2 The Proteome

2.1. The Proteome and the Genome

Each of our cells contains all the information necessary to make a complete human being. However, not all the genes are expressed in all the cells. Genes that code for enzymes essential to basic cellular functions (e.g., glucose catabolism, DNA synthesis) are expressed in virtually all cells, whereas those with highly specialized functions are expressed only in specific cell types (e.g., rhodopsin in retinal pigment epithelium). Thus, all cells express: 1) genes whose protein products provide essential functions, and 2) genes whose protein products provide unique cell-specific functions. Thus, every organism has one genome, but many proteomes.

The proteome in any cell thus represents some subset of all possible gene products. However, this does not mean that the proteome is simpler than the genome. In fact, the opposite is certainly true. Any protein, though a product of a single gene, may exist in multiple forms that vary within a particular cell or between different cells. Indeed, most proteins exist in several modified forms. These modifications affect protein structure, localization, function, and turnover.

In this chapter, we look at the proteome in five different ways. First, we briefly consider the "life-cycle" of proteins—from their appearance as translation products in ribosomes to their many modifications and their ultimate degradation. Second, we consider proteins as modular structures that can be classified in groups based on sequence motifs, domain structures, and biochemical functions. Third, we consider the distribution of the genome into functional families of proteins.

From: *Introduction to Proteomics: Tools for the New Biology*
By: D. C. Liebler © Humana Press, Inc., Totowa, NJ

Fourth, we look at the proteome through genomic sequences, which indicate the diversity and redundancy of functions in living systems. Finally, we consider the factors that dictate how much of any protein is present in a cell at any one time, and how that influences the difficulty of finding it by analytical proteomics methods.

2.2. The Life and Death of a Protein

Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes. In most cases, the initial polypeptide-translation product undergoes some type of modification before it assumes its functional role in a living system. These changes are broadly termed "posttranslational modifications" and encompass a wide variety of reversible and irreversible chemical reactions. Approximately 200 different types of posttranslational modifications have been reported. Some of these are summarized in **Fig. 1**, which depicts the life cycle of a prototypical protein.

The protein is born as a ribosomal translation product of an mRNA sequence. Folding and oxidation of cysteine thiols to disulfides confers secondary structure on the random-coil polypeptide. A number of "permanent" modifications, such as carboxylation of glutamate residues or removal of the N-terminal methionine, can occur early in the life of the polypeptide. Further processing in the Golgi apparatus often results in glycosylation. Specific delivery of the protein to specific subcellular or extracellular compartments is often achieved with leader or signal sequences, which may be proteolytically cleaved. Prosthetic groups may be added. Combination with other proteins forms multisubunit complexes. Palmitoylation or prenylation of cysteine residues assists anchoring of proteins in or on membranes. These more or less "permanent" modifications and transport ultimately result in the delivery of functional proteins to specific locations in cells.

At their cellular destinations, proteins carry out their many functions. The activities of many proteins are then controlled by posttranslational modifications. The most prominent and best-understood of these is phosphorylation of serine, threonine, or tyrosine residues. Phosphorylation may activate or inactivate enzymes, alter protein-protein interactions and associations, change protein structures, and target proteins for degradation. Protein phosphorylation regulates protein function in diverse contexts and appears to be a key switch

