Website address:
https://www.gesundheitsindustrie-bw.de/en/article/dossier/the-human-proteome-the-next-major-goal

The human proteome – the next major goal

The "Human Proteome Project", a ten-year global initiative that is making a systematic effort to map all human proteins, has moved from the planning to the experimental stage. How significant and how effective the project will be depends on how much the resources offered are used by proteome researchers and on the data that the researchers bring into the project.

The successful sequencing of the human genome under the Human Genome Project more than ten years ago was expected to facilitate the identification of disease-relevant genes and consequently find ways to create therapies for previously incurable diseases. However, it soon became clear that the expectations could not be fulfilled. Like all physiological and development stages of an organism, diseases are no more than phenotypic abnormalities that result from the expression of an organism's genes, influenced by environmental factors and the interactions between the two.

The Human Proteome Project aims to identify and characterise all human proteins. The photo shows Andreas Vesalius' drawing "De humani corporis fabrica" from 1543.

Proteins are the most important cell elements involved in phenotypic expression and interaction. As the word protein (derived from 'Proteus', the Greek god of elusive sea change) implies, proteins can undergo constant structural rearrangements. While there is broad consensus that humans have between 20,000 and 21,000 protein-coding genes, it is estimated that we have as many as one million protein isoforms. The isoforms are formed by DNA recombination, alternative splicing of primary gene transcripts and a broad range of post-translational modifications, including glycosylation and phosphorylation. Protein isoforms are also different depending on an organism's developmental state and the site of protein expression as well as on physiological, pathological and pharmacological conditions.

The Human Proteome Project (HPP), whose draft plan was published in 2008, was officially launched in September 2011 with the ambitious goal of mapping the human proteome within 10 years, bringing order to the bewildering variety of human proteins and using the map as the basis for future proteomic research.

Technological pillars and research strategies

The initiators of the Human Proteome Project, a who's who of international proteomic research, believe that the technological advances of recent years have made it possible to map the human proteome. As the Human Genome Project previously showed, such large-scale projects produce a dynamic that leads to completely unpredictable innovations that may advance and possibly also accelerate the project. The HPP is expected to have a similar effect.

The experimental strategy of the HPP is based on three technological pillars: (1) quantitative mass spectrometry combined with liquid chromatography and other separation technologies (e.g. two-dimensional gel electrophoresis; (2) the capturing and identification of proteins using antibodies ("protein capture"); (3) bioinformatics tools and databases ("knowledge bases").

To investigate the human proteome, the HPP is pursuing two different independent strategies in parallel with one another: I. the Chromosome-centric Human Proteome Project (C-HPP) and II. the Biology/Disease-driven Human Proteome Project (B/D-HPP) that focusses on the biology of the 230 or so human cell types and major human diseases. Initial reports on the current stage of C-HPP have been published and many new reports are expected to be presented at the 12th Annual Congress of the Human Proteome Organisation which will be held in Yokohama (Japan) in September 2013. These will give an idea of the explosive amount of information the project is likely to produce.

The "Chromosome-centric Human Proteome Project" C-HPP

The number of protein-coding genes can be determined relatively easily from the known DNA sequence of the human genome. According to the latest information published in December 2012, humans have 20,059 protein-coding genes. The calculations

were done by neXtProt, a specialised software resource for human proteins developed at the Swiss Institute for Bioinformatics. However, the majority of publications refer to between 20,300 and 21,000 protein-coding genes.

Human karyogramme containing 22 chromosome pairs as well as an X and a Y chromosome.

© Human Genetics, Heidelberg University Hospital

The C-HPP aims to map and annotate the entire human protein set encoded in each chromosome. An international consortium consisting of 25 teams will be covering the 24 human chromosomes (the 22 diploid autosomes and the two sex chromosomes X and Y) and the mitochrondrial genome. The teams are being coordinated by research groups or organisations from different countries or groups of countries. Scientists interested in participating in C-HPP can select a chromosome based on specific disease interests (e.g. cancer, genetic disease) or targets (e.g. biomarkers).

