Measuring the Dynamics of the Proteome

Edward M. Marcotte[1]

Department of Chemistry and Biochemistry and Institute of Cell and Molecular Biology, University of Texas, Austin, Texas 78712, USA

The modest goal of proteomics is to collect protein expression data to the same extent that one can now collect mRNA expression data with DNA microarrays. If proteomics lives up to its promise, we can expect to catalog thousands of different proteins from a sample of cells, then vary the cell growth conditions and see how the protein expression changes. Several technologies are vying to deliver on this promise, among them protein microarrays (MacBeath and Schreiber 2000). Currently, the most effective technique for cataloging the thousands of proteins in an average cell sample is mass spectrometry.

Studying the Proteome by Mass Spectrometry

The mass spectrometry proteomics experiments currently come in two general flavors: In one approach, proteins from a cell extract are first separated by two-dimensional (2D) gel electrophoresis. Then, proteins on the gel are identified by measuring their masses with MALDI-TOF mass spectrometry. The coups of this method include identifying many proteins from the yeast proteome (Shevchenko et al. 1996).

A second approach rids itself of the reliance on 2D electrophoresis and, therefore, promises to be technically simpler and more scaleable. In this approach (Fig. 1), proteins from a cell extract are first proteolytically digested into fragments (Hunt et al. 1986). The fragments are then partially purified by high-performance liquid chromatography and injected into an electrospray tandem mass spectrometer (MS/MS) or ion-trap mass spectrometer (LCQ). The mass spectrometer efficiently separates the peptide mixtures. In a continuous, automatic process, each peptide peak is in turn selected from the peptide mixture flowing into the mass spectrometer and sequenced by fragmenting the peptide in a collision cell and then measuring the masses of the peptide fragments.

Much like nucleic acid sequencing, in which the nucleic acid sequence is derived from the different mobilities of sequence fragments on a gel, the peptide sequences can be derived from the masses of their component fragments. With a partial sequence of the peptide in hand, coupled with measurements of masses of the peptide and some of its fragments, a database of protein sequences is searched to find a protein containing a matching peptide (Eng et al. 1994; Mann and Wilm 1994). Proteins whose peptides are identified in this way can be added to the growing catalog of proteins expressed under specific cell conditions. In this manner, many proteins of *Dienococcus radiodurans* and *Escherichia coli* have recently been cataloged (Jensen et al. 2000).

Finally, mass spectrometry researchers have established methods to quantitatively measure protein concentrations. In this breakthrough technique, one isotopically labels proteins harvested from cells grown under one set of conditions and then mixes those proteins with differentially labeled proteins from cells grown under different conditions (Gygi et al. 1999; Jensen et al. 2000). The result is that the quantity of each protein can be measured relative to a reference state, so quantitative changes in protein expression can be detected.

Measuring Modifications in the Protein Population

However, where mass spectrometry really promises to outshine other proteomics techniques is through its potential to detect mutations and modifications in proteins. Because peptides are sequenced during a mass spectrometry proteomics experiment, the researcher has the potential to detect specific mutated and modified amino acids among these sequenced peptides.

How common is posttranslational modification? A rough count of enzymes performing the most common modifications (Table 1) suggests that a large fraction, perhaps 5%–10%, of a genome encodes modification-catalyzing proteins. So, modifications are widespread and it is likely that the majority of proteins in the cell are deliberately posttranslationally modified, in addition to sporadic nonenzymatic modifications. In total, >200 different protein modifications are known to occur in cells (Gooley and Packer 1997), including such diverse modifications as ADP ribosylation, tyrosine nitration and sulfation, palmitoylation, and polyglycylation. Such modifications are fascinating in their modulation of the activity of cellular proteins, but unfortunately, the diversity of modifications creates significant difficulties for proteomics efforts.

Ironically, the physical measurement of modifications does not pose a problem when analyzing proteins by mass spectrometry. Instead, the difficulties are computational. When the only information one has about the identity of a peptide stems from the masses of its components, what does one do when, because of posttranslational modifications, the masses of the components change? Because of a combination of this problem and the quality of mass spectra and incomplete fragmentation of peptides, over one-half of the peptides in a typical proteomics experiment are never identified.

[1]Corresponding author.
E-MAIL marcotte@icmb.utexas.edu. FAX (512) 471–2149.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.178301.

