Absolute abundance for the masses

Christine Vogel & Edward M Marcotte

Mass spectrometry can now measure the absolute concentrations of the majority of cellular proteins without labeling.

Determining the absolute abundances of proteins on a proteome-wide scale has been a longstanding goal in systems biology. Although >1,000 proteins are routinely identified in a high-resolution mass spectrometry run, quantification is typically limited to measurements of relative protein concentrations. Now, in a breakthrough reported in *Nature*, Malmström *et al.*[1] have developed a combined approach that enabled estimation of the absolute abundances of more than half the known proteins of the pathogenic bacterium *Leptospira interrogans*. The authors integrate three methods for absolute quantification in a manner that will be widely applicable, even to mammalian systems, highlighting the ever-increasing capacity of mass spectrometry to analyze complex proteomes.

In a typical shotgun proteomics workflow[2], a protein sample (for example, a whole-cell lysate or a purified protein complex) is digested into peptides, and the resulting peptide mixture is partially separated by column chromatography and introduced into a mass spectrometer by means of electrospray ionization. Thousands of mass spectra are collected on successive samplings of the column eluate, and tandem mass spectrometry (MS/MS) spectra of the strongest peaks in each mass spectrum are collected periodically. Such peaks correspond mostly to unique peptides, having been purified both by chromatography and mass spectrometry. Usually, tens of thousands of MS/MS spectra are collected and used to computationally identify the peptides' amino acid sequences, providing a large list of peptides detected in the sample. Proteins are identified by the presence of their component peptides in this set.

Current methods can readily measure changes in relative protein concentrations; these are useful for comparing differences in protein abundances between conditions or cell types but tell us nothing about absolute protein concentrations. In principle, a peptide's signal intensity (the size of its mass spectrum peak) should be proportional to the absolute abundance of the peptide—and of the corresponding protein—in the sample. However, such estimates can be erroneous because of effects such as variable sequence-dependent peptide ionization efficiencies, suppression of neighboring signals by dominant peptides, and missing observations stemming from semi-stochastic peak selection for MS/MS analyses. As a consequence, measuring absolute abundances requires extra steps (**Fig. 1**).

One approach, termed selected reaction monitoring[3] (SRM), relies on samples spiked with isotopically labeled reference peptides for the proteins of interest. As the concentrations of the isotopically labeled reference peptides are known, relative signal intensities can be calibrated to an absolute scale. Although SRM is sensitive and highly reproducible across laboratories and platforms[4], and it can theoretically be extended to a full proteome, preparing thousands of isotopically labeled peptides of known concentration is both formidable and expensive.

Two recent computational approaches that do not require isotopic labels but rather calculate absolute abundances from data collected in routine shotgun proteomics experiments provide an inexpensive alternative to SRM[5]. The