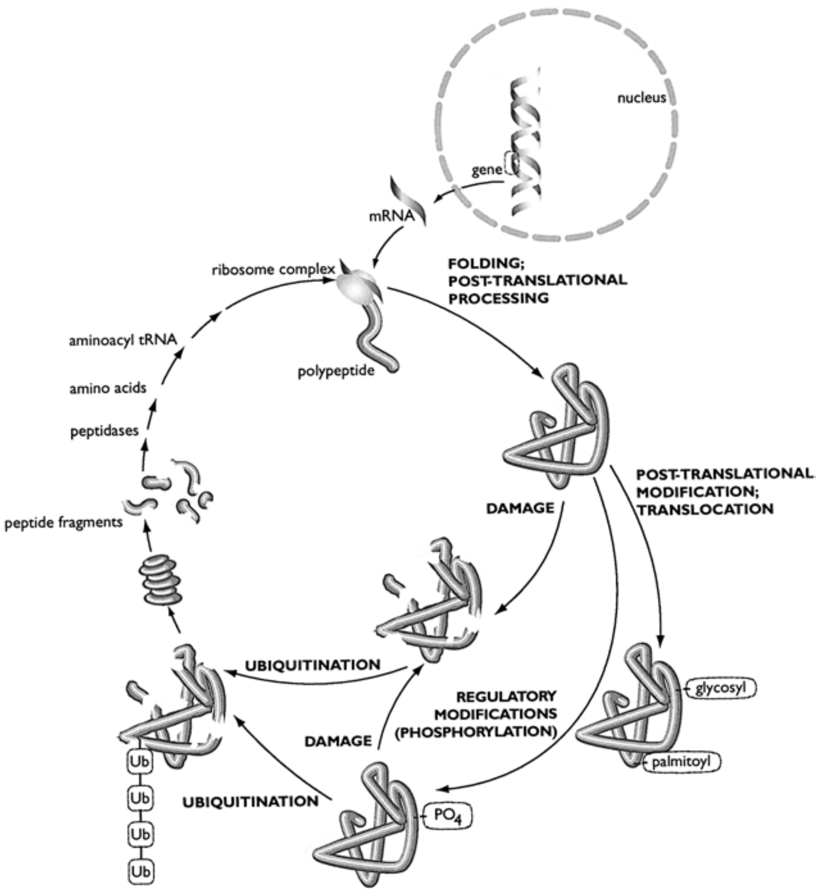

Fig. 1. The life cycle of a protein.

for rapid on-off control of signaling cascades, cell-cycle control, and other key cellular functions.

Proteins also are subject to wear and tear. The ubiquitous presence of free radicals and other oxidants in biological systems leads to oxidative protein damage. Several amino acids are susceptible to oxidation, particularly cysteine thiols. Methionine, tryptophan, histidine, and tyrosine residues also are easily oxidized. Proteins also are subject to attack by products of lipid and carbohydrate oxidation, including reactive α,β-unsaturated carbonyl compounds. In addition to these endogenous sources of protein modification, environmental agents, including radiation, chemicals, and drugs can covalently or oxidatively modify proteins. Many of these modifications can inactivate proteins, but virtually all produce some modifications of protein structure.

Protein modifications appear to be critical to initiating processes that ultimately degrade proteins. Phosphorylation of some proteins is rapidly followed by conjugation with ubiquitin, which leads to degradation by the 26S proteasomal complex. There evidently are other stimuli for protein ubiquitination and turnover, including oxidative damage and other protein modifications. Proteins also undergo degradation by lysosomal enzymes.

The foregoing sketch of the life of a protein illustrates a key point about the proteome. Any protein may be present in many forms at any one time in a cell. Collectively, the proteome of a cell comprises all of these many forms of all expressed proteins. This certainly makes the proteome bewilderingly complex. On the other hand, the status of the proteome reflects the state of the cell in all its functions.

2.3. Proteins as Modular Structures

Another way to look at proteins is to think of them as modular or mosaic structures. Certain amino acid sequences tend to form secondary structures, such as α-helices, β-sheets, or random-coil structures. However, specific amino acid sequences and secondary structures derived from these sequences also confer unique properties and functions. In this way, segments of amino acid sequences can be considered as functional building blocks or modules. From these modules, Nature has assembled a tool box from which to build proteins with diverse, yet related functions.

The modular units in proteins that confer specific properties and functions are referred to as “motifs” or “domains”. These are recognizable sequences that confer similar properties or functions when they occur in a variety of proteins. In common usage these terms often overlap. In some cases, amino acid sequences within motifs and domains are highly conserved and do not vary from protein to protein. In other cases, some key amino acids occur in a reproducible relationship to each other in a sequence, even though various substitutions in other amino acids occur.

