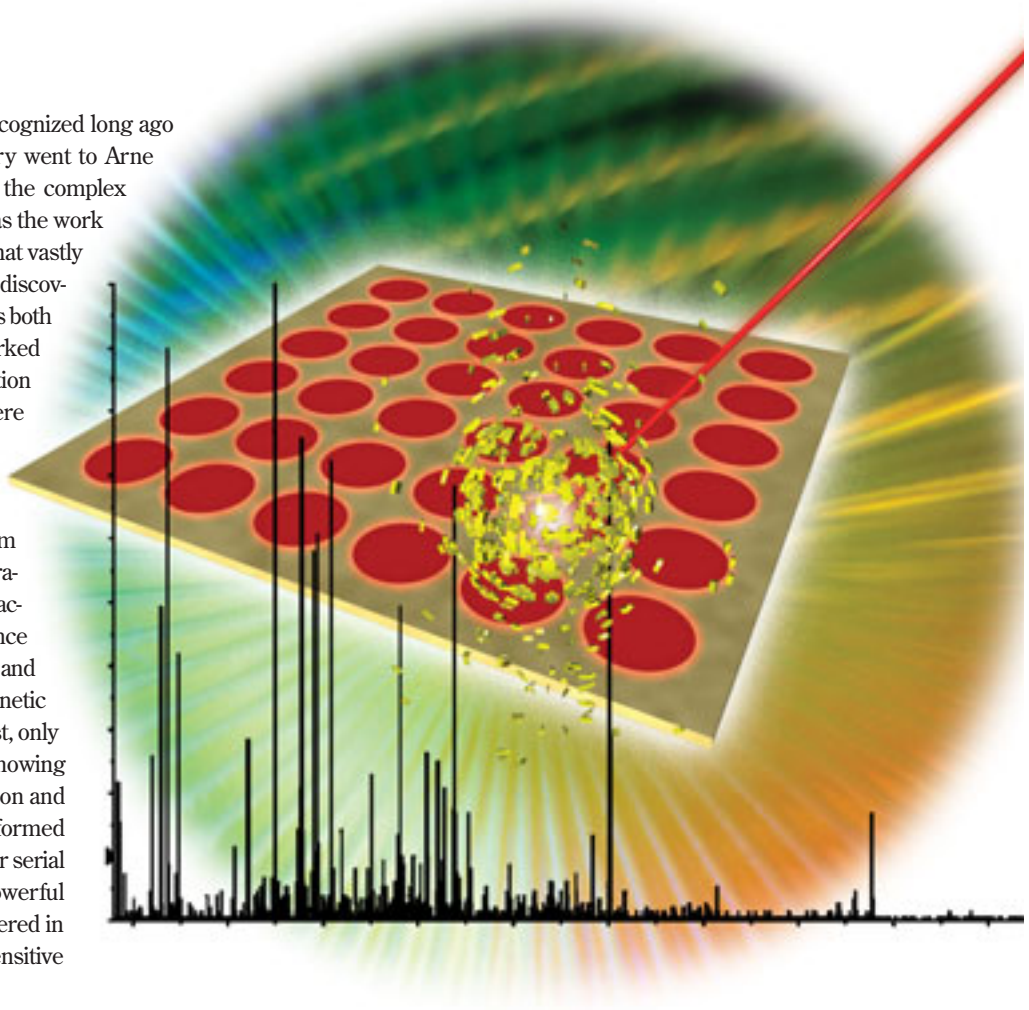
Probing proteins

Mass spectrometry is answering big questions about small molecules.

BY GARRY CORTHALS



The importance of protein science was recognized long ago when the 1948 Nobel Prize in Chemistry went to Arne Tiselius for his discoveries concerning the complex nature of serum proteins. However, it was the work of Franklin, Collins, Crick, and Watson that vastly impacted our thinking about biology. By discovering the structure of DNA, these scientists both captivated the world's attention and sparked an interest in better understanding the information encoded by our genes. We are now at a stage where scientists are sequencing genomes faster than people can download music illegally.