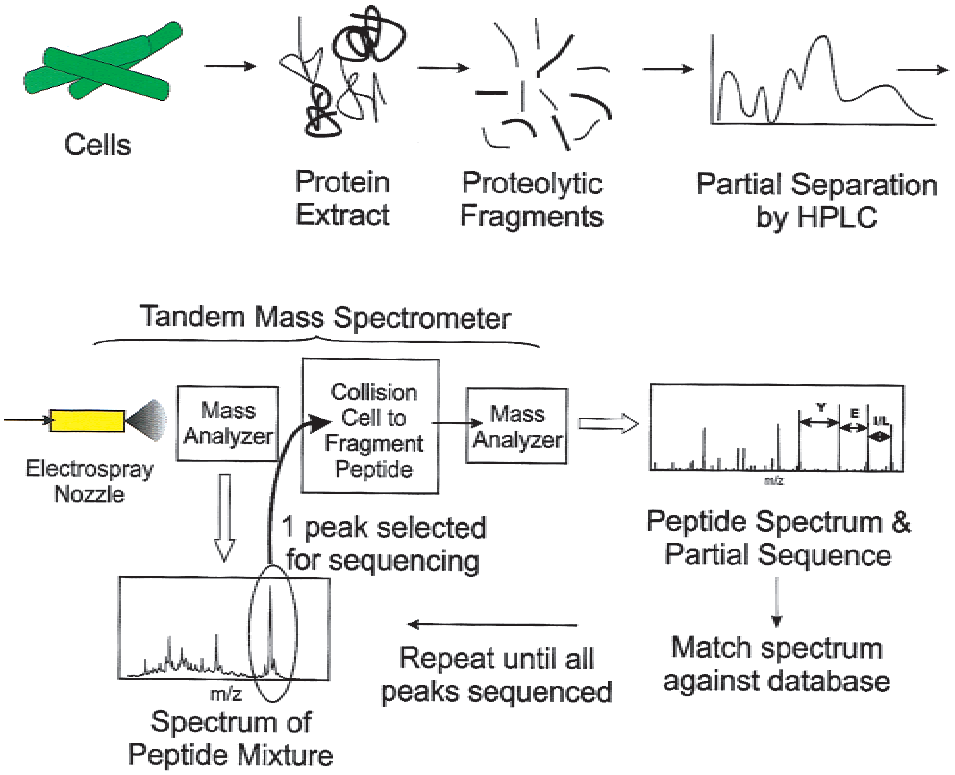

Figure 1 The combination of HPLC and tandem mass spectrometry allows the identification of potentially thousands of proteins from complex samples. In this approach to proteomics, mass spectra are generated for proteolytic peptides through the illustrated steps. Thousands of spectra may be generated, requiring sophisticated computational methods to match these spectra against spectra derived from databases of amino acid sequences.

Overcoming the Main Computational Barrier

In this issue, Pevzner et al. (2001) provide several new approaches to solve this computational problem. The essence of the problem is as follows: In a mass spectrometry experiment, a mass spectrum is generated for a particular peptide. One then attempts to match this experimental mass spectrum to mass spectra predicted from peptides in an amino acid sequence database. A good match indicates that the corresponding peptide, whose sequence is known from the database, is present in the sample. However, mass changes caused by a modification or mutation of the sample peptide change the experimental mass spectrum, and it requires clever algorithms to recognize the change.

Intuitively, to decide if two mass spectra are derived from the same peptide sequence, one might simply count the number of peaks shared by the two spectra. In fact, this simple method works for perfect matches but rapidly loses its efficacy with spectra of lower quality and with peptide mutations or modifications. A mass change caused by modification of one amino acid correspondingly changes the mass of all peptide fragments containing that amino acid. So, by looking only at the masses, one has ignored the relationships between the peaks in the spectra; the relative mass differences do not necessarily change, even though the absolute mass values may do so.

Pevzner et al. (2001) introduce three new algorithms, each of which significantly outperforms the simple counting of shared peaks. The first, spectral convolution, takes into account the mass differences between peaks in the two spectra being compared. If the peptides only differ because of the mutation/modification of one or two amino acids, the masses in the resultant spectra will also differ by a constant amount of mass. So, the algorithm looks for mass differences that show up more than expected, and the specific modification or mutation can be determined by the particular mass difference detected.

The second method, spectral alignment, represents a particularly elegant solution to the problem of aligning partially-mismatched spectra. Following their earlier work (Pevzner et al. 2000), they apply the most effective method of amino acid sequence alignment (e.g., Smith and Waterman 1981) to the problem of aligning mass spectra. Peaks in the spectra are represented as elements in a matrix, and alignments between the two spectra are then paths taken through this matrix. Using the technique of dynamic programming, one can find the best match between two spectra, given some number of allowed mismatches caused by modifications or mutations.