Even some short sequences can confer specificity for certain modifications. For example, proteins that undergo N-glycosylation tend to display a tripeptide sequence "Asn-Xaa-Ser/Thr," in which the target asparagine is followed by any amino acid and then either a serine or threonine residue. If the "Xaa" is a proline, glycosylation is blocked. Although this sequence does not ensure N-glycosylation, it does provide a signature motif that can offer clues to possible biochemical roles.

Longer amino acid sequences often form domains, which confer specific properties or functions on a protein. Some domain structures simply refer simply to sequences that confer a bulk physical property to a segment of the polypeptide, such as transmembrane domains, which simply form helices that span a lipid bilayer membrane. Other domain structures provide hydrogen bonding or other contacts for key enzyme substrates or prosthetic groups. For example, eukaryotic serine/threonine kinases display a core domain that includes a glycine-rich region surrounding a lysine residue involved in ATP binding and a conserved aspartate residue that functions as a catalytic center. In many cases, domains are made up of combinations of units of secondary structure, such as helix-loop-helix domains.

The significance of motifs and domains for proteomics is that they represent the translation of peptide sequence to protein functions. In cases where domains and motifs confer known properties or functions, their occurrence in proteins of unknown function offer hints as to their cellular roles. In short, analytical proteomics can define sequence and sequence can define biological function.

2.4. Functional Protein Families

Another way to look at the proteome is to divide it into families of proteins that carry out related functions. For example, some proteins serve structural roles, some are participants in signaling pathways, and others handle essential metabolic chores such as nucleic acid synthesis or carbohydrate catabolism. Based on classification by domain content and associated functional roles, Venter and colleagues (2001) estimated the division of protein functions in proteins encoded by the human genome (**Fig. 2**).

Enzymes involved in intermediary metabolism and nucleic acid metabolism account for about 15% of the proteins represented in the proteome. Proteins associated with structure and protein synthesis

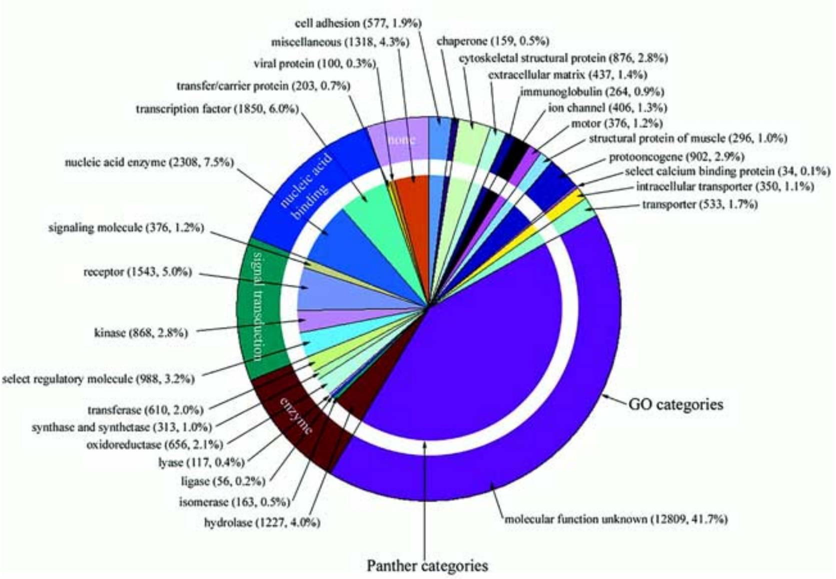

Fig. 2. Functions assigned to predicted protein products of human genes. (Reprinted with permission from Venter et al. (2001) *Science* **291:** 1304–1351. Copyright 2001, American Association for the Advancement of Science.)

and turnover (cytoskeletal proteins, ribosomal proteins, chaperones, and mediators of protein degradation) account collectively for another 15–20%. Another 20–25% consists of signaling proteins and DNA binding proteins. Although these numbers offer a useful perspective on how the genome is divided by protein functions, they do not tell us how much of any protein or protein class is expressed at any given time in a cell. Approximately 40% of the genome encodes protein products with no known function. Assigning functions to these gene products represents the most fundamental challenge for human functional genomics.