In 10 years, genomic research has delivered hundreds of complete sequences, ranging from human and mouse species to various viruses, parasites, naturally occurring plasmids, organelles, eubacteria, archaea, and fungi. The concept of sequence knowledge has enabled us to rationalize its context, and it has become clear that merely deciphering a genetic sequence does not reveal the whole picture. At best, only the potential of a system can be understood by knowing the genetic components. Systematic rapid detection and quantitation of biological activity can now be performed via differential-display PCR, cDNA microarrays, or serial analysis of gene expression (SAGE). These powerful tools, together with complete genomes, have ushered in the need for new high-throughput and highly sensitive technologies that can similarly measure proteins.

Knowledge of proteins is important, as they are the mature products of genes and represent end points of gene expression; they contribute to, and catalyze, the changes we study in a biological system. The contribution of protein measurements to existing gene-expression measurements complements and extends our knowledge. There are three areas where proteins have distinctive control and regulation over biological effects, starting with their temporal and spatial expression, which is not apparent from genomic or gene expression analysis. Second, static and dynamic protein post-translational modifications (PTMs) are richly varied, with more than 200 different types having been documented to date. Usually there is no predefined knowledge of their location, but even with the knowledge of a "consensus" sequence (a possible PTM event), there are few rules describing when and where in the cell specific PTMs occur, even for highly studied events such as phosphorylation. Finally, inducible protein–ligand interactions cannot be measured with genomic technologies.

In recent years, research has emphasized that cellular functions are not carried out by singular components but are performed by

"modules" made up of many interacting molecules, mostly proteins. These functional modules exist as a critical level of biological organization. Currently, only proteomic technologies can generate data based on interacting proteins that will lead to understanding biological regulation.

Profiling technologies

In addition to enhancing the ability and speed of making protein measurements, any analysis must be performed in a systematic, quantitative, and reproducible manner. Figure 1 shows typically used technologies that allow for global protein analysis. The mass spectrometer is central to most screening techniques because it enables outstanding speed and accuracy in protein identification.

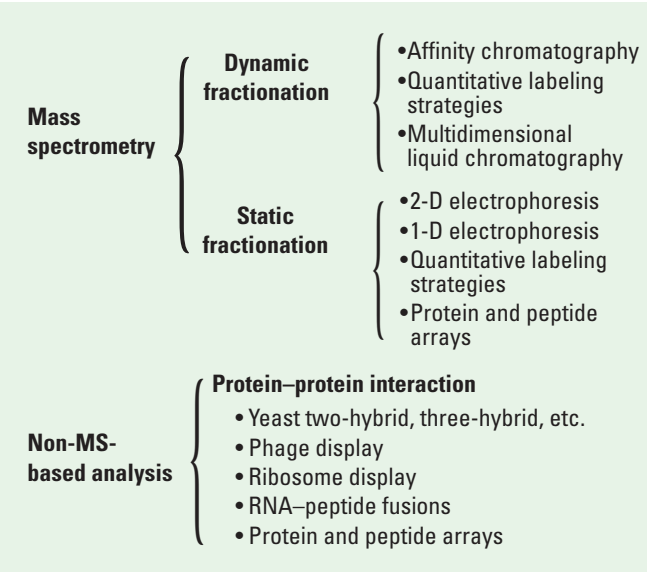


Figure 1. From single molecules to complex mixtures and protein modules. Dynamic fractionation implies a time-limited analysis, whereas static fractionation allows for a workflow that can be paused at specific steps. (Adapted with permission from Corthals, G. L.; Nelson, P. S. *Pharmacogenomics J.* **2001,** *1,* 15–19.)