Earlier efforts in this field (Yates et al. 1995) centered on building a data-

Table 1. Number of Genes Dedicated to Catalyzing Several of the Most Common Post-Translational Modifications

	Yeast	Worm	Human
Protein kinases	~117[a]–120[b]	~381[c]–411[e]	~1100[e]
Protein phosphatases	~43[b]–52[a]	~106[c]–185[e]	~300[e]
Ubiquitin system proteins	>50[a]	>50[d]	>134[a]
Glycosyl transferases	>36[a]	?	?
Non-proteasomal proteases	~50[a]–75[b]	~194[c]	~700–1000[f]

The large number of proteins encoded in a genome catalyzing post-translational modifications suggests that most proteins in the cell will be modified.
[a–f]The numbers of genes are taken from (a) the MIPS database, (b) the YPD database of Proteome, Inc., (c) the WormPD database of Proteome, Inc., (d) the Sanger Center WormPep database, (e) Plowman et al. 1999, and (f) Southan 2000.

base of all possible mutated or modified peptides to match against one's spectrum. This rapidly leads to a combinatorial explosion if one models all of the possible mutations or modifications at each position of an amino acid sequence. In fact, the advantage of spectral convolution and spectral alignment is that one does not have to build a database of all possible mutations and modifications. Instead, the data reveal which modifications have taken place and leave open the possibility of discovering new modifications. However, when the number of possible modifications is small, building a database of modifications can still be an effective approach. So, in their third algorithm, Pevzner et al. (2001) introduce a method from computer science, a branch-and-bound algorithm, to improve the efficiency of this type of database search.

And in the Future

Remaining computational challenges include the details of how to combine these approaches most successfully and how to calculate an accurate estimate of the probability that a match is correct. With these in hand, the computational side of mass spectrometry proteomics will be in much the same situation as sequence homology searches. Comparative molecular biology has flourished with the ability to rapidly search sequences against databases and to estimate the significance of the results. In much the same way, introducing such rigorous computational methods to mass spectrometry will help fulfill the promise of proteomics and will make projects such as complete sequencing of proteomes possible.

At the rate at which the field of proteomics is advancing, it is reasonable to expect that we will soon have cellular protein expression maps, much as we now have for mRNA expression through DNA chip approaches. Coupled with these protein expression data will be information about the modifications of each of the proteins. So, we can imagine a multiple-dimension mapping of all protein expression in a cell as a function of cell condition. Each protein will not only be tallied for its abundance but also for its state: On, off, membrane-anchored, oxidized, and so on. This map of protein expression and modification, combined with genomewide mRNA expression patterns and protein interaction data, will provide our first global, integrated, quantitative picture of the major cellular processes of transcription, translation, and posttranslational modification and will provide the data necessary to construct fundamental, predictive rules, not merely descriptions, about how these processes operate.

REFERENCES

Eng, J., McCormack, A.L., and Yates III, J.R. 1994. *J. Am. Mass. Spectrom.* **5:** 976–989.

Gooley, A. and Packer, N. 1997. In *Proteome research: New frontiers in functional genomics* (eds. W. Wilkins et al.) pp. 65–91). Springer, NY.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. 1999. *Nat. Biotechnol.* **17:** 994–999.

Hunt, D.F., Yates III, J.R., Shabanowitz, J., Winston, S., and Hauer, C.R. 1986. *Proc. Natl. Acad. Sci.* **83:** 6233–6237.

Jensen, P.K., Pasa-Tolic, L., Peden, K.K., Martinovic, S., Lipton, M.S., Anderson, G.A., Tolic, N., Wong, K.K. and Smith, R.D. 2000. *Electrophoresis* **21:** 1372–1380.

MacBeath, G. and Schreiber, S.L. 2000. *Science* **289:** 1760–1763.

Mann, M. and Wilm, M. 1994. *Anal. Chem.* **66:** 4390–4399.

Pevzner, P.A., Dancik, V., and Tang, C. 2000. In *RECOMB 2000, Proceedings of the fourth annual International Conference on Computational Molecular Biology*, Tokyo, Japan, April 8–11.

Pevzner, P.A., Mulyukov, Z., Dancik, V., and Tang, C.L. 2001. *Genome Res.* **11:** 290–299.

Plowman, G.D., Sudarsanam, S., Bingham, J., Whyte, D., Hunter, T. 1999. *Proc. Natl. Acad. Sci.* **96:** 13603–13610.

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H. and Mann, M. 1996. *Proc. Natl. Acad. Sci.* **93:** 14440–14445.

Smith, T.F. and Waterman, M.S. 1981. *J. Mol. Biol.* **147:** 195–197.

Southan, C. 2000. *J. Pept. Sci.* **6:** 453–458.

Yates III, J.R., Eng, J.K., McCormack, A.L., and Schieltz, D. 1995. *Anal. Chem.* **67:** 1426–1436.