Proteome

During the past 10 years, several new techniques such as cDNA microarray, yeast two-hybrid analysis, and mass spectrometry (MS) have been introduced that allow simultaneous high-throughput analysis of multiple mRNAs and proteins within the same sample. These technologies have received a great deal of attention and gradually begun to infiltrate biochemistry and cell biology laboratories. The term proteome was introduced for the first time in 1994 at the first Proteome meeting in Siena, Italy, and was used to describe the protein complement of a genome. Proteomics can be defined as "a large-scale study of protein properties, e.g., expression level, posttranscriptional modification and protein interaction, in order to obtain a global view of disease processes or cellular processes at the protein level."

Three strategies have had a strong impact in the field of biology:

(1) the generation of protein–protein linkage maps;

(2) the annotation of genomic DNA sequences by generation of MS/MS peptide sequences; and (3) the measurement of protein expression by quantitative methods. The data output from a typical proteomics experiment is huge and therefore computer-based data storage and analysis is required. Essentially, proteomics is based on protein separation, identification, and data analysis followed by biological readouts . Proteomics have a great potential to give rise to novel discoveries and to generate new testable hypotheses by choosing the appropriate study design. Therefore, different types of separation techniques as well as MS will be discussed in more detail below along with some recent proteome findings relevant for the lung.

Significance of the Proteome

What can we learn from the proteome? Since most cellular enzymatic functions, regulatory switches, signal transducers, and structural components are composed of proteins, characterizing the proteins expressed by a cell can give important clues to the function, organization, and responsiveness inherent in a cell. In addition, by defining the variation between different cells, and between cells exposed to different stimuli, we can gain an understanding of:

- cellular adaptation to environmental signals;
- mechanisms of cellular differentiation and organismal development;
- cellular aspects of disease processes;
- cellular responses to aging;
- difference between individuals within a species, i.e., the molecular basis of our individuality in physiology, disease susceptibility, and response to therapeutics and environmental exposures.

There is currently a great deal of excitement about the potential to measure gene expression levels for every gene of an organism. Extensive or complete genome sequences have made it possible to profile the levels of mRNA transcripts of all genes simultaneously by DNA microarray hybridization. Therefore, is it even necessary to study protein expression now that

gene expression is so easily measured at the mRNA level? Most scientists believe the answer is yes, because the two approaches really are quantitatively and qualitatively different. First, most DNA microarrays typically do not differentiate between variant transcripts (produced by alternative splicing, use of alternative transcription start sites or polyadenylation sites, or RNA editing). Second, protein abundance may not be accurately predicted by mRNA level since the rate of translation and protein degradation is unknown for each mRNA. Third, posttranslational modifications and proteolytic cleavages are critical for the function of a protein, but cannot be detected or predicted by mRNA level. Finally, proteins usually work in complexes and protein localization is regulated by the cell, yet neither of these properties is addressed by examining mRNA levels.

Both the significance and the complexity of studying the proteome are evident in its sheer magnitude. The proteome is many-fold larger than the genome, given the wide degree of posttranslational modifications and processing that nearly all proteins undergo. Many examples exist where a single gene (composed of many exons) can generate hundreds and possibly thousands of different protein molecules by alternative splicing and posttranslational modifications. Thus, analysis of the entire proteome presents a more daunting challenge than the genome sequencing projects.

The **proteome** is the entire set of proteins that is, or can be, expressed by a genome, cell, tissue, or organism at a certain time. It is the set of expressed proteins in a given type of cell or organism, at a given time, under defined conditions. Proteomics is the study of the proteome.

Types of proteomes

While proteome generally refers to the proteome of an organism, multicellular organisms may have very different proteomes in different cells, hence it is important to distinguish proteomes in cells and organisms.

A **cellular proteome** is the collection of proteins found in a particular cell type under a particular set of environmental conditions such as exposure to hormone stimulation.

It can also be useful to consider an organism's **complete proteome**, which can be conceptualized as the complete set of proteins from all of the various cellular proteomes. This is very roughly the protein equivalent of the genome.

The term *proteome* has also been used to refer to the collection of proteins in certain **sub-cellular systems**, such as organelles. For instance, the mitochondrial proteome may consist of more than 3000 distinct proteins.

The proteins in a **virus** can be called a *viral proteome*. Usually viral proteomes are predicted from the viral genome but some attempts have been made to determine all the proteins expressed from a virus genome, i.e. the viral proteome.[5] More often, however, virus proteomics analyzes the changes of host proteins upon virus infection, so that in effect *two* proteomes (of virus and its host) are studied.