Until about 2000, proteomics was almost exclusively performed by qualitative and quantitative display of tissue extracts or cellular protein expression profiles via two-dimensional gel electrophoresis (2-DE) followed by MS. With 2-DE, protein arrays are generated, and one can pinpoint differences between different biological states via differential pattern display. The number of applications and uses of proteomic technologies has become widespread because of this relatively simple setup. Most labs now have MS instruments with workflows incorporating 1- and 2-DE, multidimensional LC, affinity chromatography, and quantitative labeling strategies. While less common, non-MS-based workflows are important for measurements of protein–ligand interactions. Such technologies include yeast two-hybrid (three-hybrid, etc.), phage display, ribosome display, RNA–peptide fusions, and other protein and peptide arrays (for more information, see *www.biochipnet.de*).

In recent years, much effort has been directed toward developing technologies that enable global quantitative expression analysis to complement or bypass two-dimensional gels. It is generally accepted that, although 2-DE is a mature and widely practiced method, it is unable to display all the proteins within a biological system. MS-based quantitative measurements look like they will provide the key to analyze more proteins. These methods are designed to automate, accelerate, and more precisely measure protein changes between various disease states. Interestingly, they are essentially based on the venerable technique of stable isotope labeling.

A breakthrough to global quantitative MS-centered (i.e., high-speed) proteomics came five years ago, when Ruedi Aebersold and colleagues crafted new molecules to address the needs of discovery science through their clever work on the isotope-coded affinity tag (ICAT) reagents. The method using ICAT reagents (Figure 2) consists of four sequential steps. First, the side chains of cysteinyl residues in a protein mixture are reduced and alkylated with the isotopically light form of the ICAT reagent. Equivalent groups in a sample representing a second disease (or cell) state are derivatized in a similar manner with the isotopically heavy reagent. The two samples are then combined and enzymatically digested to generate peptide fragments. Third, the tagged peptides (those with a cysteine) are isolated by avidin affinity chromatography. Finally, the affinity-tagged peptides are released from the column and analyzed by LC-MS/MS. This analysis finally reveals both the quantity and sequence identity of the proteins from which the tagged peptides originated. Numerous reports have now emphasized the value of this strategy.

While their application was slow initially, they were widely discussed. In effect, these chemicals provided proof of a concept, which was to quantitatively compare two massively complex protein mixtures that were originally derived from cells, tissues, and body fluids. Indeed, this strategy captured the attention of many researchers and spurred the quest for alternative methods to apply to the countless applications in proteomics. There is now a collection of compounds that are slightly similar to this chemistry, but they are all based on the same concept, with or without affinity tag—the latest of which are Applied Biosystems' iTRAQ reagents.

What distinguishes the iTRAQ chemicals—which include four isobaric reagents—from others is the ability to measure up to four samples simultaneously, whereas other methods allow the comparison of only two. The iTRAQ labeling chemistry is peptide-oriented and involves the incorporation of an isobaric compound specific to amines in up to four different peptide mixtures. Relative or absolute quantitation—through labeling of peptides of known concentration—can be performed, and the method is compatible with all current workflows.

Nonchemical labeling strategies also exist. In SILAC (stable isotope labeling by amino acids in cell culture), essential amino acids are added to an amino-acid-deficient cell culture medium and, as cells grow, amino acids are incorporated into all proteins as they are synthesized (Figure 2). No chemical labeling or affinity purification steps are performed as with the procedures described above, and the method is likely compatible with many cell culture conditions, including primary cells. Mathias Mann and colleagues at the University of Southern Denmark have shown that incorporation can be close to complete and that cells show no phenotypic differences in the presence of labeled media. They applied SILAC

to the study of mouse C2C12 cells and followed the differentiation from myoblasts into myotubes. This process of muscle differentiation necessarily involves broad changes in protein expression levels as the cells differentiate from one cell type to another. Several proteins were found to be up-regulated during this process.

The aims of scientists are not modest, yet they must be directed toward delineating molecular networks and identifying specific targets.