2.5. Deducing the Proteome from the Genome

One of the most interesting questions facing researchers who characterize genomes in an organism is "How many genes are there?" The answer to this question can give us some idea of how many

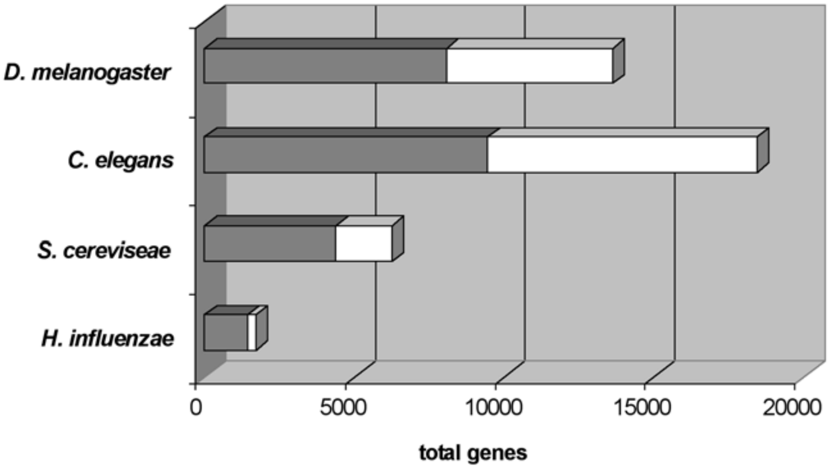

Fig. 3. Predicted protein products of genes from *H. influenzae* (1,709 genes), *S. cerevisiae* (6,241 genes), *C. elegans* (18,424 genes), and *D. melanogaster* (13,601 genes). The dark bar segments depict genes coding for unique proteins; the light bar segments depict genes coding for paralogs. (Adapted with permission from Rubin et al. (2000) *Science* **287:** 2204–2215. Copyright 2000, American Association for the Advancement of Science.)

proteins may exist in the proteome. Complete genomic sequences of several organisms have been completed and these data have allowed analysts to predict the products of all the organism's genes. Moreover, based on the predicted amino acid sequences of each gene product, these proteins have been classified on the basis of the domains and sequence motifs they contain. For example, 119 of the genes of the *Saccharomyces cerevisiae* genome encode proteins with eukaryotic protein kinase domains, whereas 47 others encode proteins with C2H2-type zinc-finger domains. Comparisons of domain-sequence characteristics with genomic sequences reveals many other protein types encoded in an organism's genome.

Recent analyses of the *S. cerevisiae*, *Caenorhabditis elegans*, and *Drosophila* genomes have revealed very interesting relationships between the size of the genomes and the predicted content of the proteomes for these organisms. Gerald Rubin and colleagues have classified the predicted protein products of the *H. influenzae, S. cerevisiae, C. elegans,* and *Drosophila* genomes based on the presence of specific domains (**Fig. 3**). Comparison of all the predicted protein products indicated the occurrence of proteins whose sequence differed only slightly from others in the genome. Correction for these redundant protein products, termed "paralogs," allowed the calculation of a "core proteome" for each organism. This core proteome represents the basic collection of distinct protein families for an organism.

A look at the the core proteomes for these organisms illustrates two interesting aspects of the proteome. First, the relationship between the complexity of an organism and the number of genes in its genome is not simple. Certainly, the yeast has more genes than the bacterium, yet fewer than the worm and the fly. However, the fly (*Drosophila melanogaster*) is a much more complicated organism than the worm (*C. elegans*), yet it has fewer genes (13,601 vs 18,424 in the worm) and a smaller core proteome (8065 distinct proteins vs 9543 in the fly). This suggests that biological complexity does not come simply from greater numbers of genes. Instead, more complex regulation of the genes and the functions of the protein products may account for the greater complexity of the fly.