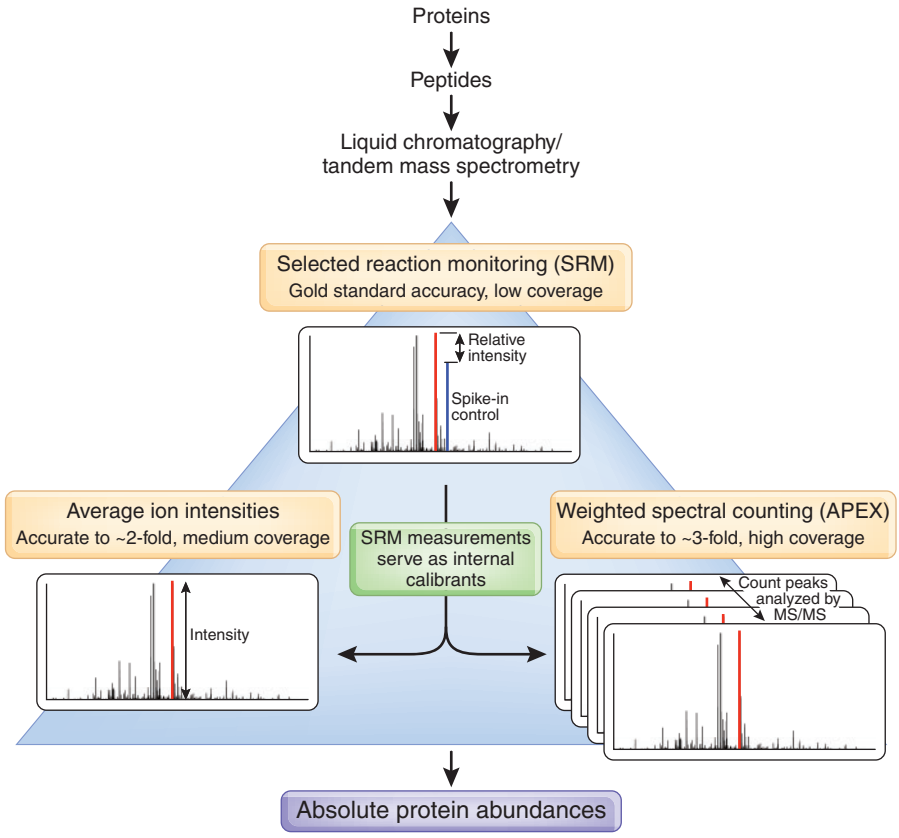

Figure 1 Large-scale measurement of absolute protein abundances by integrating three complementary methods for quantification of mass spectrometry data. Peptides analyzed by tandem mass spectrometry provide two main types of information about molecular concentrations: the intensities of each peptide's peaks in the mass spectra, and the number of times a peptide peak is observed, reflected in the count of tandem mass spectra observed for each peptide. With appropriate computational postprocessing, both types of data can be used to infer absolute concentrations of the original protein. To obtain data normalized to absolute concentrations, Malmström *et al.*[1] calibrated two large-scale methods with a small-scale, highly accurate method (SRM), which compares peak intensities of isotopically labeled and unlabeled peptides of known concentrations.

Christine Vogel and Edward M. Marcotte are at the Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, and Edward M. Marcotte is in the Department of Chemistry and Biochemistry; University of Texas at Austin, Austin, Texas, USA.
e-mail: marcotte@icmb.utexas.edu

first exploits mass spectrum signal intensities, the accuracy of which has greatly improved owing to recent advances in chromatography and ionization (for example, nanoflow electrospray ionization) and in mass spectrometers themselves (for example, the Thermo Electron Corporation LTQ/Orbitrap, which has an innovative mass analyzer[6]). As a consequence, Silva *et al.*[7] found that a protein's abundance could be well estimated from the average mass spectrum peak intensity of its three best-detected peptides. A second approach, spectral counting, analyzes the observed counts of MS/MS spectra attributable to each protein. In a recent development for large-scale absolute protein expression measurements (APEX), Lu *et al.*[8] improved the accuracy of spectral counting by incorporating differential peptide ionization propensities into the computation.

Malmström *et al.*[1] combine these three approaches—SRM measurements of a limited set of internal reference standards, the average mass spectrum signal intensities of the top three peptides selected per protein, and weighted MS/MS spectral counts—to more completely quantify the proteome (**Fig. 1**). By using the SRM measurements of reference standards to calibrate the two computational abundance calculations, they achieve abundances accurate to ~2-fold on average for 769 proteins using the approach of Silva *et al.*[7] and to ~3-fold for 1,095 more proteins with the technique of Lu *et al.*[8]. This enables them to measure abundances for >1,800 proteins, or 83% of the proteome detectable by mass spectrometry under these conditions and 51% of the *L. interrogans* proteome (based on predicted open reading frames). Combining the high accuracy of SRM with the high coverage of the two computational approaches minimizes the costs of isotopic labeling while maximizing coverage and accuracy (**Fig. 1**). The abundance estimates are validated with molecule concentrations measured by single-cell cryo-electron tomography for flagellar proteins, flagellar motors and periplasmic methyl-accepting chemotaxis protein receptors.

As with any mass spectrometry method, the techniques used by Malmström *et al.*[1] are limited by the peptides' amenability to ionization and by the mass spectrometer's ability to detect low abundance molecules. Although >200 of the ~1,000 proteins monitored after exposure of *L. interrogans* to the antibiotic ciprofloxacin changed their abundance more than twofold, the limitations of sensitivity for differentially expressed proteins may be even lower[8], depending on whether the observed quantification errors are consistent across samples and systematic in nature, which is unknown at present.