MS as turnkey

With new chemistries and bioinformatics tools, MS has quickly emerged as a potent and indispensable technology from among the collection of technical disciplines used in proteomics. Advances in MS intensify information density and improve data quality, while the range of applications steadily grows. Important contributions made over the past few years have relied on the integration of specialized LC with electrospray ionization (ESI)-MS and matrix-assisted laser desorption/ionization (MALDI)-MS workflows. However, while the variety of chemistries and approaches speaks for our ingenuity, there still is a need for general global quantitative methods that can be used and integrated with MS and various separation techniques, such as 1- and 2-DE and LC. All these workflows have typically identified proteins by peptide mass fingerprinting with MALDI-MS or direct sequence analysis of peptides through data-dependent microcapillary LC-MS/MS. Quantitation occurs mainly at the peptide level, although it would be desirable to have this performed at the MS/MS level, as with the iTRAQ chemicals, to exploit additional accuracy of measurement.

Tandem mass spectrometers have played a crucially enabling role for analyzing complex mixtures. Most instruments use quadrupoles or ion traps for their initial mass analysis, followed by further mass analysis after fragmentation. They can select ions of a particular mass (m/z) from a mixture of ions, fragment them by a process called collision-induced dissociation (CID), and then record the precise masses of the resulting fragment ions. When this process is applied to peptide ions, a peptide's amino acid sequence can be deduced. Hence, tandem MS (MS/MS) enables unambiguous protein identification as it directly confers a gene sequence to the protein sequence.

The newest entrants, MALDI time-of-flight/time-of-flight (TOF/TOF) analyzers, are faster, high-resolution tandem mass spectrometers specifically designed for rapidly sequencing peptides. TOF/TOF instruments combine the advantages of high sensitivity for peptide analysis with comprehensive peptide fragmentation. In seconds, multiple MS/MS spectra can be generated from selected peptides until enough information is obtained or the sample is consumed. TOF/TOF instruments use a different ion gate than MALDI-TOF instruments that allows improved precursor-ion selection. Different combinations of MALDI matrix and collision gas determine the amount of internal energy deposited by the MALDI and CID process, which provides control over the extent and nature of the fragment ions observed. One acquires information-rich spectra, often with high-energy fragments. These spectra can enable the validation of ambiguous database search results that are difficult to validate manually.

Nowadays, LC/LC-MS/MS procedures using an ESI source are empowering high-throughput proteomics projects with great success. Although very useful for the process, there are diminishing returns to using this method in a repetitive manner. In these workflows, one generally gathers extensive data on previously identified proteins, through the exhaustive generation of MS/MS spectra for many ions generated. Here, µLC systems are directly coupled to ESI-MS/MS instruments, which have the limited time of the LC run to perform all peptide analyses. Often, analyses are repeated to maximize information content, but repetition also generates massive data redundancy that has to be analyzed and validated. Currently, analysis and validation are a considerable bottleneck in large-scale proteomics workflows. An ideal strategy would decouple the dynamic time limitations of an LC run from the MS analysis. With the use of a MALDI-TOF/TOF mass spectrometer, one achieves this, as the instrument