Participation in the project does not require any change in a typical proteomic experiment, but does require data sets to be shared with all of the chromosome teams. Jointly agreed standards and reference proteins guarantee the comparability of data. This will result in an expansion of aggregated knowledge and more information on missing or poorly characterised proteins.

Open questions

Two-dimensional polyacrylamide gel electrophoresis used for the analysis of the human proteome.

© HUPO

Mass spectrometry and protein capture data obtained by the C-HPP team suggest that around two thirds of the expected proteins have been identified with a high degree of confidence; however, comprehensive information about function is still lacking. Little reliable information is yet available about the gene products of around 30 to 35 percent of all protein-coding genes encoded by the chromosomes. The researchers assume that finding the remaining 30 to 35% of proteins might be possible by reducing the detection limit, thus allowing better capture of published data sets. On the other hand, the reason the proteins are missing might be primarily biological:

- Many proteins are encoded by more than just one gene; an example of such a multi-gene protein is the receptor 2W1 in the olfactory epithelium, which is encoded by eight genes. Proteomics will never be able to disambiguate these gene level identifications.
- On the other hand, the high degree of polymorphisms might lead to multiple protein entries for certain gene products; a classical example of this is proteins of the major histocompatibility complex (MHC).
- Entire protein families might be missed because such protein samples are difficult to prepare and difficult to differentiate due to their extreme homology; they include for example membrane-embedded proteins, cytokeratins or olfactory receptors.
- Some proteins have not yet been discovered because they have short half-lives or act at very low abundance; these include regulatory proteins in the cell nucleus or low-abundance proteins in tissues and cells that have not yet been studied in detail.

These aspects are probably only an incomplete list of the challenges scientists will face as C-HPP progresses. It can be safely assumed that many problems can be solved while other, currently unforeseeable, issues arise.

The “Biology/Disease-driven HPP”

The second branch of the global proteome initiative, the so-called “Biology/Disease-driven HPP”, is a complementary project being carried out in parallel with C-HPP. At present, it consists of 16 subprojects that focus on specific biological processes or systems and disease areas. C-HPP teams have already been established for diabetes, cancer proteomics, mitochondria, infectious diseases and epigenetics/chromatin-associated proteins.

Other areas, including the proteomics of blood plasma, the liver, the brain, the kidneys and urinary tracts, cardiovascular diseases, stem cells as well as model organisms, are still in the planning stage. For all subprojects, experts prepare lists of proteins that are, or might be, of particular relevance. The investigation of signalling pathways, to name but one research example, would require scientists to collect information on all human protein kinases (the so-called kinome), a field of huge importance for the pharmaceutical industry. The B/D-HPP themes also offer space for the proteomic analysis of intracellular compartments and systems such as the nuclear pore complex and the endo-exocytose transport pathway as well as including the analysis of the proteomes of infection-causing parasites.

Different pilot projects will commence in 2013. Both the C-HPP and the B/D-HPP depend on cooperation between leading proteome laboratories around the world. The success of B/D-HPP requires the unrestricted availability of mass spectrometry and affinity chromatography data and samples to the scientific community.

According to the Swiss proteome researcher Ruedi Aebersold, a leading figure in the Human Proteome Project, the B/D-HPP aims to be much more than just a protein expression repository. The project delivers information about biological networks, connections with other cellular component classes and systems such as the genome, epigenome, transcriptome and metabolome, and provides reagents and information for scientists who are using proteomic analyses for their biological and clinical research.

"From this project, the research community can expect a full catalogue of proteins including novel drug targets, new diagnostic biomarkers and a parts list of the isoforms of cellular regulators such as major signalling pathways." (see J. Proteome Res., Vol. 12, Issue 1: Chromosome-centric Human Proteome Project – Editorial, publication date (Web): 20th December 2012). The structure of the HPP will result in continued evolution in the field with improved methodologies, sample collections and databases, international coordination in data management, global data sharing and improved repositories.

Dossier

13-May-2013