

PeptideAtlas was initially designed to annotate eukaryotic genomes with peptide sequences obtained from mass spectrometry (MS) experiments.

http://www.peptideatlas.org
From Peptides to Genome Annotation

- Sample
- Proteins
- Peptides
- Mass Spectrum
- Peptides
- database search
- statistical filtering
- SBEAMS
- BLAST protein database
- Map to genome
- visualization

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Peptide</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum 1</td>
<td>LGYGH</td>
<td>1.0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Spectrum N</td>
<td>EIYHKF</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Chrom</th>
<th>Start_Coord</th>
<th>End_Coord</th>
</tr>
</thead>
<tbody>
<tr>
<td>P6000007336</td>
<td>X</td>
<td>12217318</td>
<td>132217388</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
PRIDE, the 'PRoteomics IDEntifications database' is a database of protein and peptide identifications that have been described in the scientific literature.

http://www.ebi.ac.uk/pride

The PRIDE database has been developed to provide a standards-compliant repository for mass-spectrometry based proteomics data comprising identifications of proteins, peptides and post-translational modifications, together with the mass spectra that provide evidence for these identifications.
## Identifications

<table>
<thead>
<tr>
<th>Accession</th>
<th>Splice Isoform</th>
<th>Database</th>
<th>Score</th>
<th>Threshold</th>
<th>Search Engine</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPI00295313</td>
<td>IPI human 2.31</td>
<td>52.0</td>
<td>36.0</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00017340</td>
<td>IPI human 2.31</td>
<td>56.0</td>
<td>35.0</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00028128</td>
<td>IPI human 2.31</td>
<td>80.6667</td>
<td>27.3333</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00031189</td>
<td>IPI human 2.31</td>
<td>48.0</td>
<td>33.5</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00291262</td>
<td>IPI human 2.31</td>
<td>60.6667</td>
<td>41.0</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00395553</td>
<td>IPI human 2.31</td>
<td>48.0</td>
<td>36.0</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPI00328748</td>
<td>IPI human 2.31</td>
<td>43.0</td>
<td>37.0</td>
<td>Mascot 2.0.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## mzData

### Information:
- **Version:** 1.05
- **Accession:** 1

### Sample:
- **Name:** unstimulated human platelets
- **Description:**
  - **Source** | **Name** | **Value**
  - NEWT | Homo Sapiens |
  - MeSH | blood platelets |

### Source File:
- **Name:** Krs Gevaert
- **Institution:** Ghent University, Dept. of Medical Protein Research
- **Contact Information:** kris.gevaert@UGent.be

### Instrument:
- **Analyzer #1**
  - **Source** | **Name** | **Value**
  - PSI | Ionization Type | ESI
  - User | comment | Fragmentation time per ms/ms spectrum was 8 sec

### Detector:
- **Source** | **Name** | **Value**
- PSI | Analyzer Type | Q-TOF
- User | No Detector Component...

### Additional Information:
- **Source** | **Name** | **Value**
- PSI | Vendor | Micromass UK Limited, Cheshire, UK
- PSI | Model | Q-TOF I
Why study the proteome in disease?

Despite tremendous advances in our understanding of the molecular basis of diseases such as cancer, substantial gaps remain both in our understanding of disease pathogenesis and in the development of effective strategies for early diagnosis and for treatment.

The current interest in proteomics is due in part to the prospects that a proteomic approach to disease investigations will overcome some of the limitations of other approaches.
Why study the proteome in disease?

Numerous alterations may occur in proteins that are not predictable from genomic analysis, thus providing a compelling rationale for direct analysis of gene expression at the protein level.

It is clear that a better understanding of these alterations will have a substantial impact in medicine.
As a result, there is intense interest in applying proteomics to foster a better understanding of disease processes, develop new biomarkers for diagnosis and early detection of disease, and accelerate drug development.
Identification of serum biomarkers in HBV-infected patients

• HBV, a serious infectious and widespread human pathogen, represents a major worldwide health problem.

• Chronic HBV infection has a very high chance of evolving into hepatocellular carcinoma.

• The pathogenesis of HBV infection is still elusive and a definite diagnosis still relies on biopsy histological test.

To use proteomic technology to globally examine HBV-infected serum samples in a search for disease-associated proteins that can be used as serological biomarkers for diagnosis and/or target proteins for pathogenetic study.
Identification of serum biomarkers in HBV-infected patients

After a comparison with normal serum samples, we found that at least seven proteins were significantly changed in HBV-infected sera.

- Haptoglobin β and α2 chain
- Apolipoprotein A-I and A-IV
- α1-antitrypsin
- Transthyretin
- DNA topoisomerase IIβ

The alteration of these proteins presents not only in their quantities but also in their patterns, some of which can be correlated with the necroinflammatory scores (NIS).

In particular, apolipoprotein A-I displays heterogeneous change in expression level with different isoforms and this phenomenon appear to be specific to HBV infection.
An Altered Pattern of Liver Apolipoprotein A-I Isoforms Is Implicated in Male Chronic Hepatitis B Progression

Fu Yang,† Yixuan Yin,† Fang Wang,*, Ling Zhang, Yuqi Wang, and Shuhan Sun*

Department of Medical Genetics, Second Military Medical University, Shanghai, China

Received July 06, 2009

The pattern of liver Apo A-I isoforms was altered in male CHB patient sera but not in female.

Development of antibodies that specifically recognize the isoforms of Apo A-I may prove to be useful, in combination with other traditional markers, as a more efficient way to evaluate the prognosis of CHB.
Because most drug targets are proteins, it is inescapable that proteomics will enable drug discovery, development and clinical practice.

Proteomics research permits the discovery of new protein markers for diagnostic purposes and the study of novel molecular targets for drug discovery.

Applying their findings will improve our understanding of the roles of individual proteins or the entire cellular pathways in the initiation and development of disease.
The protein markers identified have a broad range of potential applications:

- they may be used for clinical diagnostic or prognostic purposes

- to help devising an optimal therapeutic treatment plan for different patient subsets and to monitor the effect of treatment.

- In this way, protein markers may be used to accelerate the speed and efficacy of clinical trials.

- If further biochemical research reveals that proteins have a causal role in disease pathology, they may have the utility as molecular targets for therapeutic intervention in the disease.
Proteomics and Personalized Medicine

Proteomics will fuel this drive toward niche drugs and targeted therapies.

Different people have different responses to treatment:

✓ Some are cured
✓ Some don't respond
✓ Some experience a bad side effect

The goal of personalized medicine is to test patients to determine which medicine will work for them without unpleasant side effects.
The abundance of information provided by proteomics research is entirely complementary with the genetic information being generated by genomics research.

The combination of genomics and proteomics will play a major role in biomedical research, and it will have a significant impact on the development of diagnostic and therapeutic products in the future.