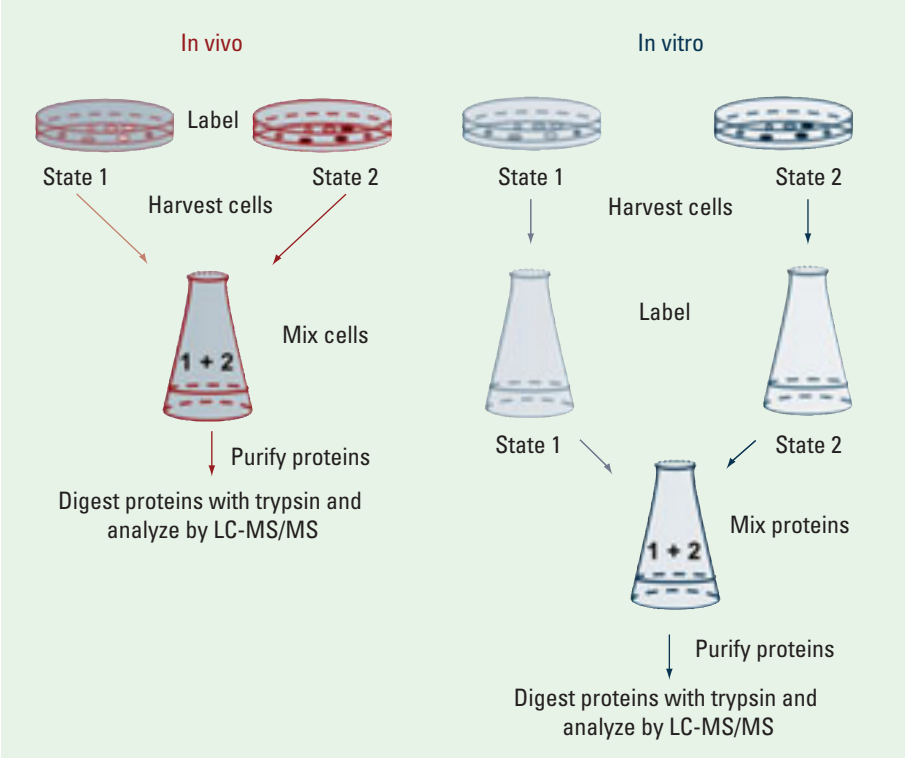


Figure 2. Labeling methods for proteome-wide quantitative analysis. Biological labeling procedures performed in vivo differ from chemical labeling procedures in that labeling of the peptide or protein is achieved by growing cells in media enriched in stable-isotope-containing amino acids. With in vitro chemical labeling, a derivatization reagent is used to "tag" proteins after harvesting from the cells.

is decoupled from the dynamic LC flow; a static flow is deposited on the MALDI target plate. Consequently the instrument can pause at any time, even between acquisition and analysis.

Recently, we have proposed a nonredundant (nr) MS strategy in which data analysis results regulate acquisition. With nrMS, data acquisition and data analysis are used consecutively. Following primary acquisition, a data analysis step is introduced. On the basis of knowledge of the first identified protein in the sample, precursor selection for analysis is then filtered to avoid repetitive identification of peptides from the same protein. Further analysis is then limited to peptides that have mass values different from the mass values of unmodified peptides from these entries. The analysis is repeated until the sample is consumed or all ions are accounted for. A fundamental component of this nrMS workflow is a TOF/TOF analyzer capable of generating peptide mass fingerprints and MS/MS information.

In our nrMS development, we have successfully used the Applied Biosystems 4700 Proteomics Analyzer, which will include this option in its next version of control software. Another advantage of this instrument is its ability to generate high-energy fragments (x-, w-, c-, and d- ions) during CID for de novo sequencing or validating ambiguous database search results. It is also equipped with a high-frequency laser (200 Hz) that enables high-speed (10 times faster than other instruments) data acquisition and therefore high-throughput analysis. The nrMS strategy can potentially be used with any kind of MS/MS platform using a MALDI ionization source.

Integrated technologies

There is enormous potential for applying wide-ranging transcriptomics and proteomics technologies in clinical and pharmaceutical research. Using these profiling technologies, one can measure gene expression of cell lines and animal models of disease, take tissue biopsies for diagnostic and prognostic purposes, monitor patient disease progress, discover new disease markers, classify diseases at the molecular level, and use and develop technologies for toxicological purposes and drug testing. Whatever the application, one must make choices and apply them diligently.

The use of a multitiered approach involving a number of complementary strategies that, when combined, broaden one's view of the overall protein flux is preferred (e.g., Figure 3). Two-dimensional gels are useful as they provide an immediate visual representation of differences between two states, and quality control of the samples analyzed, between experiments and over time. In addition, isoform changes, often due to PTMs, are readily observed using 2-DE. We use in-house-designed labels for workflows that involve 1-DE gels, as these gels target a different range of proteins than 2-DE gels do.