Importance in cancer

The proteome can be used to determine the presence of different types of cancers.

The proteome can be used in order to comparatively analyze different cancer cell lines. Proteomic studies have been used in order to identify the likelihood of metastasis in bladder cancer cell lines KK47 and YTS1 and were found to have 36 unregulated and 74 down regulated proteins. The differences in protein expression can help identify novel cancer signaling mechanisms.

Biomarkers of cancer have been found by mass spectrometry based proteomic analyses. The use of proteomics or the study of the proteome is a step forward in personalized medicine to tailor drug cocktails to the patient's specific proteomic and genomic profile. The analysis of ovarian cancer cell lines showed that putative biomarkers for ovarian cancer include "α-enolase (ENOA), elongation factor Tu, mitochondrial (EFTU), glyceraldehyde-3-phosphate dehydrogenase (G3P), stress-70 protein, mitochondrial (GRP75), apolipoprotein A-1 (APOA1), peroxiredoxin (PRDX2) and annexin A (ANXA)".

Comparative proteomic analyses of 11 cell lines demonstrated the similarity between the metabolic processes of each cell line; 11,731 proteins were completely identified from this study. Housekeeping proteins tend to show greater variability between cell lines.

Resistance to certain cancer drugs is still not well understood. Proteomic analysis has been used in order to identify proteins that may have anti-cancer drug properties, specifically for the colon cancer drug irinotecan. Studies of adenocarcinoma cell line LoVo demonstrated that 8 proteins were unregulated and 7 proteins were down-regulated. Proteins that showed a differential expression were involved in processes such as transcription, apoptosis and cell proliferation/differentiation among others.

The proteome in bacterial systems

Proteomic analyses have been performed in different kinds of bacteria to assess their metabolic reactions to different conditions. For example, in bacteria such as *Clostridium* and *Bacillus*, proteomic analyses were used in order to investigate how different proteins help each of these bacteria spores germinate after a prolonged period of dormancy. (Chen *et,al*) In order to better understand how to properly eliminate spores, proteomic analysis must be performed.

History

Marc Wilkins coined the term *proteome* in 1994 in a symposium on "2D Electrophoresis: from protein maps to genomes" held in Siena in Italy. It appeared in print in 1995, with the publication of part of his PhD thesis. Wilkins used the term to describe the entire complement of proteins expressed by a genome, cell, tissue or organism.

Size and contents

The genomes of **viruses** and **prokaryotes** encode a relatively well-defined proteome as each protein can be predicted with high confidence, based on its open reading frame (in viruses ranging from ~3 to ~1000, in bacteria ranging from about 500 proteins to about 10,000). However, most protein prediction algorithms use certain cut-offs, such as 50 or 100 amino acids, so small proteins are often missed by such predictions. In **eukaryotes** this becomes much more complicated as more than one protein can be produced from most genes due to alternative splicing (e.g. human proteome encodes about 20,000 proteins, but some estimates predicted 92,179 proteins out of which 71,173 are splicing variants.

Proteoforms. There are different factors that can add variability to proteins. SAPs (single amino acid polymorphisms) and non-synonymous single nucleotide polymorphisms (nsSNPs) can lead to different "proteoforms"or "proteomorphs". Recent estimates have found ~135,000 validated nonsynonymous cSNPs currently housed within SwissProt. In dbSNP, there are 4.7 million candidate cSNPs, yet only ~670,000 cSNPs have been validated in the 1,000-genomes set as nonsynonymous cSNPs that change the identity of an amino acid in a protein.

Dark proteome. The term dark proteome coined by Perdigão and colleagues, defines regions of proteins that have no detectable sequence homology to other proteins of known three-dimensional structure and therefore cannot be modeled by homology. For 546,000 Swiss-Prot proteins, 44–54% of the proteome in eukaryotes and viruses was found to be "dark", compared with only ~14% in archaea and bacteria.

Human proteome. Currently, several projects aim to map the human proteome, including the Human Proteome Map, ProteomicsDB, isoform.io, and The Human Proteome Project (HPP). Much like the human genome project, these projects seek to find and collect evidence for all predicted protein coding genes in the human genome. The Human Proteome Map currently (October 2020) claims 17,294 proteins and ProteomicsDB 15,479, using different criteria. On October 16, 2020, the HPP published a high-stringency blueprint covering more than 90% of the predicted protein coding genes. Proteins are identified from a wide range of fetal and adult tissues and cell types, including hematopoietic cells.