Although there is no theoretical upper limit to the size of the proteome for which this approach should be effective, current mass spectrometers and practices restrict it to a few thousand proteins; this covers the majority of proteins for simple organisms but typically represents only a fraction of the expressed proteome for higher organisms. Fractionation of samples before analysis can substantially increase the proteome coverage, but further work remains to determine how fractionation affects these quantification methods. For example, the SRM calibrants might have to be chosen appropriately to sample the different fractions. Perhaps more importantly, resolving the differential expression of splice variants, which are common in proteomes of higher organisms, is still a challenging problem in shotgun proteomics. Nonetheless, given that these approaches offer protein quantification without the need for genetic modification or extensive isotopic labeling, the combination of approaches presented by Malmström *et al.* should be widely applicable to many systems.

The availability of absolute protein concentration data will be indispensable to fulfilling the promise of systems biology. Owing to extensive post-transcriptional regulation, protein abundances are only partially correlated with the abundances of the corresponding mRNAs[8–10]. This has led many to argue that direct assessment of protein levels is often more informative of the cellular state than analysis of mRNA levels. Indeed, protein abundances seem more conserved across evolution than mRNA transcript abundances[10]. Quantitative mass spectrometry is now poised to routinely provide such data at large scale and with high accuracy—a testament to the rapid progress in quantitative shotgun proteomics over the last few years.

1. Malmström, J. *et al. Nature* **460**, 762–765 (2009).
2. Han, X., Aslanian, A. & Yates, J.R. III. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).
3. Lange, V. *et al. Mol. Syst. Biol.* **4**, 222 (2008).
4. Addona, T.A. *et al. Nat. Biotechnol.* **27**, 633–641 (2009).
5. Kito, K. & Ito, T. *Curr. Genomics* **9**, 263–274 (2008).
6. Hu, Q. *et al. J. Mass Spectrom.* **40**, 430–443 (2005).
7. Silva, J.C. *et al. Mol. Cell. Proteomics* **5**, 144–156 (2006).
8. Lu, P. *et al. Nat. Biotechnol.* **25**, 117–124 (2007).
9. Anderson, L. & Seilhamer, J. *Electrophoresis* **18**, 533–537 (1997).
10. Schrimpf, S.P. *et al. PLoS Biol.* **7**, e48 (2009).

Combinatorics and next-generation sequencing

Nick Patterson & Stacey Gabriel

The massive capacity of today's sequencing machines can be harnessed efficiently by sequencing pooled samples and decoding the results.

In the last year alone, the average yields of a single DNA sequencing instrument have increased by at least tenfold, and ten billion bases can now be obtained routinely in a single run. Indeed, for many applications, current sequencing throughput is vastly greater than what is needed to process a single sample—a situation that brings not only new opportunities but also new challenges. Two recent papers in *Genome Research*, by Erlich *et al.*[1] and Prabhu and Pe'er[2], present improved methods for exploiting this technological capability. Using ideas from a branch of mathematics called combinatorics, they show that thousands of pooled samples can be sequenced *en masse* and the results decoded.

The new sequencing technologies will have many applications[3], but here we concentrate on methods for discovery of rare mutations, which are likely to account for much of the genetic basis of disease. The high yields of the latest instruments allow us to deeply sequence genes of medical interest for thousands of individuals[3]. As only tens to hundreds of kilobases are of interest in such studies, and as even the smallest functional unit of a sequencer—a single 'lane'—generates data amounting to many thousand–fold coverage of such targets, the challenge is how to use a sequencer efficiently on samples requiring only a fraction of its capacity. An equally daunting challenge is the need to individually amplify and create sequencing templates for thousands of samples. The cost of the amplification and the difficulties of sample tracking and automation are substantial.

Pooling DNA samples promises to solve both of these challenges. Grouping many samples together in each run makes the most effective use of the high depth of sequencing coverage and alleviates the problem of handling many individual samples. Simply mixing all of the samples together, however, makes it impossible to determine which individual contributed

Nick Patterson and Stacey Gabriel are at the Broad Institute, Cambridge, Massachusetts, USA. e-mail: nickp@broadinstitute.org