1

INTRODUCTION TO PROTEOMICS AND STRATEGY FOR PROTEIN IDENTIFICATION

JERZY SILBERRING

WHY PROTEOME?

Proteomics differs, in many aspects, from other traditional methods for the isolation and identification of proteins. Using classical methods, one protein is usually isolated and an effort is made to determine its sequence, structure, and function. In contrast, proteomics deals with the simultaneous (global) analysis of all proteins, using several analytical techniques. An example of a typical working scheme is shown in Figure 1.

These two approaches also differ in respect to the instrumentation used and to the scale of analysis. It should, however, be emphasized that "classical" methods for protein characterization are still in common use, and the information on isolated single molecules obtained is much more accurate and specific. Identification procedures used in proteomics are not free of artifacts and is still concerned with the problem of inaccurate bioinformatics tools and incomplete protein databases. In many cases, even less accurate but rapid identification of possible differences in protein patterns may be the basis for further, more detailed experiments. An outstanding success of proteomics is the recent discovery that HIV-infected humans who are resistant to AIDS development have a significantly higher level of defensins, cyclic peptides produced by the immune system.

The flow of information within a cell depends on many factors, such as the protein biosynthesis pathways, whose general scheme is presented in Figure 2. Analysis of the

Proteomics: Introduction to Methods and Applications, Edited by Agnieszka Kraj and Jerzy Silberring

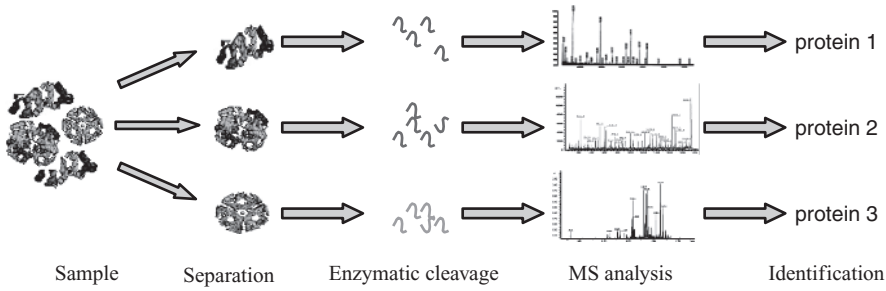

FIGURE 1 General strategy for protein identification. (*See insert for color representation of figure*.)

genome appeared to be insufficient for identification of possible causes for the occurrence of various changes in cell content. This can be visualized as shown in Figure 3.

There is no straight-line relationship between gene expression and the most important parameters characterizing proteins. For example, gene expression in the central nervous system is relatively high but the level of particular proteins is very low. Moreover, only 10% of all genes code for specific proteins. Information in the genes does not allow us to determine posttranslational modifications. A similar problem is related to protein function. Therefore, there is an urgent need for complementary investigations using the proteomic approach. This strategy makes possible additional verification of the data obtained and leads to better control of pharmacological intervention in pathophysiology.

Moreover, information provided by genes does not include the presence of low-molecular-mass components. Most often, such substances, including nitric oxide,

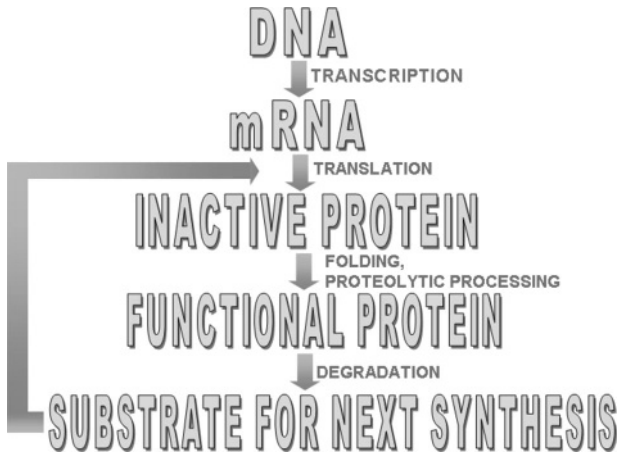

FIGURE 2 General scheme for protein biosynthesis.

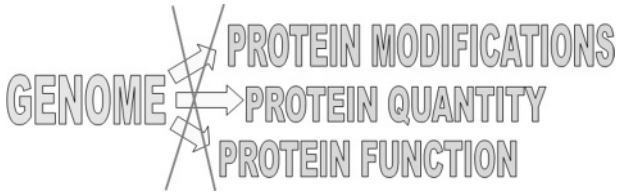

FIGURE 3 Relationships between gene expression and proteomics.

biogenic amines, prostaglandins, steroid hormones, and carbon oxide, fulfill several important functions in our organisms.

CURRENT CHALLENGES

In the postgenomic era in which almost the entire DNA sequence has been determined, the next challenge is to specify the function of each gene and to determine the role of each protein encoded by the genes. In this area, termed *functional proteomics*, identification of the primary structure (amino acid sequence) is not sufficient but is absolutely necessary in assigning the role(s) of particular proteins. It often happens that the same protein may have several functions in the same organism. For example, hemoglobin is an oxygen transporter but also transfers CO_2 to the lungs. Moreover, this protein releases shorter fragments (peptides) during proteolytic cleavage. These peptides have opioidlike properties (e.g., hemorphins) or antibacterial activity (e.g., hemocidins). It is worth noting that transgenic mice, which lack myoglobin, the protein responsible for oxygen storage in muscles, do not show any symptoms and appear to feel quite happy. It is anticipated that the function of myoglobin has been replaced by another compensatory mechanism(s). Therefore, the commonly accepted dogma of one gene–one protein–one function is no longer true. Another example is that of the proopiomelanocortin (POMC) precursor, which releases adrenocorticotrophic hormone, melanotrophic hormone, and endorphins. Each of these substances shows a distinct function in the organism, but all of them are released from the same POMC precursor.

