The Proteome: Discovering the Structure and Function of Proteins

By: Jill Adams, Ph.D. (Freelance science writer in Albany, NY) © 2008 Nature Education


Proteins are continually being synthesized, modified, and degraded, and vary among species, tissue, and even cells — how do you capture and describe this ever-changing proteome?

Proteomics is the large-scale study of the structure and function of proteins. In its strictest sense, the word "proteome" refers to the set of proteins encoded by the genome, including the added variation of posttranslational modification. You might say that DNA is the blueprint for life, and proteins are the tools that make living machines work.

The proteome is neither as uniform nor as static as the genome. Whereas the genetic code is made of four nucleotides and the sequence of these nucleotides is identical in every cell of an organism, proteins are built from 20 different amino acids, and post–translational modification adds other chemical constituents to these molecules, including sugars, fats, phosphates, and even other proteins! In addition, proteins come in different isoforms, are churned through metabolic and degradative pathways, are alternatively spliced, and often link with one another to form complexes made up of multiple proteins. Moreover, the set of proteins produced by a cell varies depending on cell type, cell shape, cell function, what tissue the cell resides in, and what signals the cell receives from its environment, not to mention what developmental stage the cell is in.

Benefits of Proteomics

Given the number of proteins that can be produced by individual organisms, it seems that proteomics may allow greater understanding of the complexity of life and the process of evolution than the study of the genetic code alone. Within the proteome, the many observed layers of complexity begin with an RNA processing mechanism called alternative splicing (Figure 1), in which a single gene can produce multiple versions of a protein. One of the most extreme examples of alternative splicing is the Down syndrome cell adhesion molecule in fruit flies, in which the Dscam gene can give rise to an

Figure 1

Figure 2
astonishing 38,000 distinct protein variants (Schmucker et al., 2000). Another example involves the production of neurexins in mammals; here, three genes give rise to over 1,000 distinct proteins within the mammalian brain (Ullrich et al., 1995). Beyond alternative splicing, post–translational modifications are another source of protein variation. According to Brett et al. (2001), "more than 200 different types [of post–translational modifications] are known, and it is predicted that, on average, for each human gene three different modified proteins with different functions are produced" (Figure 2).

Proteomics doesn't only reveal information about life's complexity, however; it also provides insight into the vibrancy of cells and their preparedness to react. Cells and tissues respond to signals and changes in their environment, and changes in the proteome must mirror that. In fact, early changes in the health of a tissue may be detectable by changes at the proteomic level. Researchers are just beginning to take advantage of measurable changes in protein profiles to assess disease; for example, assaying serum proteins using a chip–based mass spectrometry system reveals a difference in protein profiles between men who have benignly enlarge prostates and those who have prostate cancer (Adam et al., 2002). Moreover, the difference in profiles is robust enough to use as a predictive diagnostic tool.

**Studying the Proteome**

In short, the proteome is an ever–changing swarm of modified proteins that differs from cell to cell—which poses significant challenges for scientists seeking to capture and describe it. In the words of Mike Tyers and Matthias Mann (2003), two scientists who have spent their careers studying the proteome (even before the term "proteomics" was coined), "[a]ll of these difficulties render any comprehensive proteomics project an inherently intimidating and often humbling exercise." Still, Tyers, Mann, and many other researchers view these difficulties as challenges to be broached and problems to be solved as progress continues on this worthy enterprise.

The first complete eukaryotic genomic sequence was of the yeast Saccharomyces cerevisiae (Goffeau et al., 1996). Prior to sequencing, a rich knowledge of yeast biology had already been acquired through decades of hypothesis–driven research with this model organism. Today, even more genomics data and information resources are available to scientists, including the Yeast Protein Database and the Saccharomyces Genome Database. An ambitious study sought to
integrate these genomic and proteomic data after experimental manipulation of a well-studied metabolic pathway in yeast, the galactose utilization pathway. A mathematical model was created and used to predict previously unknown interactions within the pathway and with other cellular processes. Some of these predictions were then verified experimentally (Ideker et al., 2001).

Systems biology approaches will continue to detect connections between broad cellular functions and pathways that were neither apparent nor predictable despite decades of biochemical and genetic analysis of the biological system in question. For example, one study looked at quantitative protein profiling in cells with and without the oncogene Myc, one of the most frequently altered genes in human cancer (Shiio et al., 2002). Here, the researchers noted differences in the adhesion molecules associated with these cells—differences that may underlie the morphological changes that lead to unchecked proliferation in cancer. This finding represents how proteomics can enrich our understanding of the interdependence of cellular processes.

Currently, researchers’ ability to collect large proteomic data sets is greater than their ability to interpret or integrate that information. Thus, the need for bioinformatics algorithms and software tools will likely remain high for some time to come. While improvements in proteomic technologies will likely accelerate research involving single-celled organisms, the additional layers of complexity and organization in multicellular organisms will necessitate grander conceptual schemes, such as those associated with systems biology (Figure 3). Beyond relating genes to transcripts, transcripts to proteins, or proteins to functions, systems biology seeks to integrate all of these layers to achieve a fuller understanding of normal function, disease, and development.

**References and Recommended Reading**


Jensen, O. N. Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology* 6, 391–403 (2006) doi:10.1038/nrm1939 (link to...