For workflows that bypass gels altogether, using ICAT reagents is particularly convenient as they reduce sample complexity. Besides ICAT reagents, approaches that use SILAC or iTRAQ reagents may be considered. Although no data reduction is achieved with these approaches, they can be combined with affinity-based labeling strategies (e.g., a biotin tag on Cys residues). Finally, this work is combined with experimental microarray transcript data. We have reported initial results on this approach for prostate cancer, and the work has been expanded to include these new strategies. Likewise, for *Staphylococcus aureus*, a similar approach has been designed in which we are converging two data types—proteomics and transcriptomics.

Aims of scientists are not modest, yet they must be directed toward delineating molecular networks and identifying specific targets that promote the differentiation and apoptotic potential of cancer cells. Similarly, for debilitating bacterial infections, we need to gain a better understanding of the complex mechanisms of antibiotic resistance, and in doing so engage in the full characterization of these bacterial proteomes. For these projects, and most projects including discovery aspects, it is important to apply technologies that can detect molecular changes in the cell without preconceived ideas about what information will be most valuable to monitor, or which profiling platform will have the greatest impact.

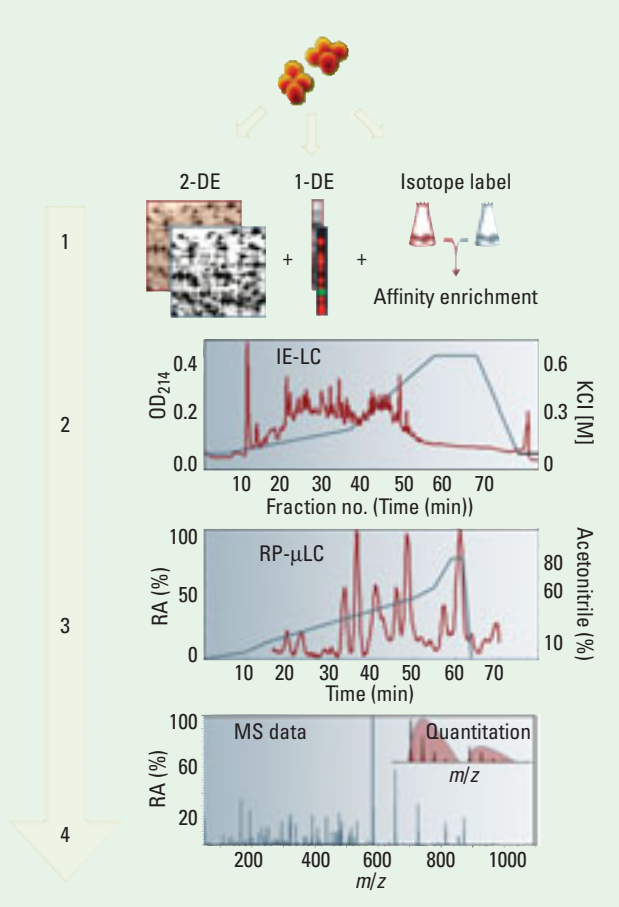


Figure 3. Comprehensive protein profiling using combined approaches. (1) Proteins are separated by various technologies targeting the characteristic and different solubilities of proteins. 2-DE and 1-DE approaches can use labels incorporating fluorescent tags for visualization, whereas quantitation of liquid-based samples for further analysis is enabled by incorporation of chemical or biological isotope tags. 1-DE is also ideal for "double-tagging," where fluorescent and chemical tags are incorporated. (2) Depending on the complexity of the sample, ion-exchange (IE) chromatography can be used prior to (3) microcapillary liquid chromatography (µLC) MS/MS that is typically interfaced with the ionization source of a mass spectrometer. For our work using MALDI-TOF/TOF instruments, we have interfaced the IE-LC/µLC to a MALDI target plate with a robot. (4) MS/MS with the 4700 Proteomics Analyzer then delivers information for protein identification and quantitation.

Garry Corthals is head of the analytical proteomics group at Geneva University Hospital in Switzerland. ■