Second, the number of paralogs increases dramatically in the worm and the fly. This reflects the fact that about half of the genes in the worm and the fly are near-duplicates of other genes. These duplicate-containing gene families often appear as gene clusters on the same chromosome.

The recent completion of the human genome sequence has provided evidence that the human genome encodes between 30,000 and 40,000 genes. In view of the tremendous difference in complexity of the human organism compared to the worm, it is indeed surprising that the human genome encodes only about twice as many genes as that of the worm. Reliable estimates of the numbers of unique genes vs paralogs are not yet available. Nevertheless, it is already becoming axiomatic that the complexity of the human organism lies in the diversity of human proteomes, rather than in the size of the human genome.

2.6. Gene Expression, Codon Bias, and Protein Levels

One of the key issues encountered by investigators who study the proteome is how much of a particular protein is expressed in a cell.

Expression levels of proteins vary tremendously, from a few copies to more than a million. It is important to realize in this context that the level of a protein expressed in a cell has little to do with its significance. Essential enzymes of intermediary metabolism or structural proteins often are present at levels in the thousands of copies per cell or more, whereas certain protein kinases involved in cell-cycle regulation are found at only tens of copies per cell. *S. cerevisiae* contains approx 6000 genes, of which about 4000 are expressed at any given time, based on measurements of mRNA levels.

The level of any protein in a cell at any given time is controlled by: 1) the rate of transcription of the gene, 2) the efficiency of translation of mRNA into protein, and 3) the rate of degradation of the protein in the cell. Gene expression certainly can dictate protein levels to a considerable extent. However, a number of studies indicate that gene expression *per se* does not really correlate that well with protein levels. This finding certainly reflects the influences of the other two factors mentioned earlier. It also is an important reminder of the limitations of gene-expression analyses (such as microarrays).

Many genes are regulated by inducible transcription factors, which are regulated in turn by a wide variety of environmental influences. However, an intrinsic determinant of the level of expression of many genes is a phenomenon referred to as "codon bias." This term describes the tendency of an organism to prefer certain codons over others that code for the same amino acid in the gene sequence. Thus, genes containing codon variants that are less preferred tend to be expressed at a lower level. Calculated codon bias values for yeast genes range from approx –0.2 to 1.0, where a value of 1.0 favors the highest level of gene expression. Most yeast genes display codon bias values of less than 0.25 and are expected to be expressed at relatively low levels.

Studies in yeast have compared protein levels, mRNA expression, and codon bias for a number of proteins. While there is some disagreement as to the particulars, the following generalizations can be drawn.

- Genes with low codon bias values tend to be expressed at low levels, whether analyzed on the basis of mRNA expression or protein levels.
- mRNA levels correlate poorly ($r < 0.4$) with protein levels when genes with codon bias values of 0.25 or less (i.e., most genes)

are considered. However, the correlation between mRNA levels and protein levels is much higher ($r > 0.85$) for the most highly expressed genes (i.e., those with codon bias values above 0.5).

- Longer-lived proteins appear to be present in higher abundance than short-lived proteins (i.e., those proteins that are degraded rapidly).

Thus, although gene-expression measurements may indicate changes in protein levels, it is difficult to infer protein expression from gene expression.

2.7. Conclusion and Significance for Analytical Proteomics

The proteome in essentially any organism is a collection of somewhere between 30 and 80% of the possible gene products. Most of these proteins are expressed at relatively low levels (10^1–10^2 per cell), although some are expressed at much higher levels (10^4–10^6 per cell). Regardless of the absolute level of expression of the polypeptide gene products, most proteins exist in multiple posttranslationally modified forms. This situation poses the greatest challenge for proteomic analysis: we must find ways to detect a large number of distinct molecular species, most of which are present at relatively low levels and many of which exist in multiple modified forms. The next section of the book describes the tools we can bring to bear on this daunting analytical problem.

Suggested Reading

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40.

Coghlan, A. and Wolfe, K. H. (2000) Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* **16,** 1131–1145.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19,** 1720–1730.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., et al. (2000) Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.