Based on information from gene sequences, we can roughly estimate that the human body contains some 60,000 to 70,000 proteins. Taking into consideration post-translational modifications (over 100 identified!) and alternative splicing, we may end up with an approximate total of 700,000 proteins. One cell contains an average of approximately 10,000 proteins, with their quantity and proportions subject to changes in space and time. Isolation and identification of a particular protein, with its search for all modifications, its three-dimensional structure, and the function of such a huge number of molecules is a real challenge for today's scientists. It is worth noting that concentrations of endogenous compounds are often extremely low, as shown in Table 1.

We reiterate that the protein content in cells is not equal and that it changes in time and space (i.e., undergoes dynamic changes). This fact causes additional methodological problems as well as problems concerning interpretation of the results.

TABLE 1 Approximate Amounts of Substances in Living Organisms

Compound	Concentration Range
Hormones	nanomolar (10^{-9} M)
Neurotransmitters	picomolar (10^{-12} M)
Neuropeptides	femtomolar (10^{-15} M)

THE BASICS OF PROTEOMICS

As mentioned earlier, the number of proteins present in living organisms, the diversity of cells and tissues, and the dynamic changes in protein levels not only in time but also during pathological processes further complicate the methodology. Moreover, various strategies for solving these problems are not unified; that is, each laboratory may obtain a distinct set of data from the same biological material. It is also worth noting that the number of methodologies used in proteomics lends itself to the production of false-positive results. This is due to the fact that the vast amount of raw data, the output of the analytical instruments, *always* displays a "positive" result. To limit artifacts, it is advisable to focus the range of investigations and models used. For example, one can analyze proteome in prostate cancer, proteome in Down's syndrome, or peptidome in neuropeptide research. The separate aspect is *metabolomics*, a proteomics variant that deals with low-molecular-mass compounds and their metabolites. Similarly, in drug-dependence studies we can specify, for example, a morphinome or a "cocainome" because morphine and cocaine have distinct mechanisms of action on the central nervous system, inducing various changes in an organism.

Proteomics is usually associated with mass spectrometry (MS). Indeed, MS techniques are one of the most important elements of the entire strategy aimed toward the identification of proteins. But in this case, identification of a particular protein is based only on a determination of its incomplete amino acid sequence (partial sequence or peptide map). The more detailed analysis must include a full amino acid sequence, possible mutations, posttranslational modifications, tertiary and quaternary structures, interactions with other molecules, and the protein's role in the organism. It can easily be seen that from such a point of view, proteomics is not and cannot be associated with mass spectrometry only but must be linked to a vast number of techniques, including molecular biology and genetics (e.g., transgenic animals, antisense probes), crystallography, pharmacology, material engineering, and others. Additionally, analytical problems are even more complex, due to the fact that we are not able to judge which concentration of a given protein is still "physiological" or how large the differences between the physiology and pathological states should be before considering such a molecule to be of value as a marker. Sometimes, a 20% elevated protein level suggests abnormal processes, and sometimes, protein concentration must exceed two to three times its basal value before indicating pathology.

STRATEGY FOR PROTEOMIC ANALYSIS

Despite the rapid development of modern analytical techniques, simultaneous and accurate identification of the huge number of known proteins is still very difficult, if not impossible. Therefore, as mentioned earlier, laboratories try to focus such analysis on particular problems, such as cancer, drug dependence, or neurodegenerative disease. In all aspects, such investigations demand ultrasensitive and precise analytical methodologies and a solid knowledge, often at the interdisciplinary level. Because of the limited amount of precious biological material and a need for its concentration, the entire analytical process is commonly performed in a single drop (i.e., less than 5 μL).

A general strategy for protein analysis that is utilized in proteomics is shown in Figure 4. The work flow consists of several phases, each of which may be critical to the overall success of the analysis. The major challenge during the identification of endogenous compounds is the work at the very edge of the sensitivity limit, which, again, is associated with the availability of biological material. One of the crucial points is nonspecific adsorption of samples on tube walls, columns, pipette tips, and so on. All important aspects of the methodologies used in proteome research are covered in the following chapters.

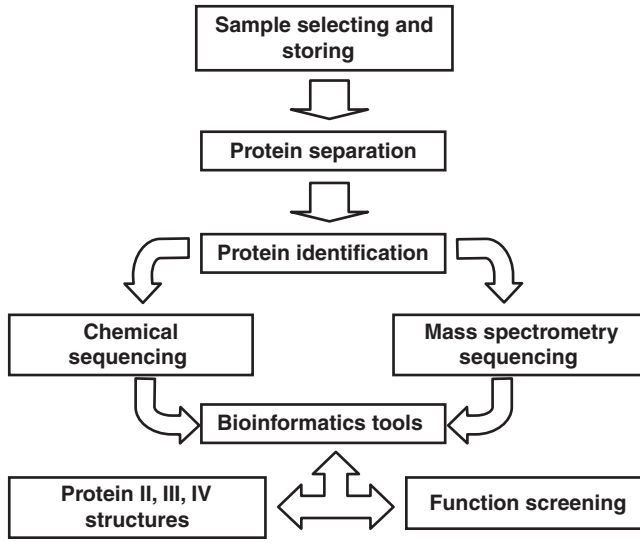

FIGURE 4 Schematic of the work flow in proteomic analysis.