MILESTONE 20

Protein discovery goes global

The term ‘proteome’ was created in the mid-1990s to describe the entire set of proteins expressed by a cell or organism at any one time. Although the importance of understanding how proteins function in biological systems was realized long before this, the high-throughput characterization of these complex molecules had only recently become possible.

In the early 1990s, Edman degradation was commonly used to resolve the sequences of gel-separated proteins. The process was slow, required large amounts of protein sample and sometimes failed to identify correct peptide sequences, particularly in the presence of protein modifications such as glycosylation or phosphorylation. Looking to speed up the protein identification process, a team led by William Henzel, John Stults and Colin Watanabe began working on a mass spectrometry–based technique to match peptide mass spectra to molecular-weight information inferred from known amino acid sequences—a process now known as peptide mass fingerprinting. Although the method was promising, its low sensitivity prevented widespread adoption.

It was the emergence of matrix-assisted laser-desorption/ionization (MALDI; **Milestone 18**) and electrospray ionization (ESI; **Milestone 15**) that propelled mass spectrometry forward as a sensitive and efficient method for protein identification. Henzel and colleagues were the first to exploit the increased sensitivity of MALDI and demonstrate a rapid peptide-mass-fingerprinting method for identifying proteins from two-dimensional gels. They developed a computational algorithm, Fragfit, that accurately matched peptide masses to the masses of known sequences generated using identical proteolytic digestion techniques. They applied Fragfit to ten proteins isolated from an *Escherichia coli* cell lysate, and, using only three peptide masses per protein, the algorithm uniquely identified each protein from 91,000 protein sequences. This study was one of the first to showcase the use of mass spectrometry for protein identification.

In the following year, ESI-based liquid chromatography (**Milestone 8**) coupled with tandem mass spectrometry (LC-MS/MS; **Milestone 13**) eclipsed the performance of MALDI-MS for protein identification. John Yates and his team developed the SEQUEST algorithm to match experimentally obtained LC-MS/MS spectra to theoretical spectra inferred from every amino acid sequence in the GenPept database. SEQUEST successfully assigned peptides derived from digested *E. coli* and yeast cells to the correct sequences in species-specific databases. This work formed the basis for future ‘shotgun’ proteomics studies by showing that global protein profiles could be obtained from digested complex protein mixtures.

Around the same time, Matthias Mann and Matthias Wilm found that fragmented peptides contained short, identifiable amino acid sequences that, together with the mass of the regions flanking the peptide, could be used to match proteins to known sequences. These so-called sequence tags were shown to have enormous discriminating potential, and, importantly, Wilm and Mann were able to quantify this potential by calculating the likelihood that sequence-tag matches were incorrect. The manual nature of this approach somewhat limited its popularity, however.

Notably, both of these studies emphasized the need for algorithms that could correlate experimental mass spectra to translated gene sequences. The importance of such algorithms became increasingly evident as advances in sequencing technologies spurred a rapid expansion in DNA sequence libraries that could be used for high-throughput protein identification.

As protein-identification tools continued to improve, the demand for accurate quantitation also became apparent. Consequently, stable isotopes were incorporated into samples prior to mass spectrometry analysis to determine the relative abundance of proteins in two samples. One approach involved chemically modifying two protein populations using isotope-coded affinity tags (ICAT); although successful, this was limited to proteins containing specific residues, was difficult to replicate and compared only two samples in a single analysis.

More sensitive quantitation came about through the use of metabolic labeling—a method introduced by Brian Chait and his team in 1999. They showed that by growing pools of cells on medium containing different metabolically labeled amino acids, all cellular proteins could be labeled as they were synthesized. A popular example of this approach, stable-isotope labeling by amino acids in cell culture (SILAC), was published three years later. The following year, a chemical labeling method using tandem mass tags was shown to improve the multiplexing capabilities of protein quantification—allowing eight samples to be analyzed simultaneously. This work gave rise to a range of isobaric tagging methods that remain popular today.

In 2002, William Henzel, John Stults and Colin Watanabe (pictured left to right) won the Distinguished Contribution in Mass Spectrometry Award for their work on peptide mass fingerprinting.
Reproduced from Robinson, C. & Gross, M., Focus on proteomics in honor of the 2002 Distinguished Contribution in Mass Spectrometry Award to W. J. Henzel, J. T. Stults, and C. Watanabe, *J. Am. Soc. Mass Spectrom.* **14**, 929–930 (2003), with kind permission from Springer Science and Business Media.

Another important component of proteomics research was the development of methods that estimate and control for erroneous assignments of mass spectra to known sequences. In particular, the target-decoy approach, introduced by Steven Gygi and colleagues in 2003, remains the gold standard for error assessment in large-scale protein identification studies.

In just over two decades, the field of proteomics has progressed from the analysis of single proteins to the ability to profile near-complete proteomes and detect post-translational modifications (**Milestone 21**). Mass spectrometry–based proteomics has proven to be an indispensable technique for understanding how complex organisms function.

Sarah Perry, Associate Editor, Nature Biotechnology

ORIGINAL RESEARCH PAPERS Henzel, W.J. *et al.* Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90**, 5011–5015 (1993) | Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994) | Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994)

FURTHER READING Wilkins, M.R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/technology* **14**, 61–65 (1996) | Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999) | Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999) | Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics* **1**, 376–386 (2002) | Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003) | Nesvizhskii, A., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003) | Peng, J. *et al.* Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003)