Methods to study the proteome

Analyzing proteins proves to be more difficult than analyzing nucleic acid sequences. While there are only 4 nucleotides that make up DNA, there are at least 20 different amino acids that can make up a protein. Additionally, there is currently no known high throughput technology to make copies of a single protein. Numerous methods are available to study proteins, sets of proteins, or the whole proteome. In fact, proteins are often studied indirectly, e.g. using computational methods and analyses of genomes. Only a few examples are given below.

Separation techniques and electrophoresis

Proteomics, the study of the proteome, has largely been practiced through the separation of proteins by two dimensional gel electrophoresis. In the first dimension, the proteins are separated by isoelectric focusing, which resolves proteins on the basis of charge. In the second dimension, proteins are separated by molecular weight using SDS-PAGE. The gel is stained with Coomassie brilliant blue or silver to visualize the proteins. Spots on the gel are proteins that have migrated to specific locations.

Mass spectrometry

Mass spectrometry is one of the key methods to study the proteome.[21] Some important mass spectrometry methods include Orbitrap Mass Spectrometry, MALDI (Matrix Assisted Laser Desorption/Ionization), and ESI (Electrospray Ionization). Peptide mass fingerprinting identifies a protein by cleaving it into short peptides and then deduces the protein's identity by matching the observed peptide masses against a sequence database. Tandem mass spectrometry, on the other hand, can get sequence information from individual peptides by

isolating them, colliding them with a non-reactive gas, and then cataloguing the fragment ions produced.

In May 2014, a draft map of the human proteome was published in *Nature*. This map was generated using high-resolution Fourier-transform mass spectrometry. This study profiled 30 histologically normal human samples resulting in the identification of proteins coded by 17,294 genes. This accounts for around 84% of the total annotated protein-coding genes.

Chromatography

Liquid chromatography is an important tool in the study of the proteome. It allows for very sensitive separation of different kinds of proteins based on their affinity for a matrix. Some newer methods for the separation and identification of proteins include the use of monolithic capillary columns, high temperature chromatography and capillary electrochromatography.[24]

Blotting

Western blotting can be used in order to quantify the abundance of certain proteins. By using antibodies specific to the protein of interest, it is possible to probe for the presence of specific proteins from a mixture of proteins.

Protein complementation assays and interaction screens

Protein-fragment complementation assays are often used to detect protein–protein interactions. The yeast two-hybrid assay is the most popular of them but there are numerous variations, both used *in vitro* and *in vivo*. Pull-down assays are a method to determine what kinds of proteins a protein interacts with.

Protein structure prediction

Protein structure prediction can be used to provide three-dimensional protein structure predictions of whole proteomes. In 2022, a large-scale collaboration between EMBL-EBI and DeepMind provided predicted structures for over 200 million proteins from across the tree of life.[26] Smaller projects have also used protein structure prediction to help map the proteome of individual organisms, for example isoform.io provides coverage of multiple protein isoforms for over 20,000 genes in the human genome. (Sommer)

Protein databases

The Human Protein Atlas contains information about the human proteins in cells, tissues, and organs. All the data in the knowledge resource is open access to allow scientists both in academia and industry to freely access the data for exploration of the human proteome. The organization ELIXIR has selected the protein atlas as a core resource due to its fundamental importance for a wider life science community.

The Plasma Proteome database contains information on 10,500 blood plasma proteins. Because the range in protein contents in plasma is very large, it is difficult to detect proteins that tend to be scarce when compared to abundant proteins. There is an analytical limit that may possibly be a barrier for the detections of proteins with ultra low concentrations(Ponomarenko *et,al)*

Databases such as neXtprot and UniProt are central resources for human proteomic data.

Literature

Chen, Yan; Barat, Bidisha; Ray, W. Keith; Helm, Richard F.; Melville, Stephen B.; Popham, David L. (2019-03-15). "Membrane Proteomes and Ion Transporters in Bacillus anthracis and Bacillus subtilis Dormant and Germinating Spores". *Journal of Bacteriology*. **201** (6). doi:10.1128/JB.00662-18. ISSN 0021-9193. PMC 6398275. PMID 30602489

*Ponomarenko, Elena A.; Poverennaya, Ekaterina V.; Ilgisonis, Ekaterina V.; Pyatnitskiy, Mikhail A.; Kopylov, Arthur T.; Zgoda, Victor G.; Lisitsa, Andrey V.; Archakov, Alexander I. (2016). "The Size of the Human Proteome: The Width and Depth". International Journal of Analytical Chemistry. **2016**: 7436849. doi:10.1155/2016/7436849. ISSN 1687-8760. PMC 4889822*