nature biotechnology
VOLUME 21 • NO 3 • MARCH 2003

A cast of thousands

In his keynote speech celebrating the discovery of the DNA double helix at last month's Miami *Nature Biotechnology* Winter Symposium, Jim Watson referred to DNA as "the script" and proteins as "the actors" in biology. This analogy certainly seems apt (if not completely accurate). These days, biology often can seem like a fast-paced, all-action, blockbuster movie—one in which the script has been recycled many times and the actors play roles that are difficult to discern. (The performance of proteins is also often adversely affected by drugs and alcohol ... but maybe that is taking the analogy too far.)

This issue focuses on the current status of technologies for analyzing the biological actors—the tens of thousands of proteins that are expressed in a cell at any given time. In it, we ask how far has quantitative protein analysis progressed and what the key challenges are for this field?

The term proteomics (note to self: congratulations for getting this far without mentioning it) was coined back in 1994 by Marc Wilkins at a conference on genome and protein maps in Sienna, Italy. Today, there are almost as many interpretations of proteomics as there are remakes of Frankenstein. It can be defined as the systematic determination of protein sequence, quantity, modification state, interaction partners, activity, subcellular localization, and structure in a given cell type at a particular time.

Proteomic approaches are currently mostly focused on addressing specific questions in biology (for example, what is the phosphorylation state of a set of particular kinases or which proteins are involved in formation of the nuclear pore?). As technologies improve in sensitivity and reproducibility, and data sets become more complete, proteomics should also facilitate the application of systems approaches for modeling complex phenomena, reveal protein biomarkers that can be used in diagnostic and predictive screens for disease, and pinpoint proteins as potential drug targets or therapeutics.

Unfortunately, putting theory into practice has proven difficult. Proteins, like actors, are notoriously temperamental and difficult to handle. In the human genome, over 35% occur as splice and translation isoforms. A protein's function can also alter depending on the proteins around it (for example, a receptor for adrenomedullin can be turned into a receptor for calcitonin-related binding protein just by changing which chaperone takes it through the Golgi apparatus). In addition, covalent modifications of proteins (for example, phosphorylation, glycosylation, or farnesylation) often determine activity. Although over 200 modifications have now been documented (it is not yet clear how many have functional relevance), we are still in the early stages of developing technologies to analyze and detect them (see p. 255). New reagents are needed to allow profiling of post-translational modifications, and novel probes and imaging technologies are required to study protein localization at the subcellular level.

Another problem facing researchers is the bewildering diversity of protein molecules and their activities. Many types of proteins (for example, membrane proteins and highly alkaline proteins) are particularly difficult to purify and study *in vitro*, requiring special fractionation or enrichment protocols. Methods for preparing membrane proteins (and for that matter any protein on chip platforms) in their native conformation have also proven elusive, although it appears that, here, progress is being made (see pp. 262 and 223).

Typically, the dynamic range of protein abundance in yeast spans five to six orders of magnitude. In contrast, human blood protein levels can differ by a factor of 10^9 (for example, 10^9 pg/ml for serum albumin compared with 0–5 pg/ml for interleukin 6). This presents a formidable challenge for those attempting to profile low-abundance proteins in human plasma. The fact that two-dimensional gel electrophoresis and tandem mass spectrometry typically have a dynamic range of only 10^2 to 10^4 illustrates the inadequacy of current technology. The absence of an equivalent of PCR for amplifying minute amounts of proteins dictates that proteomics is necessarily limited by the levels of substrate in the natural source. It is no surprise that many believe proteins of low abundance will turn out to be those of greatest clinical (and biological) interest.

Thus, we are still a long way from achieving a complete proteome map for any organism (whether a prokaryote or a human being). In yeast, for example, we still have no handle on the function of around 1,800 of the 6,200 predicted genes, despite the availability of the genome sequence since 1996. When one considers that a human cell expresses many more than 30,000 proteins at any given time, and that only a few thousand proteins can be detected on the average gel or by mass spectrometry, the magnitude of the task ahead is clear. Part of the solution to this problem will clearly lie in improving the throughput of mass spectrometry approaches by refining instrumentation and software for data analysis (see p. 221). The fact that most current methodologies produce data that are mostly qualitative or quantitative only in relation to a reference sample also will have to be addressed: assaying technologies need to be developed that can provide absolute measurements of proteins in samples.

Sydney Brenner has commented that "the more you annotate the genome, the more you make it opaque." The time is coming when proteomics research will have to move away from merely collating lists of proteins and mapping interactions to a more integrated approach in which proteomic data sets are interpreted in the context of many other types of biological data. To accomplish this, we will need to establish a data infrastructure that allows the capture and dissemination of proteomics data (see p. 247). And we will also need to embrace systems biology approaches that detect feedback loops and connections between pathways that have eluded decades of biochemical and genetic analysis carried out on an isolated, reductionist level. Actors work as a cast, not as individuals. Proteins are no different.